А.Н.ТИХОНОВ, Д.П.КОСТОМАРОВ

РАССКАЗЫ О ПРИКЛАДНОЙ МАТЕМАТИКЕ

А. Н. ТИХОНОВ, Д. П. КОСТОМАРОВ

РАССКАЗЫ О ПРИКЛАДНОЙ МАТЕМАТИКЕ

МОСКВА «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
1979

22.19
Т 46
УДК 519.6

Рассказы о прикладной математике. Тихонов А. Н., Костомаров Д. П. Наука. Главная редакция физико-математической литературы, М., 1979.

В книге в популярной форме рассказывается о прикладной математике, о применении математических методов и электронно-вычислительных машин к решению прикладных задач. Основное внимание уделяется построению математических моделей изучаемых объектов, вычислительным алгоритмам и электронно-вычислительным машинам.

Изложение построено на базе материала, который либо входит в программу средней школы, либо близко к ней примыкает.

Т $\frac{20204-089}{053(02)-79}$ 91-79. 1702070000

ОГЛАВЛЕНИЕ

Предисловие 5

Введение. **Научно-технический прогресс и математика** 7

Глава 1. **Математические модели** 12

§ 1. Пусть дано 12
§ 2. Соответствие математической модели изучаемому объекту. Критерий практики 15
§ 3. Развитие и уточнение математической модели . . 19

Глава 2. **Вычислительные алгоритмы** 30

§ 1. Понятие алгоритма 30
§ 2. Алгоритмы извлечения квадратного корня . . . 33
§ 3. Число π и его вычисление 41

Глава 3. **Электронно-вычислительные машины** 49

§ 1. От 10 пальцев к ЭВМ 49
§ 2. Как работают ЭВМ 52
§ 3. Поколения ЭВМ и проблемы общения человека и машины 61
§ 4. Применение ЭВМ 78

Глава 4. **Численные методы решения уравнений** . . . 83

§ 1. Решение уравнения в виде формулы не правило, а исключение 83
§ 2. Качественное исследование уравнений. Теорема о существовании корня у непрерывной функции 85
§ 3. Метод вилки 88
§ 4. Метод итераций (метод последовательных приближений) 93
§ 5. Метод касательных (метод Ньютона) 99
§ 6. Заключительные замечания 103

Глава 5. **Задачи оптимизации** 106

§ 1. Задача о наилучшей консервной банке 107
§ 2. Одномерные задачи оптимизации 109
§ 3. Одномерные задачи оптимизации. Продолжение 112
§ 4. Частные производные и градиент функции нескольких переменных 118
§ 5. Многомерные задачи оптимизации 121

Глава 6. Линейное программирование 131
§ 1. Если бы директором был я 131
§ 2. Математическая постановка задачи линейного программирования 141
§ 3. Симплекс-метод 145
§ 4. Снова задача о стульях 148

Глава 7. Определенный интеграл. Численное интегрирование 152
§ 1. Как подсчитать путь при неравномерном движении или работу переменной силы 152
§ 2. Формула Ньютона — Лейбница 156
§ 3. Понятие определенного интеграла 157
§ 4. Интегрируемость монотонных функций 159
§ 5. Алгоритмы численного интегрирования 166

Глава 8. Дифференциальные уравнения 176
§ 1. Задача о зеркале прожектора, о колебании маятника и некоторые другие задачи 176
§ 2. Дифференциальные уравнения первого порядка . 182
§ 3. Метод ломаных Эйлера 187
§ 4. Дифференциальные уравнения высших порядков и системы дифференциальных уравнений 192
§ 5. Задача о колебании маятника 194
§ 6. Расчет траектории ядра с учетом сопротивления воздуха 197
§ 7. Как послать космический корабль к Луне . . . 199

Предметный указатель 205

ПРЕДИСЛОВИЕ

Создание в середине XX века электронно-вычислительных машин (ЭВМ) можно в некотором смысле сопоставить с изобретением паровой машины или использованием электричества. Однако ЭВМ занимают в ряду этих величайших достижений человечества особое место: если обычные машины расширяли физические возможности людей, то ЭВМ существенно повысили их интеллектуальный потенциал. Широкое применение математических методов и ЭВМ открыло новые возможности увеличения производительности труда, дальнейшего развития производства, совершенствования управления. Вычислительные машины привели к появлению новых эффективных методов познания законов реального мира и их использования в практической деятельности людей.

Процесс математизации науки, техники, экономики потребовал подготовки высококвалифицированных специалистов, способных реализовать те огромные и пока еще далеко не исчерпанные возможности, которые дает применение ЭВМ. Вычислительные машины не работают без направляющего воздействия человека. Их использование связано с построением математических моделей изучаемых объектов и созданием вычислительных алгоритмов. Машины также должны пройти соответствующее «обучение», т. е. получить программное обеспечение как общего, так и специально ориентированного характера.

Весь этот широкий комплекс проблем является полем деятельности специалистов по прикладной математике. Для их подготовки во многих университетах и институтах созданы специальные факультеты, отделения, кафедры.

Однако сегодня среди пользователей ЭВМ наряду с профессиональными математиками-вычислителями много представителей других специальностей: инженеров, физиков, химиков, экономистов, социологов и т. д. Завтра круг людей, которым в своей производственной деятель-

ности нужно будет уметь грамотно пользоваться математическими методами и ЭВМ, станет еще шире.

Это обстоятельство послужило основной причиной, побудившей авторов написать данную книгу. В ней в популярной форме рассказывается о современной прикладной математике, о характерных особенностях применения математических методов и ЭВМ к изучению реальных «нематематических» объектов, встречающихся в прикладных задачах.

Книга начинается с трех глав, посвященных трем основным элементам прикладной математики: математическим моделям, вычислительным алгоритмам и электронно-вычислительным машинам. Последующие главы носят более специальный характер: в них разбираются типичные задачи прикладной математики и описываются методы их решения. Изложение каждой темы начинается с вопросов, которые входят в программу средней школы. Однако они подаются под таким углом зрения, чтобы от них легко можно было перейти к реальным задачам прикладной математики.

Огромную помощь в работе над 3-й главой книги оказали авторам Л. Н. Королев и Р. Л. Смелянский. Целый ряд существенных замечаний и предложений сделали Н. Н. Ефимов — по 1-й и В. Г. Карманов — по 6-й главам. Всем им авторы выражают свою самую искреннюю благодарность.

Академик А. Н. ТИХОНОВ
Профессор Д. П. КОСТОМАРОВ

Факультет вычислительной математики
и кибернетики МГУ

Введение

НАУЧНО-ТЕХНИЧЕСКИЙ ПРОГРЕСС И МАТЕМАТИКА

Одной из характерных особенностей нашего времени является широкое применение математических методов и электронно-вычислительных машин в самых различных областях человеческой деятельности. «Диагноз ставит ЭВМ», «Соавтор конструктора» — такие заголовки нередко встречаются сегодня в газетах. Бурный процесс математизации науки, техники, народного хозяйства начался в пятидесятых годах после появления и быстрого совершенствования ЭВМ. Он привел к формированию современной прикладной математики, которая включает круг вопросов, связанных с использованием математических методов и вычислительной техники. Этому научному направлению уделено большое внимание в решениях XXIV и XXV съездов КПСС. Одна из важнейших задач, поставленных XXV съездом перед советской наукой, сформулирована следующим образом: «Расширять исследования по теоретической и прикладной математике. Развивать научные работы, направленные на создание и эффективное применение в народном хозяйстве электронно-вычислительной техники».

Цель настоящей книги заключается в том, чтобы в популярной форме рассказать широкому кругу читателей и прежде всего учащейся молодежи о прикладной математике, об идеях, методах, трудностях исследований, связанных с применением математических методов и вычислительной техники к изучению законов природы и их использованию в практической деятельности людей.

Математика является одной из самых древних наук. Она зародилась на заре человеческой цивилизации под влиянием потребностей практики. Строительство, измерение площадей земельных участков, навигация, торговые расчеты, управление государством требовали умения производить арифметические вычисления и определенных геометрических знаний. В дальнейшем математика раз-

вилась в стройную логическую систему, как составная часть общего комплекса научных знаний. Потребности естествознания, техники, всей практической деятельности людей постоянно ставили перед математикой новые задачи и стимулировали ее развитие. В свое очередь прогресс в математике делал математические методы более эффективными, расширял сферу их применения и, тем самым, способствовал общему научно-техническому прогрессу.

Роль математики в различных областях человеческой деятельности и в разное время была существенно различной. Она складывалась исторически, и существенное влияние на нее оказывали два фактора: уровень развития математического аппарата и степень зрелости знаний об изучаемом объекте, возможность описать его наиболее существенные черты и свойства на языке математических понятий и уравнений или, как теперь принято говорить, возможность построить «математическую модель» изучаемого объекта.

Математическая модель, основанная на некотором упрощении, идеализации, не тождественна объекту, а является его приближенным отражением. Однако благодаря замене реального объекта соответствующей ему моделью появляется возможность сформулировать задачу его изучения как математическую и воспользоваться для анализа универсальным математическим аппаратом, который не зависит от конкретной природы объекта. Математика позволяет единообразно описать широкий круг фактов и наблюдений, провести их детальный количественный анализ, предсказать, как поведет себя объект в различных условиях, т. е. спрогнозировать результаты будущих наблюдений. А ведь прогнозирование — всегда трудная задача, и оправдывающиеся прогнозы являются предметом особой гордости любой науки.

Сложность построения и исследования математической модели существенно зависит от сложности изучаемого объекта. Математические методы давно и весьма успешно применяются в механике, физике, астрономии, т. е. в науках, в которых изучаются наиболее простые формы движения материи. Математика стала языком этих наук, относящихся к разряду «точных». Значительную роль играла также математика в технике. Этим вплоть до недавнего времени исчерпывалась сфера широкого применения математических методов. Ситуация резко измени-

лась с появлением ЭВМ. Причина этого заключается в следующем. В математике часто встречаются задачи, решение которых не удается получить в виде формулы, связывающей искомые величины с заданными. Про такие задачи говорят, что они не решаются в явном виде. Для их решения стремятся найти какой-нибудь бесконечный процесс, сходящийся к искомому ответу. Если такой процесс указан, то, выполняя определенное число шагов и затем обрывая вычисления (их нельзя продолжать бесконечно), мы получим приближенное решение задачи. Эта процедура связана с проведением вычислений по строго определенной системе правил, которая задается характером процесса и называется алгоритмом.

Такой подход к решению математических задач был известен еще до появления ЭВМ, но применялся весьма редко из-за исключительной трудоемкости больших вычислений. Когда Лаверье «открыл» за письменным столом «на кончике пера» новую планету (Нептун), рассчитав ее траекторию по возмущениям траектории планеты Уран, то это было научным подвигом, навсегда вошедшим в историю науки. Однако в большинстве случаев исследователи стремились избегать больших вычислений. Поэтому сложные математические модели, для которых не удавалось получить ответа в виде формул, либо вообще не рассматривались, либо упрощались с помощью дополнительных предположений. Упрощение модели снижало степень ее соответствия изучаемому объекту, делало результаты исследования объекта менее точными и, следовательно, менее интересными, а иногда и приводило к ошибкам.

Опытный вычислитель тратил на выполнение одного арифметического действия в среднем за рабочую смену около полминуты. Современные ЭВМ выполняют миллионы операций в секунду. Таким образом, за короткий промежуток времени, порядка 30 лет, благодаря ЭВМ скорость проведения вычислений возросла примерно в 100 миллионов раз. Такого скачка не было за всю историю человечества ни в одной сфере человеческой деятельности.

Применение численных методов на базе ЭВМ сразу существенно расширило класс математических задач, допускающих исчерпывающий анализ. Теперь уже исследователю при построении математической модели какого-то объекта не нужно стремиться к упрощениям, которые были необходимы раньше при желании полу-

чить ответ в явном виде. Его внимание, прежде всего, должно быть направлено на то, чтобы правильно учесть все наиболее существенные особенности изучаемого объекта и отразить их в математической модели. После того, как модель построена, встает вопрос о разработке алгоритма решения соответствующей математической задачи и его реализации на ЭВМ. Таким образом, ЭВМ изменили подход к применению математики как метода исследования. Они вызвали переориентацию многих сложившихся направлений математики и развитие ряда новых. Сегодня ЭВМ являются одним из определяющих факторов научно-технического прогресса. Их применение способствует ускорению развития ведущих отраслей народного хозяйства, открывает принципиально новые возможности проектирования сложных систем при значительном сокращении сроков их разработки и внедрения в производство, обеспечивает выбор оптимальных режимов производственно-технологических процессов, создает условия для совершенствования управления и повышения производительности труда. Если обычно машины брали на себя физические функции человека в процессе производства, делали его сильнее, то ЭВМ помогают человеку в умственной деятельности, делают его умнее. Они являются одним из важных факторов превращения науки в непосредственную производительную силу нашего общества. Без ЭВМ не могли бы развиваться многие крупные научно-технические проекты современности (космические исследования, атомная энергетика, сверхзвуковая авиация и т. д.).

Благодаря ЭВМ идет интенсивный процесс математизации не только естественных и технических, но также и общественных наук. Важное значение приобрело применение математических методов в экономике. Математическое моделирование начинает широко использоваться в химии, геологии, биологии, медицине, психологии, лингвистике. Большое внимание уделяется подготовке высококвалифицированных кадров, способных реализовывать те огромные возможности, которые открывает эффективное использование ЭВМ. Во многих университетах и институтах созданы факультеты прикладной или вычислительной математики. Подтверждается точка зрения К. Маркса, который, по словам П. Лафарга, считал, что «наука только тогда достигает совершенства, когда ей удается пользоваться математикой».

В заключение остановимся коротко на содержании книги. Она начинается с трех глав, посвященных трем основным элементам прикладной математики: математическим моделям, вычислительным алгоритмам и электронно-вычислительным машинам. Эти главы дают общее представление о современной прикладной математике.

Последующие главы носят более специальный характер. В них разбираются различные типичные задачи прикладной математики и методы их решения. Эти главы практически не зависят друг от друга и могут читаться в любом порядке (но обязательно после трех первых глав). Однако мы считаем, что выбранный в книге порядок, подчиненный, в частности, принципу постепенного возрастания сложности, является наиболее естественным.

Изложение материала в каждой главе начинается с вопросов, которые в той или иной степени входят в программу средней школы и вам знакомы. Однако они подаются под таким углом зрения, чтобы было легко сделать следующий шаг: перейти от школьных вопросов к реальным задачам прикладной математики.

Книга, конечно, не охватывает всех сторон современной прикладной математики. Это обусловлено и небольшим ее объемом, и тем, что она рассчитана на читателей, которые знакомы с математикой в объеме средней школы. При отборе материала какую-то роль сыграл субъективный фактор, связанный с профессиональными интересами авторов. В частности, в книге очень коротко рассказано об электронно-вычислительных машинах, остались практически не затронутыми вопросы программирования. Эти темы могли бы стать предметом отдельной книги.

Мы надеемся, что книга пробудит дополнительный интерес к математике и ее многочисленным приложениям, поможет вам по-новому взглянуть на достаточно широкий круг математических идей и понятий, которые изучаются в настоящее время в школе, научит лучше использовать эти знания на практике. Мы также рассчитываем на то, что у некоторых из вас появится желание стать специалистами в этой очень интересной области. Что же, факультет вычислительной математики и кибернетики МГУ и многочисленные родственные факультеты других вузов нашей страны ждут вас. Добро пожаловать.

Глава 1
МАТЕМАТИЧЕСКИЕ МОДЕЛИ

§ 1. Пусть дано...

Такими словами явно или неявно обычно начинаются формулировки теорем и условий математических задач. Затем на языке строго определенных математических понятий следует полное изложение исходных предпосылок, которое воспринимается совершенно одинаково любым математиком, являющимся специалистом в соответствующей области.

Иначе обстоит дело с прикладными задачами. В них непосредственно задается реальный «нематематический» объект: явление природы, производственный процесс, конструкция, система управления, экономический план и т. д. Исследование начинается с формализации объекта, с построения соответствующей математической модели: выделяются его наиболее существенные черты и свойства и описываются с помощью математических уравнений. Только после того, как построена математическая модель, т. е. задаче придана математическая форма, мы можем воспользоваться для ее изучения математическими методами.

Вы знакомы с математическими моделями, хотя, может быть, раньше и не встречали этого термина. Представьте себе, что нужно определить площадь комнаты или, если быть более точным, площадь пола комнаты. Для выполнения такого задания измеряют длину и ширину комнаты, а затем перемножают полученные числа. Эта элементарная процедура фактически означает следующее. Реальный объект — пол комнаты — заменяется абстрактной математической моделью — прямоугольником. Прямоугольнику приписываются размеры, полученные в результате измерения, и площадь такого прямоугольника приближенно принимается за искомую площадь пола.

Вспомните задачи по физике. В них обычно задается некоторая физическая система и условия, в которых она

находится. Вы сами должны сделать предположения о возможной идеализации этой системы (например, рассматривать некоторое реальное тело как материальную точку), выделить физические законы, которые нужно принять во внимание при ее изучении, и записать их в виде математических уравнений. Это и есть математическая модель рассматриваемой физической системы.

Обсудим в качестве примера следующую задачу по механике. Телу на Земле сообщили начальную скорость v_0, направленную под углом α к ее поверхности. Требуется найти траекторию движения тела и вычислить расстояние между ее начальной и конечной точками.

Чтобы сделать задачу более конкретной, предположим, что речь идет о камне, брошенном с помощью катапульты. Это уточнение определяет характерные размеры тела, его вес и возможную начальную скорость. Построим для данного случая математическую модель, основанную на следующих предположениях:

1) Земля — инерциальная система отсчета;

2) ускорение свободного падения g постоянно;

3) кривизной Земли можно пренебречь и считать Землю плоской;

4) влиянием воздуха на движение камня можно пренебречь.

Здесь сформулированы только наиболее важные предположения. Попытка оговорить абсолютно все увела бы нас слишком далеко от цели.

Введем систему координат. Ее начало совместим с катапультой, ось x направим горизонтально в сторону движения камня, ось y — вертикально вверх. При сделанных предположениях камень будет двигаться вдоль оси x равномерно со скоростью $v_x = v_0 \cos \alpha$. Движение камня в вертикальном направлении равноускоренное с ускорением $a_y = -g$ и начальной скоростью $v_y = v_0 \sin \alpha$. Таким образом, характер движения камня определяется формулами:

$$x = tv_0 \cos \alpha, \tag{1}$$

$$y = tv_0 \sin \alpha - gt^2/2, \tag{2}$$

которые дают математическую модель задачи при предположениях 1) — 4).

Полученная модель является очень простой, и с ее помощью легко получить ответ на поставленный вопрос.

Выразим из формулы (1) время t через координату x:

$$t = \frac{x}{v_0 \cos \alpha}$$

и подставим в формулу (2). В результате получим уравнение траектории камня:

$$y = x \operatorname{tg} \alpha - x^2 \frac{g}{2v_0^2 \cos^2 \alpha}, \qquad (3)$$

представляющей собой параболу (см. рис. 1). Эта пара-

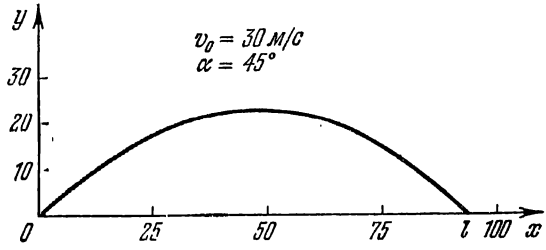

Рис. 1. Параболическая траектория движения камня.

бола пересекает ось x в двух точках: $x = 0$ и $x = l$, где

$$l = \frac{v_0^2}{g} \sin 2\alpha. \qquad (4)$$

Первая точка является началом траектории, в ней камень вылетает из катапульты. Вторая точка соответствует месту падения камня на Землю. Формула (4) в рамках принятой модели определяет искомое расстояние l. Эта формула вам хорошо известна: она выводится и подробно обсуждается в учебнике по физике для 8-го класса.

В прикладных задачах построение математической модели — это один из наиболее сложных и ответственных этапов работы. Опыт показывает, что во многих случаях правильно выбрать модель — значит решить проблему более чем наполовину. Трудность данного этапа состоит в том, что он требует соединения математических и специальных знаний. При решении школьных задач по физике вы выступаете одновременно как физик и математик. Однако для больших проблем, которые рассматриваются в прикладной математике, такое совмещение профессий нетипично. Обычно над математической моделью

совместно работают математики и специалисты из той области, к которой относится изучаемый объект. Для успеха их деятельности очень важно взаимопонимание, которое приходит тогда, когда математики обладают специальными знаниями об объекте, а их партнеры — определенной математической культурой, опытом применения математических методов исследования в своей области. В противном случае совместная работа легко может превратиться в диалог глухих со слепыми. Так нередко бывало, особенно на раннем этапе развития современной прикладной математики, пока обе стороны не научились лучше понимать друг друга.

§ 2. Соответствие математической модели изучаемому объекту. Критерий практики

Математическая модель никогда не бывает тождественна рассматриваемому объекту, не передает всех его свойств и особенностей. Основанная на упрощении, идеализации, она является его приближенным отражением. Поэтому результаты, которые получаются при анализе модели, всегда носят для объекта приближенный характер. Их точность определяется степенью соответствия, адекватности модели и объекта. Вопрос о точности, о достоверности результатов — это один из самых тонких вопросов прикладной математики.

Наиболее просто он решается в случае, когда хорошо известны законы, определяющие поведение и свойства объекта и имеется большой практический опыт их применения. Тогда можно априори (до опыта, здесь до начала решения математической задачи) оценить точность результатов, которую обеспечивает рассматриваемая модель.

Например, для расчета траектории советского космического корабля Луна-1, который доставил впервые в истории человечества на Луну вымпел с изображением герба Советского Союза, была использована математическая модель, основанная на законах механики и всемирного тяготения. (Мы подробно познакомим вас с этой моделью в последнем параграфе главы 8). До этого полета не было прямой экспериментальной проверки справедливости такой модели для космических аппаратов, созданных человеком. Однако многовековой опыт изучения движения небесных тел в Солнечной системе говорил о том, что

данная модель позволяет очень точно рассчитывать их траектории. Поэтому, в силу универсальности законов природы, еще до начала данного космического эксперимента не было никаких оснований сомневаться в справедливости траекторных расчетов.

Более сложная ситуация возникает тогда, когда наши знания об изучаемом объекте недостаточны. В этом случае при построении математической модели приходится делать дополнительные предположения, которые носят характер гипотез. Выводы, полученные в результате исследования такой гипотетической модели, носят для изучаемого объекта условный характер. Они справедливы для него настолько, насколько правильны исходные предположения. Для их проверки необходимо сопоставить результаты исследования модели со всей имеющейся информацией об изучаемом объекте. Степень близости расчетных и экспериментальных данных позволяет судить о качестве гипотетической модели, о справедливости или несправедливости исходных предположений. Таким образом, вопрос применимости некоторой математической модели к изучению рассматриваемого объекта не является чисто математическим вопросом и не может быть решен математическими методами. Основным критерием истинности является эксперимент, практика в самом широком смысле этого слова. Критерий практики позволяет сравнить различные гипотетические модели и выбрать из них такую, которая является наиболее простой и в то же время в рамках требуемой точности правильно передает свойства изучаемого объекта.

Для иллюстрации этих соображений вернемся к задаче о траектории полета камня, выбрасываемого катапультой, и продолжим ее обсуждение. В предыдущем параграфе мы построили математическую модель движения камня, основанную на четырех упрощающих предположениях, и получили формулу (4) для дальности броска. Теперь нам нужно оценить точность этой формулы, установить пределы ее применимости. Для такого анализа вовсе не обязательно делать по старым чертежам катапульту, доставать какой-нибудь прибор для измерения начальной скорости камня (например, кинокамеру, делающую 10 000 кадров в минуту), рулетку и идти на ближайший пустырь для метания камней. Нет, интересующим нас вопросам в науке уделялось самое пристальное внимание, по ним накоплен огромный экспериментальный

и теоретический материал, так что нужно только умело воспользоваться им для анализа поставленной задачи.

Перечитайте еще раз упрощающие предположения, на основании которых была построена математическая модель движения камня, и вдумайтесь в их смысл. Пусть катапульта может метать камни на расстояние до 100 м, для чего она должна сообщить им скорость порядка 30 м/с. При этом камень поднимется на высоту 20—30 м и пробудет в воздухе около 5 с. В этих условиях первые три предположения выглядят совершенно оправданными, и нам нужно проанализировать четвертое предположение о влиянии воздуха.

На всякое тело, движущееся в воздухе, он действует с некоторой силой $\boldsymbol{F}$. Ее величина и направление зависят от формы тела и скорости движения. Силу $\boldsymbol{F}$ можно разложить на две составляющие: параллельную и перпендикулярную скорости движения тела v.

Перпендикулярная составляющая возникает только при наличии асимметрии тела по отношению к направлению движения. Наиболее характерным ее проявлением является подъемная сила, действующая на крыло самолета, без которой была бы невозможна авиация. Для того чтобы эта сила могла оторвать самолет от земли и поддерживать его в воздухе, крылу придают специальную форму и располагают его под определенным «углом атаки» к набегающему воздушному потоку. Однако для камня, форма которого близка к сферической, перпендикулярная составляющая силы $\boldsymbol{F}$ мала, и ею можно пренебречь (для шара из-за его симметрии она точно равна нулю).

Параллельная составляющая силы $\boldsymbol{F}$ возникает всегда. Она направлена в сторону, противоположную движению, и стремится затормозить тело. Ее называют лобовым сопротивлением. Таким образом, в интересующем нас случае

$$\boldsymbol{F} \approx \boldsymbol{F}_{\text{л}}. \tag{5}$$

Величину лобового сопротивления (модуль вектора $\boldsymbol{F}_{\text{л}}$) можно представить в виде

$$F_{\text{л}} = CS\,\frac{\rho v^2}{2}, \tag{6}$$

где S — площадь поперечного сечения тела, ρ — плотность воздуха, v — скорость движения тела, C — безразмерный множитель, который называют коэффициентом лобового сопротивления.

Коэффициент лобового сопротивления зависит от формы тела и характеристики процесса обтекания, которую называют числом Рейнольдса:

$$C = C(\mathrm{Re}), \quad \mathrm{Re} = \frac{vd\rho}{\mu}. \tag{7}$$

Здесь d — характерный размер тела, ρ и μ — плотность и вязкость воздуха ($\rho = 1{,}3$ кг/м^3, $\mu = 1{,}7 \cdot 10^{-5}$ кг/м с).

Оценим величину числа Рейнольдса в интересующем нас случае, когда $v = 30$ м/с, $d = 0{,}2$ м $= 20$ см:

$$\mathrm{Re} = \frac{30 \cdot 0{,}2 \cdot 1{,}3}{1{,}7 \cdot 10^{-5}} = 4{,}6 \cdot 10^5 \tag{8}$$

(число Рейнольдса — величина безразмерная).

Экспериментальные и теоретические исследования показывают, что для шара в широком диапазоне чисел Рейнольдса, включающих значение (8):

$$3 \cdot 10^5 \leqslant \mathrm{Re} \leqslant 7 \cdot 10^6, \tag{9}$$

коэффициент лобового сопротивления C очень слабо зависит от своего аргумента Re, и его можно считать постоянным:

$$C \approx 0{,}15. \tag{10}$$

Подставляя (10) в (6) и полагая $S = \pi R^2$, получим простую формулу для величины лобового сопротивления шара при условии (9):

$$F_л = \frac{C\pi}{2} R^2 \rho v^2. \tag{11}$$

Эта формула, в частности, указывает на квадратичную зависимость $F_л$ от скорости.

Для того чтобы оценить влияние сопротивления воздуха на характер движения, сравним его с основной силой в рассматриваемой задаче, с силой тяжести:

$$\mathrm{P} = mg = \frac{4\pi}{3} R^3 \rho_0 g,$$

где ρ_0 — плотность камня ($\rho_0 = 2{,}3 \cdot 10^3$ кг/м^3). Заменим в формуле (6) v^2 на lg и составим отношение $F_л/\mathrm{P}$:

$$\frac{F_л}{\mathrm{P}} = \frac{\frac{1}{2} C\pi R^2 \rho l g}{\frac{4\pi}{3} R^3 \rho_0 g} = \frac{3C}{8} \cdot \frac{\rho l}{\rho_0 R}. \tag{12}$$

При $l = 100$ м, $R = 0{,}1$ м, считая $C \approx 0{,}15$, получаем:

$$F_л/P = 0{,}03.$$

Таким образом, если мы работаем в рассматриваемом диапазоне параметров и если нам не требуется высокая точность в определении l, так что ошибка порядка 2—3% (2—3 м) допустима, то применение модели без сопротивления воздуха оправдано. В противном случае такую модель следует отвергнуть и заменить более сложной моделью, учитывающей сопротивление воздуха.

В заключение подчеркнем, что наша оценка рассмотренной математической модели не является абсолютной: модель годится или модель не годится. Проведенный анализ дал возможность установить для нее условия применимости, связав их с диапазоном изменения параметров задачи и требуемой точностью. Он также позволил наметить путь уточнения модели в случае, когда условия ее применимости перестают выполняться.

§ 3. Развитие и уточнение математической модели

Исследование прикладных задач обычно начинается с построения и анализа простейшей, наиболее грубой математической модели рассматриваемого объекта. (Характерным примером может служить модель параболической траектории полета тела, получившего на поверхности Земли начальную скорость v_0.) Однако в дальнейшем часто возникает необходимость уточнить модель, сделать ее соответствие объекту более полным. Это может быть обусловлено разными причинами: требованием более высокой точности, появлением новой информации об объекте, которую нужно отразить в математической модели, расширением диапазона параметров, выводящим за пределы применимости исходной модели, и т. д. При построении новой модели полезно максимально полно использовать опыт и результаты, полученные на первом этапе. Часто процесс последовательного развития и уточнения модели повторяется неоднократно.

Для иллюстрации этих соображений вернемся снова к задаче о движении тела, брошенного с поверхности Земли, и рассмотрим ее применительно к внешней баллистике. Так называют науку о движении снаряда, вылетевшего из ствола орудия. Мы выбрали баллистику не только потому, что ее задачи интересны с математической

и важны с практической точек зрения. Не менее существенна другая сторона вопроса: на этом примере мы сможем наглядно показать процесс постепенного совершенствования и уточнения математической модели рассматриваемого явления, который исторически продолжался более 300 лет.

Древние воины, которые использовали катапульты для разрушения неприятельских укреплений, не знали законов механики и не могли теоретически рассчитать траекторию полета камня даже в рамках очень простой модели. Впрочем, это и не было нужно: камнеметание велось на небольшие расстояния «на глазок». Положение изменилось с изобретением пороха и появлением артиллерии, существенно увеличившей дальность, интенсивность и эффективность обстрела.

Исследования по внешней баллистике были начаты в XVII веке Г. Галилеем, который разработал теорию параболического движения снаряда, рассмотренную в § 1. Следующий шаг связан с именем И. Ньютона. В своем основном труде «Математические начала натуральной философии» (1687 г.) он предпринял попытку решить задачу баллистики с учетом сопротивления воздуха. Этот шаг был совершенно необходим: из результатов § 2 ясно, что при всем несовершенстве орудий XVII века воздух должен был оказывать заметное влияние на движение ядер, вызывая отклонение их траектории от параболической (3).

Действительно, проведем с помощью формулы (12) оценку роли сопротивления воздуха для чугунного ядра, полагая $l = 1$ км $= 10^3$ м, $R = 0{,}07$ м, $\rho_0 = 7 \cdot 10^3$ кг/м^3 (плотность чугуна), $C = 0{,}15$. При указанных значениях параметров будем иметь:

$$F_{\text{л}}/\text{P} = 0{,}15,$$

т. е. пренебрежение сопротивлением воздуха приводит к ошибке в определения дальности l порядка 15% (около 150 м).

Рассмотрим математическую модель баллистики пушечного ядра, учитывающую сопротивление воздуха. При ее построении мы по-прежнему будем считать справедливыми первые три предположения предыдущей модели, а четвертое предположение переформулируем следующим образом:

4) воздух действует на ядро при его движении с силой $\boldsymbol{F}_л$, ее величина определяется формулой (11), а направление противоположно направлению скорости $\boldsymbol{v}$.

Уравнение движения ядра (второй закон Ньютона) при этих предположениях имеет вид

$$m\boldsymbol{a} = \boldsymbol{P} + \boldsymbol{F}_л. \tag{13}$$

Для его анализа введем такую же систему координат, как в предыдущем случае, и запишем разложение векторов $\boldsymbol{P}$ и $\boldsymbol{F}_л$ по единичным базисным векторам $\boldsymbol{i}$, $\boldsymbol{j}$, направленным вдоль координатных осей x и y:

$$\boldsymbol{P} = -mg\boldsymbol{j}, \tag{14}$$

$$\boldsymbol{F}_л = -F_л(v)\frac{v_x}{v}\boldsymbol{i} - F_л(v)\frac{v_y}{v}\boldsymbol{j} =$$

$$= -\frac{C\pi}{2}\cdot R^2\rho(vv_x\boldsymbol{i} + vv_y\boldsymbol{j}). \tag{15}$$

Формула (14) очевидна: сила тяжести $\boldsymbol{P}$ направлена вертикально вниз и равна по величине mg. Формула (15) следует из рис. 2. На нем показаны в некоторый момент

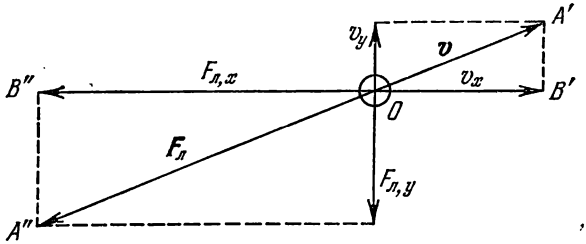

Рис. 2. Определение проекций лобового сопротивления $\boldsymbol{F}_л$ на координатные оси x, y.

времени t вектор скорости $\boldsymbol{v}$ и вектор лобового сопротивления $\boldsymbol{F}_л$, имеющий противоположное направление. Величины проекций $F_{л,x}$ и $F_{л,y}$ могут быть найдены из подобия треугольников $OA'B'$ и $OA''B''$.

Заменим векторное уравнение (13), разделив его предварительно на массу ядра m, двумя скалярными уравнениями для проекций на координатные оси x и y:

$$a_x = -\frac{C\pi}{2}\frac{R^2\rho}{m}vv_x,$$

$$a_y = -g - \frac{C\pi}{2}\frac{R^2\rho}{m}vv_y. \tag{16}$$

Пусть $x(t)$, $y(t)$ — искомые функции, определяющие координаты ядра в любой момент времени его полета, тогда производные этих функций дают скорость ядра:

$$v_x = x'(t), \quad v_y = y'(t),$$

$$v = \sqrt{v_x^2 + v_y^2} = \sqrt{(x')^2 + (y')^2}. \tag{17}$$

Повторное дифференцирование позволяет подсчитать ускорение:

$$a_x = x''(t), \quad a_y = y''(t). \tag{18}$$

Подставляя (17) и (18) в систему (16), получим систему уравнений

$$x'' = -\frac{C\pi}{2}\frac{R^2\rho}{m}\sqrt{(x')^2 + (y')^2}\,x',$$

$$y'' = -g - \frac{C\pi}{2}\frac{R^2\rho}{m}\sqrt{(x')^2 + (y')^2}\,y'. \tag{19}$$

Уравнения подобного типа, связывающие искомые функции и их производные, в математике называются дифференциальными уравнениями. Мы получили систему двух дифференциальных уравнений для функций $x(t)$, $y(t)$, представляющую собой в рамках сформулированных выше предположений математическую модель баллистики ядра с учетом сопротивления воздуха.

Дифференциальным уравнениям в нашей книге посвящена глава 8. В ней обсуждаются особенности таких уравнений и методы их решения. В частности, § 6 главы 8 специально посвящен системе (19). Сейчас же, не вдаваясь в детали, обсудим результат некоторых расчетов для этой системы.

На рис. 3 и 4 приведены две траектории. В обоих случаях предполагалось, что чугунное ядро радиуса $R = 0{,}07$ м вылетает под углом 45° к поверхности Земли. Начальная скорость v_0 принималась равной 60 м/с для траектории на рис. 3 и 90 м/с — для траектории на рис. 4. Штриховыми линиями даны параболические траектории (3) в модели § 1. Сравнение сплошных и штриховых линий наглядно демонстрирует влияние сопротивления воздуха на движение ядра при различных начальных скоростях.

Рис. 5, 6, 7 являются суммирующими для рассматриваемой задачи. На них сплошными и штриховыми линиями показаны зависимости дальности l, максимальной

высоты h и времени полета T от начальной скорости v_0 с учетом и без учета сопротивления воздуха. При отсутствии сопротивления воздуха высота h и время T связаны простым соотношением

$$h = \frac{1}{8} gT^2. \qquad (20)$$

Результаты проведенных расчетов показывают, что эта формула с высокой степенью точности выполняется при

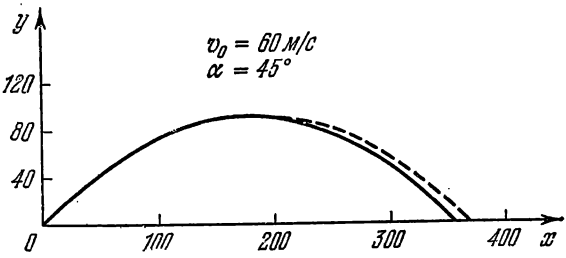

Рис. 3. Влияние сопротивления воздуха на движение пушечного ядра, вылетающего с начальной скоростью 60 м/с. Сплошная линия — траектория, рассчитанная с учетом сопротивления воздуха, штриховая — без учета.

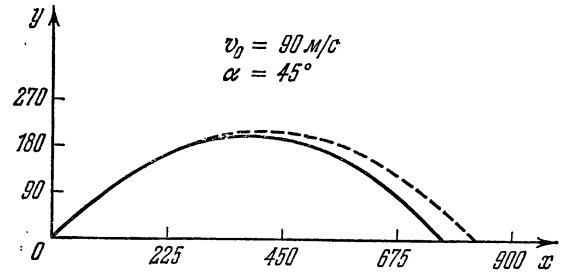

Рис. 4. Влияние сопротивления воздуха на движение пушечного ядра, вылетающего с начальной скоростью 90 м/с. Сплошная линия — траектория, рассчитанная с учетом сопротивления воздуха, штриховая — без учета.

движении ядра в воздухе, несмотря на заметное отличие общей картины движения. Этим свойством широко пользуются в баллистике.

Переход в XIX веке от гладкоствольного к нарезному оружию позволил существенно увеличить дальность и точность стрельбы. Появилась возможность вести огонь

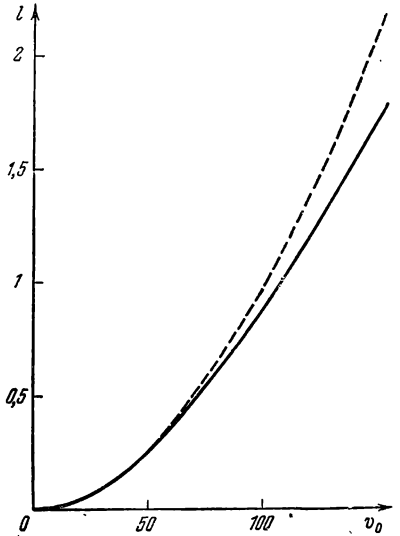

Рис. 5. Зависимость дальности полета пушечного ядра от начальной скорости, рассчитанная с учетом (сплошная линия) и без учета (штриховая линия) сопротивления воздуха.

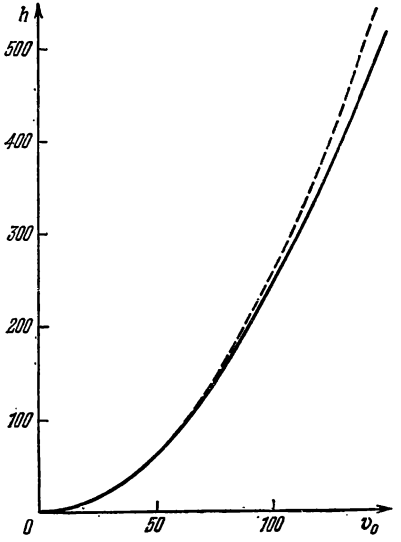

Рис. 6. Зависимость максимальной высоты, на которую поднимается пушечное ядро, от начальной скорости, рассчитанная с учетом (сплошная линия) и без учета (штриховая линия) сопротивления воздуха.

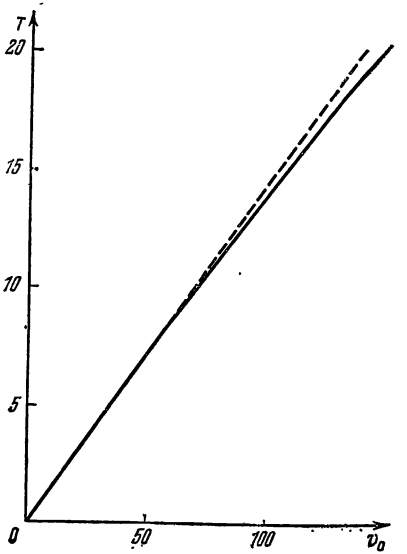

Рис. 7. Зависимость времени движения пушечного ядра от начальной скорости, рассчитанная с учетом (сплошная линия) и без учета (штриховая линия) сопротивления воздуха.

с закрытых позиций по приборам, не видя непосредственно цели. Это потребовало совершенствования прицельных приспособлений, поставило новые задачи перед баллистикой и привело к дальнейшему уточнению используемых математических моделей.

Наиболее характерной особенностью баллистики снарядов нарезного оружия является сверхзвуковая скорость их движения. При таких скоростях около снаряда образуется волна сильного сжатия, которая расходится от него в виде конуса (так называемый конус Маха) с углом раствора $\varphi = \arcsin \frac{c}{v}$, где v — скорость снаряда, c — скорость звука (при нормальных условиях $t = 15°$, $p = 760$ мм рт.ст., $c = 340$ м/с). Эта волна хорошо видна на рис. 8, где приведено изображение снаряда, движущегося со скоростью $v = 2{,}48\ c$.

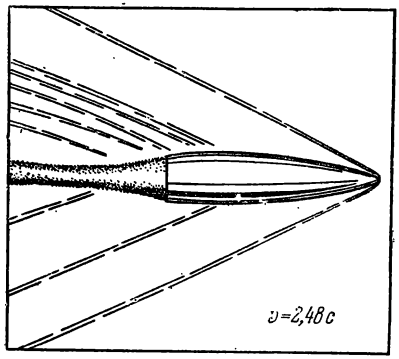

Рис. 8. Волна сжатия вокруг снаряда, движущегося со сверхзвуковой скоростью $v = 2{,}48c$.

Лобовое сопротивление при сверхзвуковом движении является, в первую очередь, волновым. Оно связано с тем, что снаряд должен непрерывно расходовать свою кинетическую энергию на образование волны. Для его уменьшения снаряду придают специальную форму вытянутого тела вращения с заостренным носом (см. рис. 8).

Величину лобового сопротивления, как и в предыдущем случае, удобно представить в виде формулы (11). Однако для сверхзвуковых скоростей ($v > c$) коэффициент C не будет постоянным. Он оказывается функцией скорости $C = C(v)$, резко возрастающей при ее увеличении. Обычно значения коэффициента лобового сопротивления $C(v)$ при различных скоростях v измерялись экспериментально в аэродинамической трубе. Только в последнее время благодаря ЭВМ появилась возможность теоретического расчета.

Если коэффициент лобового сопротивления $C(v)$ тем или другим способом найден, то, подставляя его в систему дифференциальных уравнений (19), мы получим матема-

тическую модель сверхзвукового движения снаряда, с помощью которой можно рассчитать траекторию.

До сих пор мы предполагали, что все величины, которые входят в условия задачи баллистики, известны абсолютно точно. На самом деле это не так. Снаряды нельзя изготовить абсолютно одинаковыми: они немного отличаются друг от друга по весу, по величине порохового заряда, и при выстреле получают разные начальные скорости. Невозможно дважды выстрелить под одним и тем же углом возвышения α. Эти и целый ряд других неконтролируемых факторов приводят к тому, что два снаряда, выпущенные из орудия при одинаковых, на наш взгляд, условиях, никогда не попадут в одну и ту же точку: из-за действия случайных факторов снаряды рассеиваются. Это означает, что любое предсказание баллистического расчета носит не строго определенный, детерминированный характер, а является вероятностным. В конце XIX — начале XX веков в математических моделях баллистики начали использовать методы теории вероятности.

Сложность задач баллистики делает невозможным их решение непосредственно на поле боя. В помощь артиллеристам стали создаваться специальные таблицы, получившие название артиллеристских таблиц. В них для определенного типа орудий и снарядов (т. е. для заданной начальной скорости v_0) приводились основные характеристики траекторий при разных углах возвышения α: дальность l, максимальная высота h, время полета снаряда T. Такие таблицы были уже широко распространены к началу первой мировой войны.

Дальнейшее увеличение скорости снарядов и дальности стрельбы привело к необходимости внести в математическую модель баллистики еще одно уточнение — учесть вращение Земли вокруг своей оси. Тем самым пришлось отказаться от предположения 1), считавшего Землю инерциальной системой отсчета. Анализ показывает, что из-за вращения Земли всякое движущееся тело должно в северном полушарии отклоняться немного вправо, в южном — влево. С этим эффектом связаны, в частности, два хорошо известных явления: у наших рек правый берег обычно подмывается сильнее левого, а на железных дорогах быстрее снашивается правый рельс.

Мы не будем останавливаться на деталях математического описания данного эффекта и его включения в об-

щую модель баллистики снаряда. Приведем только один поучительный эпизод из истории первой мировой войны.

В 1914 году около Фольклендских островов, которые расположены в Атлантическом океане вблизи побережья Южной Америки (51° S, 60° W), произошел морской бой между английской и немецкой эскадрами, закончившийся победой англичан. Однако в начале боя залпы английских кораблей систематически ложились метров на 100 левее цели и потребовалась специальная корректировка в установках прицела, чтобы устранить отклонение. Произошло это по следующей причине. Прицельные приспособления на английских кораблях автоматически вносили поправку на вращение Земли. Однако она была рассчитана на средние широты северного полушария (50° N), бой же произошел на той же широте южного полушария, где отклонение снарядов из-за вращения Земли должно иметь противоположный знак. В результате коррекция, вносимая прицельным устройством, не компенсировала смещение, а удваивала его, и сдвиг достиг величины порядка 100 м.

Заканчивая обзор математических задач баллистики, остановимся еще на одном событии первой мировой войны. В ходе войны немцы сумели подойти к Парижу на расстояние меньше 100 км, но у них не хватило сил овладеть столицей Франции. Тогда командование германской армии приняло решение подвергнуть Париж артиллерийскому обстрелу. Для этой цели специально были изготовлены сверхдальнобойные орудия «Колоссаль». Они представляли собой огромные установки (длина ствола — 34 м, калибр — 210 мм), которые монтировались на специальной платформе и имели дальность стрельбы свыше 100 км. Снаряд посылался под фиксированным углом возвышения $\alpha = 52{,}5°$, круто уходил вверх и основную часть пути летел на высоте, превышающей 20 км, практически не испытывая сопротивления воздуха. Дальность стрельбы регулировалась величиной порохового заряда, вес которого доходил до 200 кг. Ствол из-за своей огромной длины легко деформировался и выдерживал небольшое число выстрелов.

Стрельба на такие большие расстояния потребовала сложных баллистических расчетов на основе усовершенствованной математической модели. В ней учитывались зависимость плотности воздуха и ускорения свободного

падения от высоты, топогеофизические данные о местности, включающие кривизну земной поверхности (Земля в задачах баллистики перестала быть «плоской»), метеорологические данные о скорости и направлении ветра, давлении воздуха и т. д. Несмотря на тщательную подготовку каждого выстрела, наблюдалось большое рассеивание снарядов: из 303 снарядов, выпущенных за время войны по Парижу, 120 упало за его пределами. Существенного влияния на ход и исход первой мировой войны обстрел Парижа не оказал.

Подведем итоги обсуждения математических моделей задач баллистики. Оно началось с простейшей модели Г. Галилея, которая описана в школьном учебнике по физике и основана на предположениях, сильно упрощающих задачу. Такая модель справедлива в очень узком диапазоне начальных скоростей: $v_0 \leqslant 30$ м/с.

Затем была рассмотрена модель, учитывающая сопротивление воздуха. При этом предполагалось, что коэффициент лобового сопротивления в формуле (6) является постоянным ($C \approx 0{,}15$). Эта модель справедлива в диапазоне дозвуковых скоростей: v_0 не больше 250—300 м/с.

Следующая модель позволила перейти к сверхзвуковым скоростям. Однако для ее реализации необходимо знать зависимость коэффициента лобового сопротивления рассматриваемого тела от скорости.

Далее в модель были также включены эффект вращения Земли, учет сферической формы ее поверхности и неоднородности g, метеорологических данных, случайных факторов, которые приводят к рассеянию снарядов и придают изучаемым закономерностям статистический характер. В результате была построена модель баллистики, применимая во всем диапазоне скоростей, которые были достигнуты с помощью огнестрельного оружия. Подчеркнем, что в ней оказались пересмотренными все исходные упрощающие предположения модели Г. Галилея.

Каждая модель, применявшаяся в баллистике, проходила проверку практикой. С появлением нарезного оружия, существенно увеличившего дальность и точность стрельбы, начали играть важную роль полигонные испытания. На них проверялось не только оружие, но и баллистические расчеты.

В пятидесятых годах нашего века с появлением ракет различного назначения возникли новые задачи — задачи ракетной баллистики. Наиболее сложная из них состоит

и расчете траектории ракеты с работающими двигателями в земной атмосфере с учетом программы управления. Подчеркнем: для того чтобы управлять, нужно перерабатывать поступающую информацию в режиме реального, «живого» времени. Это можно сделать только с помощью ЭВМ. Данная задача чрезвычайно сложна, и мы не имеем возможности на ней останавливаться. Более простой является задача расчета траектории ракеты за пределами земной атмосферы, когда двигатели прекратили работу. Такая задача космической баллистики обсуждается в последнем параграфе главы 8.

В заключение отметим еще раз, что математические модели позволяют свести исследование реального «нематематического» объекта к решению математической задачи, воспользоваться для его изучения универсальным математическим аппаратом и получить благодаря этому об объекте детальную количественную информацию. В этом заключаются огромные возможности математики как метода познания законов реального мира и их использования в практичской деятельности людей.

Глава 2

ВЫЧИСЛИТЕЛЬНЫЕ АЛГОРИТМЫ

Построение модели рассматриваемого объекта позволяет поставить задачу его изучения как математическую. После этого наступает второй этап исследования — поиск метода решения сформулированной математической задачи. Следует иметь в виду, что в прикладных работах нас, как правило, интересуют количественные значения искомых величин, т. е. ответ должен быть доведен «до числа». Все расчеты проводятся с числами, записанными в виде конечных десятичных дробей, поэтому результаты вычислений всегда принципиально носят приближенный характер. От этого никуда не уйти, и важно только добиться того, чтобы ошибки укладывались в рамки требуемой точности.

§ 1. Понятие алгоритма

В большинстве задач, с которыми вы встречались до сих пор в математике, ответ давался в виде формулы. Формула определяла последовательность математических операций, которую нужно выполнить для вычисления искомой величины. Например, формула корней квадратного уравнения позволяет найти их по значениям коэффициентов этого уравнения, формула Герона выражает площадь треугольника через длины его сторон и т. д.

Однако вам известны задачи, для которых ответ легко может быть найден, хотя он и не записывается в виде формулы. Вспомните младшие классы, когда вы учили целые числа и арифметические операции над ними. Можно ли назвать «формулой» правило вычисления суммы нескольких чисел с помощью поразрядного сложения столбиком? В то же время это правило полностью решает поставленную задачу: оно определяет последовательность математических операций, которую нужно выполнить для вычисления искомой величины.

Рассмотрим еще один известный вам пример — задачу отыскания наибольшего общего делителя (сокращенно НОД) двух целых чисел n_1 и n_2 (для определенности предположим, что $n_1 > n_2$). Общей формулы для решения этой задачи, которая выражала бы НОД через заданные числа n_1 и n_2, не существует. Однако можно указать универсальные методы, которые позволяют найти НОД любых двух целых чисел.

Один из таких методов заключается в последовательном переборе чисел n_2, $n_2 - 1$, $n_2 - 2$ и т. д. Процесс продолжается до тех пор, пока мы не обнаружим число, являющееся одновременно делителем n_1 и n_2. Такой подход всегда приведет к решению поставленной задачи, хотя вряд ли нужно доказывать, что он является трудоемким и неэффективным.

Рассмотрим другой, гораздо более интересный метод решения той же задачи. Разделим n_1 на n_2 в целых числах с остатком. Пусть при таком делении получены частное r_1 и остаток n_3 ($0 \leqslant n_3 < n_2$). Это означает, что число n_1 представимо в виде

$$n_1 = r_1 n_2 + n_3. \qquad (1)$$

Если остаток n_3 равен нулю, то задача решена: число n_2 — НОД пары чисел n_1, n_2. В противном случае ($n_3 \neq 0$) из формулы (1) вытекает следующее утверждение: всякий общий делитель чисел n_1, n_2 является одновременно общим делителем чисел n_2, n_3 и наоборот. Отсюда, в частности, следует, что НОД чисел n_1, n_2 равен НОД чисел n_2, n_3.

Будем теперь искать НОД чисел n_2, n_3. Для этого снова разделим n_2 на n_3 в целых числах с остатком. Пусть при этом получился остаток n_4 ($0 \leqslant n_4 < n_3 < n_2$). Если он равен нулю, то искомый НОД равен n_3. В противном случае заменим пару чисел n_2, n_3 на пару чисел n_3, n_4 и продолжим процесс. В результате после конечного числа шагов, не превышающего n_2, мы придем к решению рассматриваемой задачи.

Итак, мы видим, что для решения математической задачи важно указать систему правил, которая задает строго определенную последовательность математических операций, приводящих к искомому ответу. Такую систему правил называют *алгоритмом*. Понятие алгоритма в его общем виде относится к числу основных понятий математики.

В простейшем случае последовательность математических операций, с помощью которых можно вычислить искомые величины, определяется формулами. Так, формула Герона является алгоритмом вычисления площади треугольника по его сторонам. Однако описанный выше метод нахождения НОД дает пример алгоритма, который не сводится к формуле. Его называют *алгоритмом Евклида.*

Алгоритмы решения многих математических задач, для которых не удается получить ответ в виде формулы, основаны на следующей процедуре: строится бесконечный процесс, сходящийся к искомому решению. Он обрывается на некотором шаге (вычисления нельзя продолжать бесконечно), и полученная таким образом величина приближенно принимается за решение рассматриваемой задачи. Сходимость процесса гарантирует, что для любой заданной точности $\varepsilon\,(\varepsilon > 0)$ найдется такой номер шага N, что при этом шаге ошибка в определении решения задачи не превысит ε. Пусть слова «приближенное решение» не означают для вас «решение второго сорта». Если вы получите в какой-нибудь задаче ответ в виде формулы и захотите подсчитать по ней значение нужной величины, то, как мы уже подчеркивали в начале главы, из-за представления чисел при вычислениях конечными десятичными дробями вы сможете получить для нее лишь приближенное значение.

Например, формула, выражающая площадь треугольника через длины двух сторон и угол между ними, является точной. Однако при вычислении синуса угла вы всегда будете вынуждены ограничиться его приближенным значением. То же самое можно сказать о формуле для корней квадратного уравнения. Входящее в нее выражение $\sqrt{b^2 - 4ac}$, как правило, придется вычислять приближенно. Помните также, что в прикладных исследованиях точное решение математической задачи все равно дает лишь приближенные результаты для изучаемого объекта, о чем подробно говорилось в предыдущей главе.

Проблема применения алгоритмов, использующих бесконечный сходящийся процесс,— не в приближенном характере ответа, а в большом объеме необходимых вычислений. Не случайно такие алгоритмы принято называть вычислительными алгоритмами, а основанные на них методы решения математических задач — численными методами. Широкое применение вычислительных алгоритмов стало возможным только благодаря ЭВМ. До

их появления численные методы использовались редко и только в сравнительно простых случаях из-за чрезвычайной трудоемкости вычислений вручную.

Заканчивая параграф, сделаем несколько замечаний:

1) при разработке вычислительных алгоритмов особенное внимание уделяется тому, чтобы они были удобны для машинного счета;

2) опыт показывает, что гораздо выгоднее развивать универсальные алгоритмы для решения широкого класса типичных математических задач, чем строить частные алгоритмы для решения каждой задачи по отдельности;

3) изучение объектов самой различной природы часто приводит к одним и тем же математическим задачам. Поэтому имеется благоприятная возможность выделить задачи, которые часто встречаются в приложениях, изучить их особенности, разработать эффективные алгоритмы и реализовать эти алгоритмы в виде стандартных программ для ЭВМ.

Во втором и третьем параграфах мы разберем алгоритмы решения двух хорошо вам известных задач: извлечение квадратного корня и вычисление π. Это сравнительно частные, хотя и важные задачи, и мы начинаем с них «для затравки», для первого знакомства с вычислительными алгоритмами. В следующих главах будут рассматриваться более сложные и общие по своей постановке математические задачи. При этом каждый раз существенное внимание будет уделяться универсальным алгоритмам их решения.

§ 2. Алгоритмы извлечения квадратного корня

Вычисление квадратного корня из положительного числа может быть проведено разными способами. Обсудим два алгоритма решения этой задачи.

Алгоритм последовательного определения цифр в представлении числа $c = \sqrt{a}$ с помощью десятичной дроби. Алгоритм сложен для описания. Поэтому, излагая последовательность математических операций, мы будем иллюстрировать каждый шаг на примере извлечения корня из числа $a = 81725{,}3$ с точностью до 10^{-2}.

Переходя к формулировке алгоритма, будем считать, что значение корня вычисляется с точностью до 10^{-k}, т. е. что требуется найти цифры в десятичном представлении искомого числа до k-го знака после запятой включительно.

1. Цифры, входящие в целую часть числа a, разбиваются справа налево (от низших разрядов к высшим) на группы по две цифры. В случае, когда общее число цифр в целой части a нечетное, первая группа слева будет состоять только из одной цифры. Общее число образовавшихся при этом групп m определяет число цифр в целой части искомого корня:

$$c = \sqrt{a} = c_1c_2 \ldots c_m,\ c_{m+1}c_{m+2} \ldots$$

Цифры, образующие дробную часть числа a, также разбиваются на группы по две цифры, но слева направо (от высших разрядов к низшим). При нечетном числе цифр в дробной части к последней цифре справа следует приписать нуль, так что все группы разбиения дробной части окажутся полными.

Если число групп в дробной части a больше k, то лишние группы справа нужно отбросить, если же оно меньше k, то недостающие группы следует образовать из нулей.

В результате такой процедуры цифры числа a оказываются разбитыми на $m + k$ групп. Все группы нумеруются слева направо. Числа, образованные цифрами этих групп, обозначаются через $a_1, a_2, \ldots, a_{m+k}$. Число a_1 может состоять из одной или двух цифр, все остальные числа двузначные.

В нашем примере разбиение числа a на группы имеет вид

$$a = 8\ 17\ 25,\ 30\ 00.$$

Первая группа оказалась неполной: в нее входит только одна цифра. Из целой части числа a образовались три группы цифр ($m = 3$), следовательно, целая часть корня будет состоять из трех цифр. После запятой стояла только одна цифра. Чтобы вычислить корень с точностью до 10^{-2} ($k = 2$), мы приписали к ней справа три нуля и получили две группы. Числа, образованные цифрами групп, имеют в данном случае вид: $a_1 = 8$, $a_2 = 17$, $a_3 = 25$, $a_4 = 30$, $a_5 = 00$. Итак,

$$c = \sqrt{8\ 17\ 25,\ 30\ 00} = c_1c_2c_3,c_4c_5 \qquad (2)$$

2. Первая цифра c_1 числа $c = \sqrt{a}$ определяется как целочисленное значение корня из a_1 с недостатком, т. е.

$$c_1 \leqslant \sqrt{a_1} < c_1 + 1. \qquad (3)$$

При этом будем иметь:

$$10^m c_1 \leqslant c = \sqrt{a} < 10^m (c_1 + 1). \qquad (4)$$

В рассматриваемом примере c_1 — целочисленное значение с недостатком числа $\sqrt{8}$, т. е.

$$c_1 = 2. \qquad (5)$$

3. Вторая цифра c_2 определяется из двухстороннего неравенства

$$(10c_1 + c_2)^2 \leqslant 100a_1 + a_2 < (10c_1 + c_2 + 1)^2, \qquad (6)$$

которое удобно переписать в виде

$$(20c_1 + c_2)\cdot c_2 \leqslant b_2 < (20c_1 + c_2 + 1)(c_2 + 1), \qquad (7)$$

где

$$b_2 = 100\cdot(a_1 - c_1^2) + a_2. \qquad (8)$$

Неравенство (6) означает, что

$$10^m c_1 + 10^{m-1} c_2 \leqslant c = \sqrt{a} < 10^m c_1 + 10^{m-1}(c_2 + 1). \qquad (9)$$

В рассматриваемом примере $b_2 = 100\cdot(8-4) + 17 = 417$, и неравенство (7) принимает вид

$$(40 + c_2)\cdot c_2 \leqslant 417 < (40 + c_2 + 1)(c_2 + 1). \qquad (10)$$

Для подбора c_2 заменим левую часть условия (10) более грубым неравенством: $40\cdot c_2 \leqslant 417$. Ему удовлетворяет самое большое однозначное число $c_2 = 9$. Однако при подстановке его в (10) получается противоречие: $(40 + 9)\ 9 = 441 > 417$. Учитывая это, уменьшим предполагаемое значение c_2 на единицу, т. е. попробуем взять $c_2 = 8$. Подставляя это число в (10), будем иметь: $384 < 417 < 441$. Неравенство (10) выполняется. Таким образом,

$$c_2 = 8. \qquad (11)$$

4. Третья цифра c_3 определяется из двухстороннего неравенства

$$(100c_1 + 10c_2 + c_3)^2 \leqslant 10\,000a_1 + 100a_2 + a_3 < (100c_1 + 10c_2 + c_3 + 1)^2, \qquad (12)$$

которое удобно переписать в виде

$$(20\cdot(10c_1 + c_2) + c_3)\cdot c_3 \leqslant b_3 < (20\cdot(10c_1 + c_2) + c_3 + 1)(c_3 + 1), \qquad (13)$$

где

$$b_3 = 10\,000a_1 + 100a_2 + a_3 - (100c_1 + 10c_2)^2 = \\ = 100\cdot(b_2 - (20c_1 + c_2)\cdot c_2) + a_3. \quad (14)$$

Неравенство (12) означает, что

$$10^m c_1 + 10^{m-1}c_2 + 10^{m-2}c_3 \leqslant c = \sqrt{a} < \\ < 10^m c_1 + 10^{m-1}c_2 + 10^{m-2}(c_3 + 1). \quad (15)$$

В рассматриваемом примере $b_3 = 100\,(417-384) + 25 = 3325$, и неравенство (13) принимает вид

$$(560 + c_3)\cdot c_3 \leqslant 3325 < (560 + c_3 + 1)(c_3 + 1). \quad (16)$$

Для подбора c_3 заменим левую часть условия (16) более грубым неравенством: $560\,c_3 \leqslant 3325$. Ему удовлетворяет $c_3 = 5$. Подставляя это число в (16), будем иметь: $2825 < 3325 < 3396$.

Неравенство (16) выполняется. Таким образом,

$$c_3 = 5. \quad (17)$$

5. Вычисление последующих цифр числа c проводится по тем же правилам и прекращается после определения последней нужной цифры c_{m+k}. В результате число $c = \sqrt{a} = c_1 \ldots c_m, c_{m+1} \ldots c_{m+k}$ оказывается вычисленным с точностью $\varepsilon = 10^{-k}$.

Применение описанной процедуры к рассматриваемому примеру дает следующий результат: $\sqrt{8\ 17\ 25,\ 30\ 00} = 285{,}87$ или точнее: $285{,}87 < \sqrt{81725{,}3} < 285{,}88$.

Теперь, когда алгоритм извлечения квадратного корня сформулирован, на примере подсчета $\sqrt{81725{,}3}$ покажем схему, по которой удобнее всего проводить вычисления и делать записи.

$$\sqrt{8\ \ 17\ \ 25,\ 30\ \ 00} = 285{,}87$$

	-4
48×8	$4\ \ 17$
	$-3\ \ 84$
565×5	$33\ \ 25$
	$-28\ \ 25$
5708×8	$5\ \ 00\ \ 30$
	$-4\ \ 56\ \ 64$
57167×7	$43\ \ 66\ \ 00$
	$-\ 40\ \ 01\ \ 69$
	$3\ \ 64\ \ 31\ \ 00$

Поясним эту схему. Легко определить первую цифру: $c_1 = 2$. Записываем ее в ответ, потом возводим в квадрат, вычитаем из числа $a_1 = 8$ и сносим к результату $a_1 - c_1^2 = 4$ число $a_2 = 17$. При этом разность $a_1 - c_1^2$ смещается на две позиции влево, т. е. умножается на 100. В результате мы получаем число b_2:

$$b_2 = 100 \cdot (a_1 - c_1^2) + a_2 = 417,$$

записанное в третьей строчке.

Теперь можно перейти к определению второй цифры c_2. Для этого слева от числа $b_2 = 417$ проведем вертикальную прямую и напишем за ней на месте десятков число $2 \cdot c_1 = 4$, а на место единиц подберем с помощью неравенства (10) цифру c_2. Мы уже знаем, что данному условию удовлетворяет $c_2 = 8$.

Заносим ее в ответ, а также ставим на место единиц около 4, берем произведение $48 \cdot 8 = (20 \cdot c_1 + c_2) \cdot c_2 = 384$ и пишем результат справа от вертикальной черты под числом $b_2 = 417$. Вычисляем разность и сносим к ней число $a_3 = 25$. При этом разность $b_2 - (20 \cdot c_1 + c_2) \cdot c_2$ смещается на две позиции влево, т. е. умножается на 100. В результате мы получаем число b_3:

$$b_3 = 100\,(b_2 - (20 \cdot c_1 + c_2) \cdot c_2) + a_3 = 3325,$$

записанное в пятой строчке.

Дальнейшие вычисления проводятся аналогично. Они приведены на схеме. Мы последовательно получаем: $c_3 = 5$, $b_4 = 50\,030$, $c_4 = 8$, $b_5 = 436\,600$, $c_5 = 7$, $b_6 = 3\,643\,100$. Последнее число непосредственно нам не нужно, однако оно потребовалось бы для продолжения вычислений и подсчета следующих цифр.

Описанный алгоритм достаточно сложен, и определение каждой новой цифры связано с возрастающим объемом вычислений. Изучение этого метода вычисления квадратного корня входило в старую школьную программу по математике, однако из новой программы он исключен.

Алгоритм, основанный на построении рекуррентной монотонной последовательности. Данный метод коротко описан в качестве примера на монотонные последовательности в учебнике по алгебре и началам анализа для IX класса. Мы остановимся на нем подробнее.

Выберем за x_0 произвольное положительное число и рассмотрим последовательность $\{x_n\}$, определенную

с помощью рекуррентной формулы *):

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right), \qquad n = 0, 1, 2, \ldots, \tag{18}$$

где a — положительное число, из которого нужно извлечь квадратный корень. Анализ этой последовательности позволяет установить следующие свойства.

1) Члены последовательности x_n при $n \geqslant 1$ удовлетворяют неравенству

$$x_n \geqslant \sqrt{a}. \tag{19}$$

Для доказательства этого утверждения перепишем рекуррентную формулу (18) в виде

$$x_n = \sqrt{a} \cdot \frac{1}{2}\left(\frac{x_{n-1}}{\sqrt{a}} + \frac{\sqrt{a}}{x_{n-1}}\right) = \sqrt{a} \cdot \frac{1}{2}\left(t_{n-1} + \frac{1}{t_{n-1}}\right),$$

$$n = 1, 2, 3. \ldots,$$

где $t_{n-1} = x_{n-1}/\sqrt{a}$. Функция $f(t) = \frac{1}{2}(t + 1/t)$ при любом положительном значении t удовлетворяет неравенству $f(t) \geqslant 1$. Отсюда вытекает (19).

2). Члены последовательности x_n при $n \geqslant 1$ не возрастают, т. е.

$$x_{n+1} \leqslant x_n, \quad n \geqslant 1. \tag{20}$$

Разделив (18) на x_n и воспользовавшись неравенством $(a/x_n^2) \leqslant 1$, получим:

$$\frac{x_{n+1}}{x_n} = \frac{1}{2}\left(1 + \frac{a}{x_n^2}\right) \leqslant 1,$$

т. е. неравенство (20).

Итак, мы видим, что при $n \geqslant 1$ рекуррентная последовательность (18) является монотонно невозрастающей (20) и ограниченной снизу (19). Следовательно, по теореме Вейерштрасса о монотонных последовательностях она

*) Рекуррентная формула (от латинского recurrens — возвращающийся) — формула, позволяющая выразить $(n + 1)$-й член последовательности через значения ее первых n членов. В данном случае мы имеем простейшую рекуррентную формулу, в которой $(n + 1)$-й член выражается прямо через n-й. При наличии рекуррентной формулы поопоследовательность полностью определяется выбором нулевого члена x_0. Способ задания последовательностей с помощью рекуррентных формул является очень распространенным. Он прост и удобен для расчетов на ЭВМ.

имеет предел:

$$\lim_{n\to\infty} x_n = c. \tag{21}$$

Воспользовавшись этим, перейдем в формуле (18) к пределу при $n \to \infty$. В результате будем иметь

$$c = \frac{1}{2}\left(c + \frac{a}{c}\right)$$

или

$$c^2 = a, \quad c = \sqrt{a}. \tag{22}$$

Итак, мы доказали, что при выборе в качестве нулевого приближения x_0 любого положительного числа рекуррентная последовательность (18) сходится к $\sqrt{a}$:

$$\lim_{n\to\infty} x_n = \sqrt{a}, \tag{23}$$

монотонно приближаясь к своему пределу сверху.

Сходимость (23) означает, что для любой заданной точности $\varepsilon > 0$ можно указать такой номер N, что член последовательности x_N, а также все дальнейшие члены, удовлетворяют неравенству

$$0 \leqslant x_N - \sqrt{a} < \varepsilon \quad \text{или} \quad \sqrt{a} \leqslant x_N < \sqrt{a} + \varepsilon,$$

т. е. x_N определяет $c = \sqrt{a}$ с ошибкой меньше ε.

О достигнутой точности на n-м шаге можно судить по следующей оценке:

$$x_n - \sqrt{a} = \frac{x_n^2 - a}{x_n + \sqrt{a}} < \frac{x_n^2 - a}{2c_0}, \tag{24}$$

где c_0 — какое-нибудь число, удовлетворяющее неравенству $0 < c_0 \leqslant c = \sqrt{a}$. Практически за c_0 удобно взять грубое приближенное значение $\sqrt{a}$ по недостатку.

Число итераций *) N, необходимое для достижения точности ε, зависит как от требуемой точности, так и от близости нулевого приближения к искомому корню c.

Для оценки скорости сходимости метода положим $x_n = \sqrt{a} + \delta_n$ ($\delta_n = x_n - \sqrt{a} \geqslant 0$ при $n \geqslant 1$) и под-

*) Итерация (от латинского iteratio — повторение) — результат неоднократного применения какой-нибудь математической операции, в данном случае вычислений по рекуррентной формуле (18).

ставим в формулу (18):

$$\sqrt{a} + \delta_{n+1} = \frac{1}{2}\left(\sqrt{a} + \delta_n + \frac{a}{\sqrt{a} + \delta_n}\right)$$

или

$$\delta_{n+1} = \frac{\delta_n^2}{2(\sqrt{a} + \delta_n)} \leqslant \frac{\delta_n^2}{2\sqrt{a}}. \qquad (25)$$

Такая рекуррентная оценка погрешности говорит о высокой скорости сходимости процесса.

Для иллюстрации данного метода возьмем тот же пример, что и в предыдущем случае: вычислим $\sqrt{81725{,}3}$. За нулевое приближение примем $x_0 = 300$. Это — значение корня с избытком, которое получается при замене подкоренного числа на 90 000.

Нулевое приближение выбрано достаточно близким к корню: его погрешность δ_0 не превышает 15. По формуле (25) легко получить оценки погрешностей следующих приближений:

$$\delta_1 < \frac{225}{2 \cdot 250} < 0{,}5,$$

$$\delta_2 < \frac{0{,}25}{2 \cdot 250} = 0{,}0005.$$

(для упрощения оценок $\sqrt{a}$ в знаменателе формулы (25) заменен на меньшее число: 250). Мы видим, что уже второе приближение дает весьма высокую точность: она превышает точность, на которой были остановлены вычисления с помощью первого алгоритма.

После этого анализа выпишем несколько первых приближений:

$$x_1 = 286{,}208\ 833\ 334,$$
$$x_2 = 285{,}876\ 564\ 976,$$
$$x_3 = 285{,}876\ 371\ 881,$$
$$x_4 = 285{,}876\ 371\ 881,$$
$$x_5 = 285{,}876\ 371\ 881.$$

Обратите внимание на то, что после третьего шага числа x_n перестали изменяться, процесс «останавливается». В этом проявляется принципиальная особенность вычислений с конечным числом значащих цифр. Когда на третьем шаге мы достигли точности, превышающей 10^{-9}, то при расчетах с девятью знаками после запятой становится невозможно уловить разницу между x_{n+1} и x_n,

ложащую за пределами ошибки округления. Чтобы продолжить расчеты дальше и получить значение корня с более высокой степенью точности, нужно перейти к вычислениям с бóльшим числом значащих цифр.

Сравнение приведенных алгоритмов вычисления корня, бесспорно, говорит в пользу второго. Это — типичный вычислительный алгоритм, очень простой с точки зрения организации вычислительного процесса и обладающий высокой скоростью сходимости. Он лучше приспособлен для машинного счета: гораздо легче составить программу вычисления нескольких итераций по рекуррентной формуле (18) (такая программа приведена в следующей главе на стр. 56—58), чем «объяснять» ЭВМ сложное правило выбора цифр c_i в десятичном представлении искомого корня c в соответствии с первым алгоритмом.

Интересно отметить, что исторически второй метод намного старше первого. Он был описан около 2000 лет назад древнегреческим математиком Героном. Первый же метод, существенно использующий позиционную систему записи чисел, разработан только в XV веке.

§ 3. Число π и его вычисление

В этом параграфе мы обсудим круг вопросов, связанных с задачей о длине окружности и вычислении числа π.

За длину окружности принимается предел, к которому стремятся периметры правильных вписанных и описанных многоугольников при неограниченном увеличении числа их сторон. Таким образом, по определению, длина окружности связывается с бесконечным сходящимся процессом.

Доказывается, что такие пределы существуют и совпадают между собой. Тем самым устанавливается, что каждая окружность имеет длину, т. е. является спрямляемой линией (в математике известны примеры линий, не имеющих длины,—неспрямляемых линий). Доказывается также, что периметры вписанных и описанных многоугольников стремятся к своему пределу, монотонно возрастая и убывая. Таким образом, периметр любого правильного вписанного многоугольника дает для длины окружности оценку снизу, а правильного описанного — оценку сверху: длина окружности «зажата» между ними.

Устанавливается также важное свойство окружностей: отношение длины окружности к ее диаметру есть постоян-

ная величина, которая обозначается через π (обозначение введено Эйлером в первой половине XVIII века):

$$\frac{C}{2R} = \pi, \quad C = 2\pi R. \tag{26}$$

Это свойство сводит задачу определения длины любой окружности по ее радиусу R к вычислению универсальной постоянной π.

Вы помните, что π приближенно равно 3,14. Если нужно более точное значение, то его можно найти в справочнике. Однако сейчас речь пойдет не об этом. Для того чтобы π появилось в справочниках, его нужно было подсчитать. История вычисления числа π тесно связана с общим прогрессом математики и потребностями практики.

Алгоритм вычисления числа π, основанный на формуле удвоения. Исторически первый и в течение длительного времени единственный алгоритм вычисления числа π был основан на подсчете периметров правильных вписанных и описанных многоугольников с помощью формулы удвоения. Эта формула связывает длины сторон правильных n- и $2n$-угольников, вписанных в окружность радиуса R. Положим диаметр рассматриваемой окружности равным единице: $d = 2R = 1$ (длина такой окружности равна π), тогда формула удвоения примет вид

$$a_{2n} = \frac{1}{2}\sqrt{2 - 2\sqrt{1 - a_n^2}}. \tag{27}$$

Умножим и разделим выражение под знаком общего корня на сопряженное. В результате получим:

$$a_{2n} = \frac{\sqrt{4 - 4(1 - a_n^2)}}{2\sqrt{2 + 2\sqrt{1 - a_n^2}}} = \frac{a_n}{\sqrt{2 + 2\sqrt{1 - a_n^2}}}. \tag{28}$$

Введем периметр многоугольника $p_n = na_n$ и подставим выражения сторон a_n и a_{2n} через периметры в формулу (28), тогда она преобразуется к виду *)

$$p_{2n} = \frac{p_n}{\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{1 - \frac{p_n^2}{n^2}}}}. \tag{29}$$

*) Формула (29), как и формула удвоения (27), является рекуррентной формулой. С такими формулами мы уже встречались в предыдущем параграфе при описании алгоритма извлечения корня (см. подстрочное примечание на стр. 38).

Сторона правильного описанного n-угольника b_n при $d = 2R = 1$ выражается через сторону вписанного n-угольника a_n с помощью соотношения

$$b_n = a_n/\sqrt{1 - a_n^2}. \tag{30}$$

Перейдем в нем от сторон a_n и b_n к соответствующим периметрам, обозначая периметр описанного многоугольника через q_n: $q_n = nb_n$. В результате будем иметь:

$$q_n = \frac{p_n}{\sqrt{1 - p_n^2/n^2}}. \tag{31}$$

Как мы уже отмечали выше,

$$\lim_{n\to\infty} p_n = \lim_{n\to\infty} q_n = \pi. \tag{32}$$

Вычисление числа π с помощью данного метода можно начинать с какого-нибудь простого правильного многоугольника, например, с шестиугольника, для которого $p_6 = 3$, $q_6 = 3/\sqrt{1-1/4} = 2\sqrt{3} = 3{,}464101\ldots$ Далее процесс строится следующим образом: по рекуррентной формуле (29) находятся последовательно p_{12}, p_{24}, p_{48}, p_{96}, p_{192}, . . ., по ним вычисляются с помощью формулы (31) соответствующие значения q_n. Двухстороннее неравенство

$$p_n < \pi < q_n \tag{33}$$

позволяет утверждать, что найденные на некотором шаге числа дают приближенные значения π с недостатком и избытком с ошибкой $\varepsilon_n < \Delta_n = q_n - p_n$, которая стремится к нулю при $n \to \infty$.

Используя описанные идеи и проводя сложнейшие для своего времени вычисления, великий древнегреческий ученый Архимед дошел до правильного 96-угольника и получил для π двухстороннюю оценку:

$$3\frac{10}{71} < \pi < 3\frac{1}{7},$$

$$\left(3\frac{10}{71} = 3{,}14084\ldots,\ 3\frac{1}{7} = 3{,}14285\ldots,\ \Delta \approx 0{,}002\right).$$

Чтобы правильно понять результат Архимеда, нужно иметь в виду, что в то время не было привычной для нас позиционной записи чисел, десятичных дробей и хорошо разработанной техники извлечения квадратных корней,

хотя такую операцию Архимеду приходилось выполнять дважды на каждом шаге.

Мы не имеем возможности подробно останавливаться на очень интересной истории вычисления π и отметим только еще один результат. В первой половине XV века в обсерватории узбекского хана Улугбека под Самаркандом его придворный астроном Аль-Каши вычислил π с 17 знаками после запятой. Он сделал 27 удвоений числа сторон и дошел до $3 \cdot 2^{28}$-угольника. Это был уникальный для своего времени расчет. Вычисления Аль-Каши были связаны с составлением таблицы синусов с шагом в $1'$, нужной для астрономических наблюдений. Для сравнения укажем, что в Европе французский математик Ф. Виет (вы знаете его теорему о корнях квадратного уравнения) спустя 150 лет (в 1593 году) нашел лишь 9 правильных цифр числа π после запятой с помощью $3 \cdot 2^{17}$-угольника (16 удвоений числа сторон). Только на рубеже XVI и XVII веков, т. е. спустя 250 лет, результат Аль-Каши был повторен, а затем и превзойден.

Заканчивая обсуждение этих вопросов, приведем табл. 1. В ней даны периметры вписанных и описанных многоугольников, которые получаются из правильного шестиугольника в результате 16 удвоений числа сторон (повторение результата Ф. Виета).

ТАБЛИЦА 1

k	$n = 6 \cdot 2^k$	p_n	q_n
0	6	3, 000 000 000 000	3, 464 101 615 138
1	12	3, 105 828 541 230	3, 630 002 002 236
2	24	3, 132 628 613 281	3, 245 155 564 194
3	48	3, 139 350 203 047	3, 166 557 423 678
4	96	3, 141 031 950 890	3, 147 778 817 495
5	192	3, 141 452 472 285	3, 143 135 797 312
6	384	3, 141 557 607 912	3, 141 978 227 840
7	768	3, 141 583 892 148	3, 141 689 033 932
8	1 536	3, 141 590 463 228	3, 141 616 747 849
9	3 072	3, 141 592 105 999	3, 141 598 677 103
10	6 144	3, 141 592 516 692	3, 141 594 159 465
11	12 288	3, 141 592 619 365	3, 141 593 030 058
12	24 576	3, 141 592 645 034	3, 141 592 747 706
13	49 152	3, 141 592 651 034	3, 141 592 677 119
14	98 304	3, 141 592 653 055	3, 141 592 659 472
15	196 608	3, 141 592 653 456	3, 141 592 655 060
16	393 216	3, 141 592 653 556	3, 141 592 653 957

Для оценки точности определения π с помощью этих расчетов составим разность периметров q_n и p_n, приведенных в последней строке ($n = 6 \cdot 2^{16} = 393\,216$):

$$\varepsilon_n < \Delta_n = q_n - p_n = 0,000\,000\,000\,401.$$

Первые 10 знаков у чисел p_n и q_n совпадают, они дают первые 10 знаков числа π:

$$\pi = 3,141\,592\,653... \qquad (34)$$

Алгоритмы вычисления π, основанные на разложении в ряд арктангенса. После создания в XVII веке математического анализа было установлено много формул, содержащих π, которые могут быть использованы для его вычисления. Рассмотрим некоторые из них, связанные с разложением в ряд арктангенса.

Возьмем функцию $f(t) = 1/(1 + t^2)$. При $|t| < 1$ это выражение можно рассматривать как сумму бесконечной геометрической прогрессии со знаменателем $q = -t^2$, $|q| < 1$:

$$\frac{1}{1+t^2} = 1 - t^2 + t^4 - t^6 + \ldots \qquad (35)$$

Проинтегрируем данное равенство от 0 до x, заменяя в правой части интеграл от бесконечной суммы бесконечной суммой интегралов:

$$\int_0^x \frac{1}{1+t^2}\,dt = \int_0^x dt - \int_0^x t^2\,dt + \int_0^x t^4\,dt - \int_0^x t^6\,dt + \ldots$$

В результате будем иметь:

$$\operatorname{arctg} x = x - \frac{1}{3}x^3 + \frac{1}{5}x^5 - \frac{1}{7}x^7 + \ldots \qquad (36)$$

Мы не можем останавливаться на обосновании законности такой процедуры почленного интегрирования и отметим только, что равенство (36) справедливо при $|x| \leqslant 1$.

При $x = 1$ формула (36) принимает вид

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots$$

или

$$\pi = 4\left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots\right). \qquad (37)$$

Для вычисления суммы ряда (37) нужно составить последовательность его частичных сумм:

$$S_1 = 4 \cdot 1,$$
$$S_2 = 4\left(1 - \frac{1}{3}\right),$$
$$S_3 = 4\left(1 - \frac{1}{3} + \frac{1}{5}\right),$$
$$S_4 = 4\left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7}\right),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots,$$
$$S_n = 4\left(1 - \frac{1}{3} + \frac{1}{5} - \ldots + \frac{(-1)^{n-1}}{2n-1}\right),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots,$$

которые удобно находить по рекуррентной формуле:

$$S_{n+1} = S_n + \frac{(-1)^n}{2n+1}, \qquad n = 1, 2, 3, \ldots \tag{38}$$

Согласно (37) предел последовательности $\{S_n\}$ существует и равен π:

$$\pi = \lim_{n\to\infty} S_n. \tag{39}$$

Рассмотрим частичные суммы с четными номерами. Для них

$$S_{2n+2} = S_{2n} + \frac{4}{4n+1} - \frac{4}{4n+3} > S_{2n},$$

т. е. эта последовательность монотонно возрастает и приближается к своему пределу снизу: $S_{2n} < \pi$.

С другой стороны, для последовательности частичных сумм с нечетными номерами получаем:

$$S_{2n+1} = S_{2n-1} - \frac{4}{4n-1} + \frac{4}{4n+1} < S_{2n-1},$$

т. е. эта последовательность монотонно убывает и приближается к своему пределу сверху: $S_{2n+1} > \pi$.

Таким образом, при любом $n \geqslant 1$ имеем:

$$S_{2n} < \pi < S_{2n+1},$$

причем

$$S_{2n+1} - S_{2n} = \frac{4}{4n+1} < \frac{1}{n}.$$

Мы видим, что сумма S_{2n} определяет число π с недостатком, а сумма S_{2n+1} — с избытком с ошибкой, не превышающей $1/n$.

Мы привели и подробно обсудили формулу (37) из-за ее исключительной простоты. Однако практически для вычисления π она малопригодна, потому что процесс сходится слишком медленно. Действительно, чтобы подсчитать π по формуле (37) с тремя знаками после запятой (точность результата Архимеда $\pi \approx 3\frac{10}{71}$), нужно просуммировать 2000 слагаемых. Однако с помощью ряда (36) для арктангенса можно получить другие более эффективные, хотя и менее простые формулы для π. К одной из них мы придем, полагая в соотношении (36) $x = 1/\sqrt{3}$:

$$\frac{\pi}{6} = \frac{1}{\sqrt{3}}\left(1 - \frac{1}{3}\cdot\frac{1}{3} + \frac{1}{5}\cdot\left(\frac{1}{3}\right)^2 - \frac{1}{7}\cdot\left(\frac{1}{3}\right)^3 + \ldots\right)$$

или

$$\pi = 2\sqrt{3}\left(1 - \frac{1}{3}\cdot\frac{1}{3} + \frac{1}{5}\cdot\left(\frac{1}{3}\right)^2 - \frac{1}{7}\cdot\left(\frac{1}{3}\right)^3 + \ldots\right). \tag{40}$$

Частичные суммы этого ряда вычисляются по рекуррентной формуле:

$$S_{n+1} = S_n + \frac{2\sqrt{3}}{2n+1}\left(-\frac{1}{3}\right)^n. \tag{41}$$

Как и в предыдущем случае, частичные суммы с четными номерами монотонно возрастают и приближаются к пределу, равному π, снизу, а частичные суммы с нечетными номерами монотонно убывают и приближаются к π сверху:

$$S_{2n} < \pi < S_{2n+1}, \tag{42}$$

причем

$$\Delta_{2n} = S_{2n+1} - S_{2n} = \frac{2\sqrt{3}}{4n+1}\,\frac{1}{3^{2n}}. \tag{43}$$

Теперь ошибка стремится к нулю быстрее членов геометрической прогрессии со знаменателем $q = 1/3$, т. е. скорость сходимости намного выше, чем для ряда (37).

Результаты вычисления π с помощью ряда (40) приведены в табл. 2. Как и в предыдущем случае, мы продолжали расчеты до тех пор, пока не нашли 10 первых знаков π. Для этого нам пришлось подсчитать частичные суммы вплоть до S_{18} и S_{19} (9-я строка в таблице). Полученный результат: $\pi = 3,\ 141\ 592\ 653\ldots$, естественно, совпадает с (34).

ТАБЛИЦА 2

n	S_{2n}	S_{2n+1}
0		3, 464 101 615 138
1	3, 079 201 435 678	3, 156 181 471 570
2	3, 137 852 891 596	3, 142 604 745 663
3	3, 141 308 785 463	3, 141 674 312 699
4	3, 141 568 715 942	3, 141 599 773 811
5	3, 141 590 510 938	3, 141 593 304 503
6	3, 141 592 454 288	3, 141 592 715 020
7	3, 141 592 634 547	3, 141 592 659 522
8	3, 141 592 651 734	3, 141 592 654 173
9	3, 141 592 653 406	3, 141 592 653 648

Мы не будем останавливаться на других методах вычисления. Отметим, лишь, что к концу XIX века английский математик Вильям Шенкс вычислил 707 знаков числа π, потратив на это более 20 лет. Этот результат по праву получил славу рекорда вычислений XIX века. Однако в 1945 году было обнаружено, что Шенкс допустил ошибку в 520-м знаке, и все его дальнейшие вычисления пошли насмарку.

В настоящее время с помощью ЭВМ число π вычислено с фантастической точностью — более пятисот тысяч знаков. Продолжительность расчетов подобного типа зависит от используемого алгоритма и быстродействия машины. Для современных ЭВМ она измеряется несколькими часами.

Сравните, с одной стороны, 707 знаков и 20 лет, с другой — пятьсот тысяч знаков и несколько часов. Добавьте сюда большую вероятность ошибки при ручном счете, которая практически исключена при расчете на ЭВМ, и комментарии будут излишними. Этими замечаниями мы и закончим главу.

Глава 3
ЭЛЕКТРОННО-ВЫЧИСЛИТЕЛЬНЫЕ МАШИНЫ

§ 1. От 10 пальцев к ЭВМ

Применение математических методов для решения практических задач неизбежно связано с проведением расчетов, с доведением ответа до «числа». Без хорошо разработанной техники вычислений стройное здание математики превратится в дом без окон и дверей, изолированный от внешнего мира. Поэтому проблема упрощения и ускорения вычислений всегда имела первостепенное значение. Один из путей ее решения был связан с усовершенствованием методов счета. Самым важным результатом здесь явился переход от аддитивной *) (например, римской) системы записи чисел к принятой в настоящее время позиционной системе.

В римском счислении MCXXXVI означает пятьсот + + сто + десять + десять + десять + пять + один. В ней с увеличением изображаемых чисел нужно неограниченно увеличивать число используемых символов. Однако главный недостаток аддитивной системы заключается в сложности выполнения арифметических операций при такой форме записи чисел. Это «искусство» требовало специальной профессиональной подготовки, которую в то время могли получить лишь немногие.

Позиционная система обладает по сравнению с аддитивной двумя существенными преимуществами.

Во-первых, в ней любое число записывается с помощью небольшого числа символов (в общепринятой десятичной системе ими являются десять арабских цифр: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9; в двоичной системе, которая используется в большинстве вычислительных машин, такими символами являются 0 и 1.) В десятичной системе число MCXXXVI

*) Аддитивный (от латинского additivus — придаточный) — полученный путем сложения.

представляется в виде

$$636 = 6 \cdot 10^2 + 3 \cdot 10 + 6,$$

при этом один и тот же символ, цифра 6, в зависимости от своего положения имеет разное значение: в одном случае он означает число сотен, в другом — число единиц.

Во-вторых, для позиционной системы правила выполнения арифметических операций намного проще, чем для аддитивной. Они сводятся к запоминанию таблиц сложения и умножения однозначных чисел.

Важное значение для упрощения вычислений сыграло также изобретение логарифмов. Вот что писал по этому поводу французский математик и астроном П. С. Лаплас, которому приходилось проводить большие расчеты в связи с астрономическими исследованиями: «Изобретение логарифмов, сокращая вычисления нескольких месяцев в труд нескольких дней, словно удваивает жизнь астронома».

Наряду с упрощением методов счета постоянно велись поиски путей ускорения процесса вычислений. 10 пальцев на руках человека были самым первым «техническим средством» в этой области, оставившим нам десятичную систему счисления и поговорку «пересчитать по пальцам».

Все вы видели обычные канцелярские счеты, которые до недавнего времени были обязательным атрибутом на столе у любого бухгалтера и кассира. Их история насчитывает несколько веков.

Очень удобна для проведения простых вычислений, не требующих высокой точности, логарифмическая линейка. Ею широко пользовалось не одно поколение инженеров и научных работников.

В XVII в. начали появляться механические вычислительные машины, главная особенность которых состояла в том, что алгоритмы выполнения арифметических операций закладывались в само устройство. Их конструкция непрерывно совершенствовалась. В первой половине XX в. для проведения вычислений с многоразрядными числами широко использовались созданные по этому принципу ручные арифмометры и настольные клавишные машины, которые приводились в действие электромотором. При работе на таких машинах вычислителю было нужно набирать исходные данные на клавиатуре, записывать полученные результаты в специально приготовленную таблицу и следить за порядком выполнения арифме-

тических операций, который определялся алгоритмом решения задачи. Сами же операции выполнялись машиной. Машина в несколько раз увеличивала скорость счета и существенно снижала нервное напряжение работника по сравнению со счетом вручную. Грубо это можно сравнить с ездой на лошади: и быстрее, чем пешком, и не так утомительно.

Однако развитие науки и техники в первой половине XX в., особенно в тридцатые — сороковые годы, существенно расширило круг прикладных задач, поставленных перед математикой. Их сложность не позволяла, как правило, получить ответ в явном виде и требовала применения численных методов. Для проведения расчетов стали создаваться специальные группы вычислителей. Объем вычислений и потребность в вычислителях резко нарастали. Стало ясно, что идти дальше по экстенсивному *) пути нельзя. Необходим был принципиально новый шаг, и этот шаг был сделан: появились ЭВМ.

Создание ЭВМ можно сравнить с самыми выдающимися достижениями — такими, как изобретение колеса, освоение выплавки металлов, создание паровой машины, освоение электричества, использование атомной энергии. Хорошо известна роль так называемых «великих открытий» в судьбе человечества: за сравнительно короткий срок они существенно изменяли производительные силы общества, оказывая большое влияние на условия его жизни.

Однако в этом ряду ЭВМ занимает особое место: если обычные машины расширяли физические возможности людей, то ЭВМ существенно повысили их интеллектуальный потенциал. Они способствовали развитию новых эффективных методов познания и использования законов природы, явились одним из наиболее важных факторов научно-технического прогресса последних десятилетий.

Мы уже говорили во введении, что за тридцать с небольшим лет своего существования ЭВМ увеличили скорость проведения вычислений в 10^8 раз. Человеческое воображение плохо воспринимает такие большие числа. Представьте себе, что все 250-миллионное население Советского Союза, включая грудных младенцев, превратилось в опытных вычислителей, вооружилось кла-

*) Экстенсивный (от латинского extensivus — расширяющий, удлиняющий) — в противоположность интенсивному означает не качественное, а лишь количественное изменение.

вишными машинами и приступило к расчетам. Тогда их суммарная производительность труда будет примерно эквивалентна трем большим ЭВМ. Если же мы учтем 8-часовой рабочий день человека и 24-часовой — машины, то для замены всего населения СССР в качестве вычислителей будет достаточно одной ЭВМ.

Эта глава посвящена краткому описанию принципов работы и возможностей использования ЭВМ.

§ 2. Как работают ЭВМ

Для того чтобы ответить на этот вопрос, обсудим сначала, как работали вычислители, вооруженные клавишными машинами типа арифмометров, до появления ЭВМ.

Пусть нужно вычислить $\sqrt{a} = \sqrt{81\,725,\,3}$ (пример взят из § 2 главы 2). Вычислитель решает воспользоваться методом, основанным на рекуррентной формуле (18) стр. 38, выбрав за нулевое приближение $x_0 = 300$. После этого он заготовляет следующую таблицу.

ТАБЛИЦА 1

n	a/x_{n-1}	$x_{n-1} + a/x_{n-1}$	$x_n = \frac{1}{2}\left(x_{n-1} + a/x_{n-1}\right)$
0			300
1			
2			
3			

Выполняя расчеты, вычислитель заполняет в таблице строчку за строчкой, пока не получит результата требуемой точности (об этом можно судить либо по неравенству (24) на стр. 39, либо просто по установлению нужного числа десятичных знаков у последовательности x_n).

Приведем теперь таблицу в заполненном виде после завершения вычислений. (В отличие от данных на стр. 40 главы 2, здесь результаты содержат не 12, а только 8 значащих цифр в соответствии с разрядностью клавишных машин.)

Клавишные машины — это механические устройства. Считали они медленно: на выполнение одного действия уходило несколько секунд. Однако самым медленным звеном в вычислительном процессе была не машина, а чело-

ТАБЛИЦА 2

n	a/x_{n-1}	$x_{n-1}+a/x_{n-1}$	$x_n=\frac{1}{2}(x_{n-1}+a/x_{n-1})$.
0			300
1	272, 417 66	572, 417 66	286, 208 83
2	285, 544 29	571, 753 12	285, 876 56
3	285, 876 18	571, 752 74	285, 876 37
4	285, 876 37	571, 752 74	285, 876 37

век. Основное время уходило не на счет, а на набор с помощью клавиш данных для очередной операции и на запись результатов в таблицу.

Принципиально новый шаг, который был сделан при создании ЭВМ, состоял в полной автоматизации вычислений, в их проведении без участия человека. Для этого нужно заранее описать в «понятной» для машины форме всю последовательность арифметических операций, необходимую для решения задачи (такое описание называется программой), и задать исходные данные.

Чтобы программу и исходные данные можно было ввести в машину, выполнить необходимые вычисления и вывести полученные результаты, любая ЭВМ, независимо от ее конкретных конструктивных особенностей, должна иметь следующие узлы:

1) Устройство управления (УУ): управляет порядком выполнения операций, координирует работу всех узлов машины.

2) Арифметическое устройство (АУ): служит для выполнения арифметических и логических операций. (О логических операциях будет рассказано ниже.)

3) Запоминающее устройство (ЗУ) или просто «память»: предназначено для хранения программы, исходных данных и результатов вычислений. Вся «память» разбита на ячейки или «машинные слова». Ячейки перенумерованы, номер ячейки называется ее адресом. В каждой ячейке может храниться одно число, которое является либо настоящим числом, либо командой программы. (Команды записываются в виде чисел, которые определяют, откуда нужно взять данные, какую операцию над ними выполнить и куда поставить результат.)

Ячейки состоят из элементов, в них помещаются отдельные разряды хранимого числа. Количество разрядов

в ячейке (т. е. «длина» машинного слова) определяется конструктивными особенностями машины и одинаково для всех ее ячеек.

Мы уже говорили, что обычно в ЭВМ используется двоичная система счисления. В соответствии с этим элементы ячеек памяти могут находиться в одном из двух состояний (есть ток — нет тока, намагничено — не намагничено и т. д.). Одно из них интерпретируется как цифра 0, второе — как цифра 1. Проанализировав состояние всех элементов ячейки, можно прочитать помещенное в ней число. Меняя состояние элементов, мы заменим одно число другим.

Различают оперативную (внутреннюю) и внешнюю памяти. Разница между ними состоит в том, что УУ и АУ могут непосредственно получать и обрабатывать информацию, находящуюся только в оперативной памяти. Для работы с информацией, хранящейся во внешней памяти, ее нужно сначала поместить в оперативную память. Это осуществляется с помощью специальных команд.

Внешняя память современных ЭВМ состоит из магнитных лент, дисков, барабанов. Объем этих носителей в тысячи, десятки тысяч раз больше объема оперативной памяти. Однако для того, чтобы получить доступ к нужному элементу данных, расположенному во внешней памяти, его предварительно надо переслать в оперативную память, а эта операция занимает намного больше времени, чем прямое обращение в оперативную память.

4) Устройство ввода-вывода (УВВ): предназначено для ввода в машину программы, исходных данных и вывода результатов счета, т. е. для обмена информацией между человеком и машиной.

Ввод программы и исходных данных в ЭВМ обычно осуществляется с помощью перфокарт. Перфокарты — это специальные картонные карты стандартных размеров и формы. Нужная информация наносится (перфорируется *)) на них в виде набора отверстий (см. рис. 9). Перфорация осуществляется заранее с помощью специального устройства — перфоратора (см. рис. 10), не связанного непосредственно с ЭВМ. Подготовленные перфокарты с отперфорированной на них программой и исходными данными ставятся в читающее устройство ЭВМ, которое про-

*) Перфорировать (от латинского perforare — буравить) — пробивать отверстия, просверливать.

сматривает их одну за другой, прочитывает закодированную на них информацию и записывает ее в оперативную память. После того как программа и исходные данные введены, машина может приступать к вычислениям.

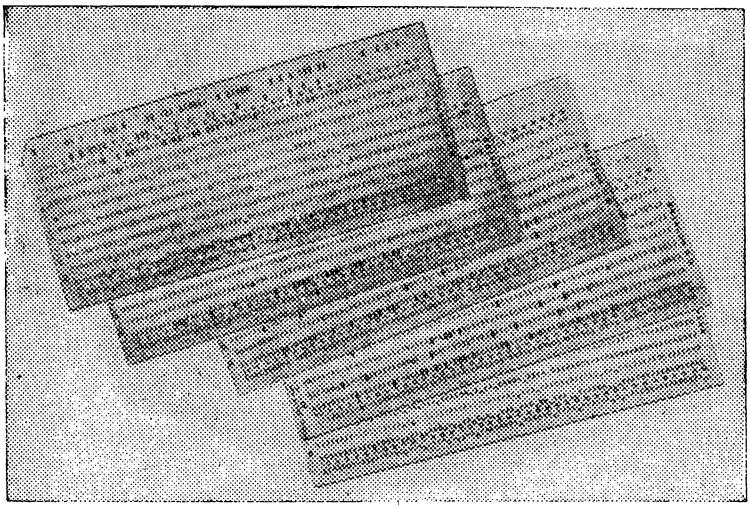

Рис. 9. Перфокарты.

Результаты расчетов, предписанные программой, печатаются на бумаге специальным печатающим устройством (см. рис. 11).

Отметим, что числа вводятся в ЭВМ и выводятся из нее на печать в привычной десятичной форме. Машина сама переводит их при вводе из десятичной системы в двоичную, а при выдаче на печать — из двоичной в десятичную. Пользователь ЭВМ может вообще ничего не знать ни о двоичной системе счисления, ни о том, что она используется для проведения вычислений в ЭВМ. Это никак не отразится на его работе.

Устройство управления, арифметическое устройство и оперативная память объединены в общий комплекс, который называется центральным процессором, остальные — внешними устройствами или периферийным оборудованием.

Теперь, когда мы познакомились с основными элементами ЭВМ и их назначением, приведем в качестве примера

программу вычисления квадратного корня из положительного числа a с заданной точностью ε. При этом, как и в § 2 главы 2, будем обозначать через x_0 нулевое приближение (за него можно принять, как мы знаем, любое

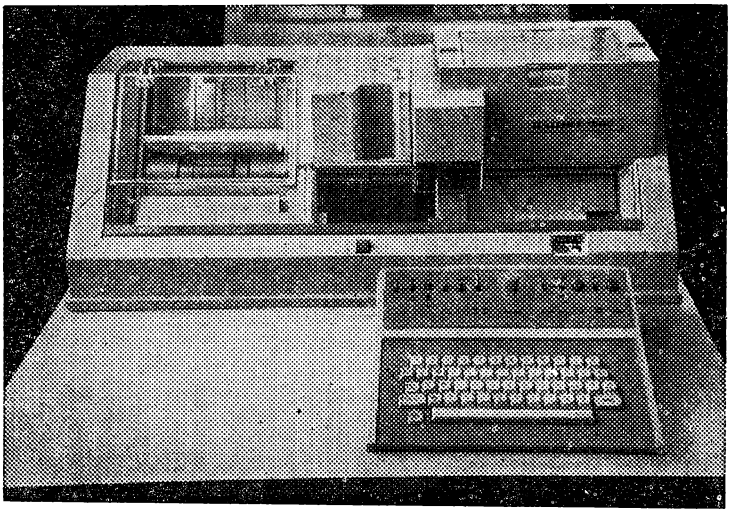

Рис. 10. Перфоратор — устройство, с помощью которого нужная информация наносится в виде отверстий на перфокарты.

положительное число) и через c_0 — какое-нибудь число, удовлетворяющее условию $0 < c_0 \leqslant \sqrt{a}$. Оно нужно для оценки достигнутой точности с помощью неравенства (24) на стр. 39.

Программа вычисления $\sqrt{a}$

1) Введи числа a, ε, x_0, c_0 и переходи к следующей команде.

2) Присвой величине x значение x_0 и переходи к следующей команде.

3) Присвой величине y значение a/x и переходи к следующей команде.

4) Присвой величине y значение $x + y$ и переходи к следующей команде.

5) Присвой величине x значение $(1/2)\,y$ и переходи к следующей команде.

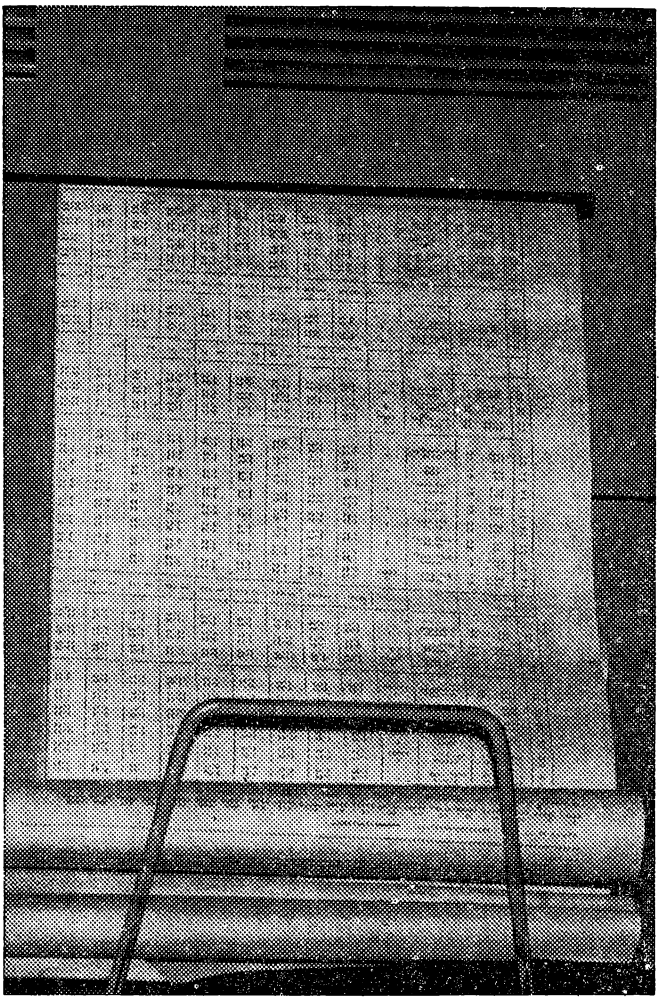

Рис. 11. Результаты расчетов, которые алфавитно-цифровое печатающее устройство (АЦПУ) выдает в виде готовой таблицы со словесным описанием приведенных величин.

6) Присвой величине y значение x^2 и переходи к следующей команде.

7) Присвой величине y значение $y - a$ и переходи к следующей команде.

8) Присвой величине y значение y/c_0 и переходи к следующей команде.

9) Присвой величине δ значение $(1/2)\,y$ и переходи к следующей команде.

10) Сравни δ и ε. Если $\delta > \varepsilon$, то переходи к команде 3), иначе переходи к следующей команде.

11) Напечатай числа x, a, ε.

12) Прекрати вычисления.

Поясним составленную программу. Команда 2) помещает значение начального приближения x_0 в ячейку памяти, в которой хранятся значения переменной величины x (на каждом этапе вычислений в этой ячейке хранится значение x, равное значению одного из членов рекуррентной последовательности x_n).

Команды 3), 4), 5) вычисляют по числу x число $(x + a/x)/2$, т. е. делают очередную итерацию. Найденное число помещается в ячейку памяти, в которой хранится значение величин x, при этом старое значение переменной величины x теряется безвозвратно. Это значит, что если бы нам понадобилось старое значение переменной величины x, то для его восстановления нам бы пришлось провести все вычисления заново.

Команды 6), 7), 8), 9) вычисляют величину δ:

$$\delta = \frac{x^2 - a}{2c_0},$$

с помощью которой оценивается сверху разность $x - \sqrt{a}$ (см. неравенство 24 на стр. 39 главы 2).

Важное значение имеет команда 10). По этой команде не производится вычислений, а сравниваются между собой два числа: δ и заданная точность ε. По результату сравнения УУ «принимает решение», что делать дальше. Если $\delta > \varepsilon$, то УУ вернет вычислительный процесс к команде 3) и заставит делать еще одну итерацию. В противном случае, когда требуемая точность достигнута, машина напечатает полученный результат и прекратит вычисления по данной программе.

Заканчивая комментарии к программе, обратим ваше внимание на переменную y. Ей мы присваиваем результаты промежуточных вычислений, которые нужны только

для выполнения следующей операции. Мы их не храним долго в памяти машины, а, использовав, тут же «затираем» результатом следующих вычислений. Такой подход позволяет экономить память машины, не загромождая ее ненужной информацией. Для данной задачи это не важно: она слишком проста, но для больших задач экономное распределение памяти машины имеет принципиальное значение, иначе задача в ней может не поместиться.

Составленная программа элементарна, однако она дает возможность обсудить ряд вопросов, связанных с процессом составления программы (программированием) и работой самой ЭВМ.

1. Выше мы говорили, что программа должна определять последовательность вычислений в «понятной» для машины форме. Данная программа не удовлетворяет этому требованию: она написана в форме, понятной человеку, а не машине. Каждая машина имеет свой язык. Он определяется набором операций («системой команд»), которые может выполнять данная машина, и формой их записи, т. е. способом кодировки. Для каждой ЭВМ система команд и способ кодировки свои. Чтобы программировать для конкретной машины, нужно знать ее язык. Программа, написанная для одной машины, будет непонятна другой машине, если их языки, т. е. система команд и форма их представления, различны.

2. Команды типа 10) называются логическими или командами передачи управления. Они не выполняют никаких арифметических операций, а только анализируют результаты выполнения предыдущих команд и в зависимости от них определяют, как вести вычисления дальше. Логические команды играют очень важную роль в организации вычислений на ЭВМ. Благодаря им мы имеем возможность оборвать бесконечный сходящийся процесс после достижения результата заданной точности, как это делает команда 10) в нашей программе.

Логические команды необходимы для решения задач, для которых алгоритм не может предопределить заранее однозначно весь ход вычислений: он допускает разветвления, и выбор нужной ветви приходится делать в процессе решения в зависимости от результатов расчетов. Таких задач очень много, они есть даже в школьной математике.

Возьмем в качестве примера квадратное уравнение

$$ax^2 + bx + c = 0.$$

Его свойства зависят от знака дискриминанта $\delta = b^2 - 4ac$. При $\delta > 0$ существуют два действительных корня, которые находятся по известной формуле; при $\delta = 0$ корень один: $x = -b/(2a)$; наконец, при $\delta < 0$ формула «не работает»: под знаком квадратного корня стоит отрицательное число, уравнение не имеет действительных решений. Выбрать нужный вариант действия можно только, вычислив δ, т. е. в ходе решения задачи, а не заранее.

Второй пример — система двух линейных уравнений с двумя неизвестными

$$a_1x + b_1y = c_1,$$
$$a_2x + b_2y = c_2,$$

свойства которой определяются равенством или неравенством нулю определителя системы $\Delta = a_1b_2 - a_2b_1$. Только после вычисления Δ и анализа результата ($\Delta = 0$ или $\Delta \neq 0$) можно указать ход дальнейших расчетов.

Если даже в простейших школьных задачах мы встречаемся с разветвленными алгоритмами, так что же говорить о реальных больших задачах, решение которых проводится с помощью ЭВМ. Без логических команд выбрать нужную ветвь алгоритма нельзя.

3. Благодаря тому, что алгоритм вычисления сводится к многократному применению рекуррентной формулы, число команд в программе намного меньше числа фактически выполняемых операций. Формула описана в программе только один раз. Это очень важное обстоятельство. Если бы нам пришлось писать программы, в которых число команд равно числу выполняемых операций, то применение ЭВМ оказалось бы практически бессмысленным: время, затраченное на создание и отладку программы, превышало бы время решения задачи вручную. Рекуррентные формулы, часто встречающиеся в алгоритмах решения различных задач, очень удобны для расчетов на ЭВМ.

4. Команды программы хранятся в памяти машины как обычные числа. Это дает возможность выполнять над ними различные операции, т. е. в ходе решения задачи программа может сама себя перестраивать: изменять код операций, адреса ячеек, из которых берутся или в которые засылаются данные, убирать одни и добавлять другие команды, менять команды местами и т. д. Все это с учетом возможностей логических команд делает процедуру счета

на ЭВМ очень гибкой. Не случайно в 50-х годах при появлении первых ЭВМ их называли вычислительными машинами с гибким программным управлением. Такое название правильно передает наиболее существенную особенность этих замечательных устройств.

§ 3. Поколения ЭВМ и проблемы общения человека и машины

Во введении и первых параграфах этой главы мы подчеркивали, что появление ЭВМ — не дело случая, а результат настойчивого поиска, стимулированного развитием науки и техники, потребностями практики.

Поиск средств автоматизации вычислений велся постоянно, так что многие идеи и принципы, использованные в современных ЭВМ, были высказаны и частично реализованы раньше в других вычислительных устройствах. Поэтому прежде чем обсуждать эволюцию ЭВМ, пути их совершенствования и перспективы дальнейшего развития, полезно, хотя бы очень коротко, рассказать предысторию. Идея создания вычислительных машин с программным управлением была высказана Чарльзом Бэббиджем (1791—1871 гг.), профессором математики Кембриджского университета. Бэббидж занимался составлением навигационных таблиц, имевших важное значение для такой морской державы, как Англия. Внимательное изучение процедур вычислений привело его к мысли о возможности их автоматизации. Ч. Бэббидж разработал проект механического вычислительного автомата, названного им аналитической машиной, у которого были все основные узлы ЭВМ, включая принцип управления с помощью запоминаемой программы. Этот грандиозный проект, опередивший свое время, остался нереализованным. Бэббиджем была создана только небольшая «разностная машина» со сравнительно простой и жестко заданной программой (1812 год). Труды Бэббиджа были опубликованы в 1888 году после его смерти, и о них забыли. По достоинству его идеи были оценены значительно позднее.

В 80-х годах прошлого века Герман Холлериз (США) для проведения переписи населения 1890 года сконструировал машину, автоматизировавшую процесс обработки данных, и использовал в качестве носителей информации перфокарты. В 1896 году он основал фирму по выпуску перфокарт и счетно-перфорационных машин. В дальней-

шем она была преобразована в фирму IBM (International Business Mashines), которая является в настоящее время крупнейшим производителем ЭВМ.

В 1937 году профессор университета штата Айова Атанасов с группой сотрудников начал работу по созданию электронно-вычислительной машины. Принципиально новая идея Атанасова состояла в том, что вычислитель (арифметическое устройство) работал в двоичной системе. Были созданы специальные электромеханические блоки для перевода чисел из десятичной системы в двоичную и наоборот. Вторая мировая война не дала довести работу над проектом до конца.

Параллельно с Атанасовым работу над вычислительным автоматом вел в Гарвардском университете Айткен. В 1937 году им был предложен проект релейной электромеханической машины. Машина была построена фирмой IBM в 1944 году и названа Марк-1. Она еще не имела гибко изменяющейся программы и по современным представлениям была очень медленной (операция умножения выполнялась за 3 с). Однако возможность создания автомата, состоящего из многих тысяч логических элементов, была доказана.

Примерно в то же время в Германии была создана вычислительная электромеханическая машина с программным управлением (машина Цузе).

Первая электронная вычислительная машина ENIAC построена в 1945 году Эккертом и Моучли в Пенсильванском университете. Это была ламповая машина на электронных реле. В отличие от Марк-1, ее можно назвать машиной с автоматическим программным управлением, хотя она еще не имела внутренней памяти.

В 1949 году в Англии была построена машина EDSAC, которая уже обладала всеми необходимыми компонентами современных ЭВМ.

В 1947 году были начаты, а в 1951 году завершены работы над первой советской ЭВМ, названной МЭСМ. Руководил этой работой академик С. А. Лебедев.

Наступила эра ЭВМ. В 1952—1953 годах число ЭВМ в мире исчислялось десятками, в 1965 году оно достигло 40 тысяч, в 1970 году — более 100 тысяч, сейчас — более 0,5 млн.

Параллельно с количественным ростом числа ЭВМ не менее быстро шел процесс их совершенствования. В зависимости от элементной базы, технических характерис-

тик (быстродействия, емкости памяти), способов управления и обработки информации принято делить ЭВМ на поколения, число которых к настоящему времени достигло четырех. (Надо сказать, что это деление достаточно условно, и некоторые авторы насчитывают не четыре, а шесть поколений ЭВМ.) Различие в технических возможностях машин разного поколения накладывало свой отпечаток на стиль, особенности общения человека с машиной.

Машины первого поколения. Первое поколение ЭВМ — это машины, построенные на радиолампах, с быстродействием порядка 10^2—10^4 операций в секунду, оперативной памятью порядка 2000—4000 слов и бедным набором средств ввода-вывода. Из серийных советских ЭВМ к этому поколению относятся машины «Стрела» (1953 г.), БЭСМ-2 (1959 г.), М-20 (1959 г.), «Минск-1» (1960 г.) и другие.

Приведем в качестве примера характеристики одной из наиболее производительных машин первого поколения — машины М-20: быстродействие — 20 000 операций в секунду, длина слов — 45 двоичных разрядов, объем оперативной памяти — 4К слов, где $К = 2^{10} = 1024$ — «двоичная тысяча».

Машина М-20 имела также внешнюю память, которая состояла из промежуточной памяти в виде 3-х магнитных барабанов по 4К слов каждый и 4-х магнитофонов с емкостью магнитных лент по 75К слов. Скорость обмена информацией оперативной памяти с барабанами (промежуточная память) была сравнительно высокой, с магнитными лентами — медленной, требовала несколько десятков секунд.

Быстродействие машины М-20 превышало скорость работы вычислителя с настольной клавишной машиной, равную примерно 2 операции в минуту, в 600 000 раз. Чтобы получить представление об объеме ее памяти, воспользуемся следующими соображениями. В русском алфавите $32 = 2^5$ букв. Все буквы можно закодировать с помощью 5-разрядных двоичных чисел. В одном 45-разрядном машинном слове можно разместить 9 букв, а вся оперативная память вмещает $9 \cdot 4096 = 36\,864$ буквы или 18 с лишним страниц текста, содержащего 2000 букв на странице (40 строк по 50 букв). Если дополнительно принять во внимание емкость 3-х магнитных барабанов, то это число нужно учетверить.

При работе на машинах первого поколения программист писал программу непосредственно на языке

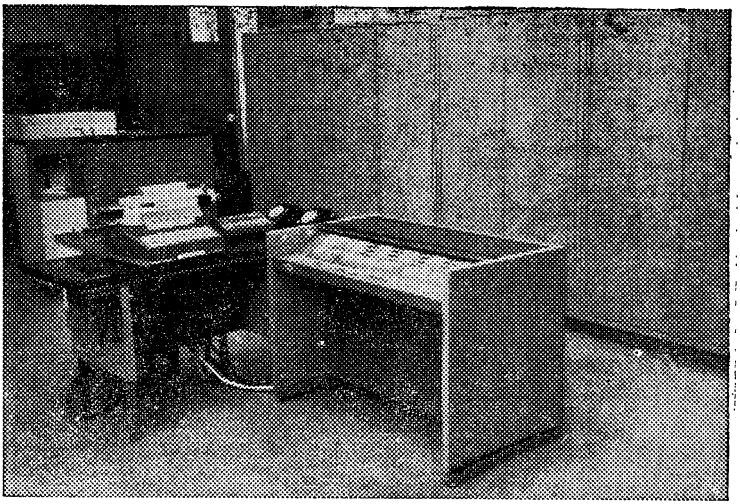

Рис. 12. ЭВМ «Минск-32».

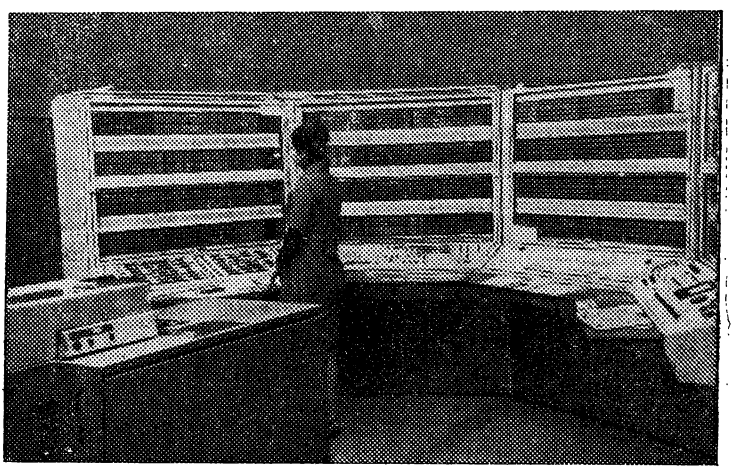

Рис. 13. ЭВМ БЭСМ-6. На переднем плане в левом нижнем углу — читающее устройство, предназначенное для считывания информации с перфокарт и ввода ее в оперативную память машины.

машины, самостоятельно распределяя ячейки оперативной памяти под программу, исходные данные, результаты счета. Разобраться в чужой программе без подробных пояснений автора о структуре программы и распределении памяти было практически невозможно, поэтому передача программ, обмен программами практиковались редко.

Для этих машин был характерен открытый режим использования: программист приходил в отведенное ему время в машинный зал, садился за пульт ЭВМ и сам «пропускал» свою программу. Очень много времени и сил занимали «отладки» новых программ, т. е. выискивание и исправление ошибок, проверка программ с помощью пробных тестовых расчетов. Хотя результаты этих расчетов никакого практического интереса, как правило, не представляли (расчеты делались для контроля программы), они «съедали» до 50% дефицитного машинного времени.

В эпоху машин первого поколения, охватившую 50-е годы, началась разработка стандартных и типовых программ, составление библиотек программ во внешних носителях памяти с инструкциями по их использованию. Столкнувшись с типичной задачей или типичным элементом в большой задаче, математик мог прямо воспользоваться готовой программой.

Машины второго поколения. Согласно принятой классификации, ко второму поколению относятся машины, построенные на транзисторах с быстродействием порядка 10^4—10^5 операций в секунду, оперативной памятью около 10^4 слов, с расширенными возможностями ввода-вывода, с более развитыми методами общения человека с ЭВМ. Различных типов машин второго поколения, в том числе и советских, очень много. Из отечественных серийных машин к ним относится ряд малых машин, например «Мир», ряд машин средней производительности, например «Минск-23», и ряд ЭВМ большой производительности, таких, как «Минск-32» (см. рис. 12), М-220, БЭСМ-4, БЭСМ-6 (см. рис. 13).

Рассмотрим на примере ЭВМ БЭСМ-6 какие возможности скрывались за данной выше общей характеристикой машин второго поколения. БЭСМ-6, серийный выпуск которой начался в 1967 году, является одной из самых больших и совершенных машин второго поколения в мире. Ее быстродействие — миллион операций в секунду (оно превышает указанные выше цифры «среднего» быстродей-

ствия машин второго поколения), оперативная память 128К слов, длина слова — 48 двоичных разрядов. Таким образом, БЭСМ-6 в 50 раз быстрее машины М-20, а ее оперативная память примерно соответствует 1 200 000

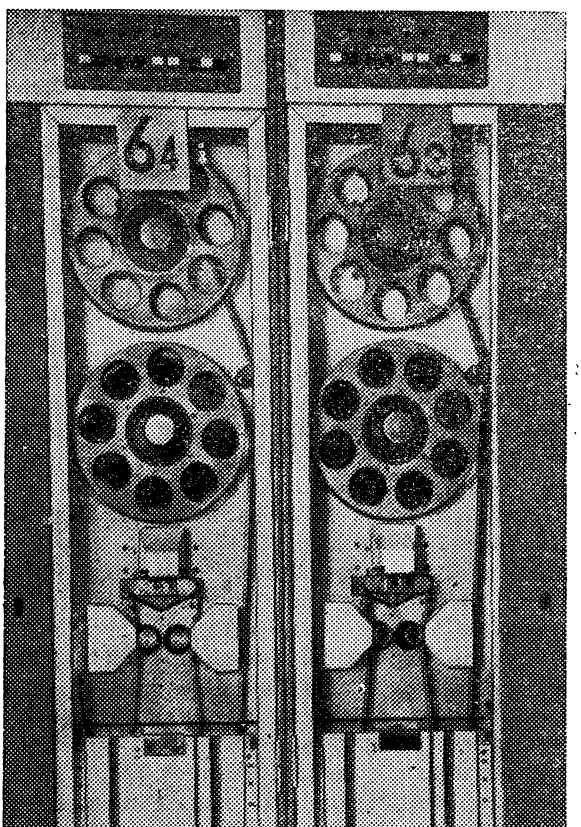

Рис. 14. Магнитные ленты.

буквам или книге в 600 страниц. Объем промежуточной памяти на магнитных барабанах — 512К, т. е. в четыре раза больше оперативной памяти. К центральному процессору могут быть подключены 32 магнитофона с емкостью каждой ленты до миллиона машинных слов (см. рис. 14). На такую ленту можно записать в закодирован-

ном виде 5 тысяч страниц текста, т. е. 10-томное издание. С 1972 года в качестве промежуточной памяти стали также использовать магнитные диски, емкость которых значительно превышает емкость барабанов (см. рис. 15).

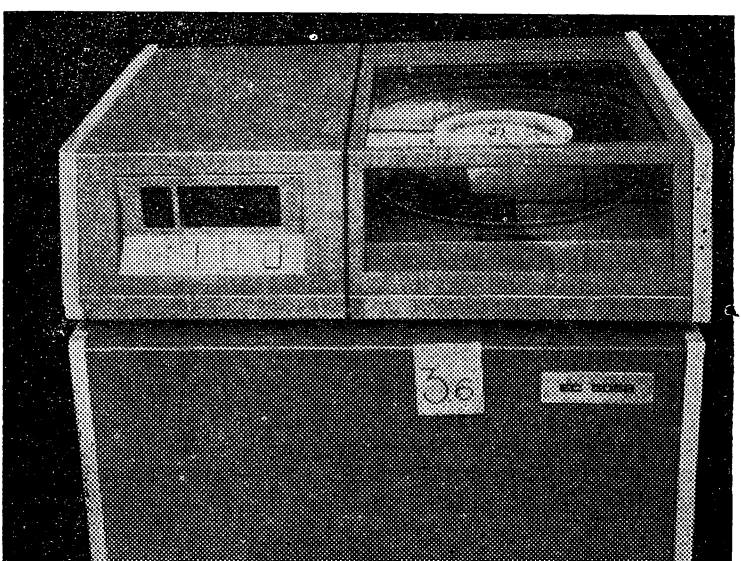

Рис. 15. Магнитный диск.

Существенное повышение технических характеристик машин второго поколения, особенно таких, как БЭСМ-6, расширило их возможности, позволило решать большие и сложные задачи. Однако усложнение задач неизбежно связано с усложнением программ. Программирование и отладка грозили стать самым «узким» местом, снижающим эффективность использования ЭВМ. Возникла проблема дальнейшей автоматизации работы, включения в нее не только вычислительного процесса, но также процесса программирования и отладки программ. Эта проблема была решена благодаря переходу от программирования на языке машины к программированию на формальных алгоритмических языках.

В настоящее время разработано несколько таких языков, различных по своему уровню и назначению. Наиболее распространенными среди них являются фортран и алгол.

Мы не будем останавливаться на описании этих языков: данная книга не учебник по программированию. Расскажем только о тех новых возможностях, которые появились благодаря их введению.

Алгоритмические языки наглядны: они максимально полно используют привычную математическую символику и другие легко понимаемые изобразительные средства. Фразы этих языков состоят из нужных формул, записанных в обычном, понятном любому математику виде, и из нескольких стандартных терминов на английском языке. Важным достоинством алгоритмических языков является их универсальность и наличие международного стандарта, они совершенно не зависят от конкретного типа машины, для которой предназначена написанная программа. Программист, работающий на алгоритмическом языке, может вообще не знать ее систему команд, ему практически не нужно переучиваться при переходе с одной машины на другую.

Алгоритмический язык в силу его лаконичности и строгости правил построения фраз дисциплинирует мышление. Логическое несовершенство метода, скрытые изъяны часто обнаруживаются при попытке его описания на алгоритмическом языке.

Для того чтобы программа, написанная на алгоритмическом языке, могла быть исполнена ЭВМ, она должна быть сначала переведена с этого универсального языка на собственный язык машины. Делает это сама ЭВМ с помощью специальной программы — транслятора *). Транслятор проводит логический анализ программ, написанных на алгоритмическом языке, и осуществляет их перевод на язык машины. Транслятор — очень сложная программа, его создание — большая работа, которую проводит группа высококвалифицированных специалистов по «системному программированию». Однако важно подчеркнуть, что обычные пользователи ЭВМ могут ничего не знать ни о самом трансляторе, ни о принципах его работы. Это не помешает им писать программы на алгоритмическом языке и проводить по ним расчеты.

Программирование на алгоритмических языках, в силу их естественности для человека, гораздо проще программирования на машинном языке. Многие элементы программирования, например, распределение памяти, орга-

*) Транслятор (по-английски — translator) — переводчик.

низация ввода-вывода, вычисление элементарных функций и т. п., транслятор берет на себя. Поэтому при составлении программы математики стали делать существенно меньше ошибок. Транслятор проводит подробный анализ написанной программы и при наличии в ней ошибок в правилах использования языка сообщает об этом программисту, точно указывая их место. Все это существенно упрощает отладку программы, сокращая ее сроки и расход машинного времени. Не может транслятор обнаружить арифметические ошибки, например, неправильно указанный знак арифметической операции в формуле. Такие ошибки выявляет сам программист с помощью пробных тестовых расчетов.

Алгоритмические языки — не только средство общения человека с машиной, но и средство общения людей между собой. Программа, написанная на одном из алгоритмических языков, может быть сравнительно легко понята любым математиком, знающим этот язык и характер задачи. Текст такой программы не зависит от конкретного вида ЭВМ, он универсален. Поэтому переход к программированию на алгоритмических языках существенно упростил обмен программами между научными центрами и группами.

Машины второго поколения обладали более развитой и совершенной системой ввода-вывода. Появились быстродействующие читающие устройства, способные пропускать до 1000 перфокарт в минуту, алфавитно-цифровые печатающие устройства (АЦПУ), графикопостроители. АЦПУ дали возможность гибко менять форму выдачи результатов, например, печатать их в виде таблиц со словесным описанием приведенных величин (см. рис. 12), графикопостроители — оформлять их в виде готовых графиков. Все это существенно облегчило обработку результатов расчетов, повысило производительность труда человека, вооруженного вычислительной машиной.

Итак, в эпоху ЭВМ второго поколения в развитии вычислительной техники, в повышении ее быстродействия и эффективности был достигнут значительный прогресс. Он сопровождался быстрым увеличением общего числа ЭВМ. Однако рост потребностей в вычислительной технике шел еще более быстрыми темпами, опережая возможности ее производства. Вызвано это было расширением сферы применения ЭВМ, увеличением числа пользователей, существенным усложнением решаемых задач.

Машинное время, которое было дефицитным в эпоху машин первого поколения, стало еще более дефицитным.

Одно из слабых мест ЭВМ первого и отчасти второго поколения состояло в том, что при вводе и выводе данных, т. е. при «общении» машины с человеком, она не считала, ее «мозг» — центральный процессор — бездействовал, а работало только периферийное оборудование. Механические устройства ввода и вывода, несмотря на существенные усовершенствования, не могли угнаться за быстродействием электронного центрального процессора, и потери его рабочего времени от простоев при вводе и выводе данных были значительными.

Жертвой этого конфликта в условиях острого и постоянно растущего дефицита машинного времени, его высокой стоимости, оказался человек. Началась жесткая борьба за сокращение «контактов» пользователей с машиной, в которой на первое место ставились «интересы» машины, ее центрального процессора, а не удобства пользователей. Одним из ее проявлений стал переход от «открытого» режима работы на ЭВМ к «закрытому» режиму: математиков-вычислителей «отлучили» от ЭВМ и перестали пускать в машинный зал. Задачи теперь пропускали операторы по инструкциям, составленным авторами программ.

Это новшество внедрялось в вычислительных центрах, как картошка на Руси, при яростном сопротивлении пользователей ЭВМ. Они доказывали, что без них задача считаться не будет, что операторы все перепутают (так, действительно, иногда бывало), что в инструкции нельзя оговорить всех тонкостей и т. д. Однако закрытый режим позволил более эффективно использовать вычислительную технику, существенно сократить непроизводительные потери машинного времени, и это предопределило исход борьбы: сопротивление пользователей было сломлено.

Преимущество закрытого режима состояло в том, что теперь на машине постоянно работали профессиональные операторы. Они знали пульт гораздо лучше большинства математиков-вычислителей, действовали более четко и грамотно, допускали меньше «операторского» брака. Пользователи, лишенные возможности присутствовать при прохождении своих задач через машину и вносить прямо на месте коррективы и исправления в программы, были вынуждены более внимательно готовить задания, заранее предусматривать возможные осложнения и отра-

жать их в инструкциях операторам. Это также давало значительную экономию машинного времени.

Скоро инструкции операторам заменили инструкции самой машине в виде набора стандартных перфокарт, которые добавлялись к программе. Был введен режим «пакетной» обработки: программы собирались одна за другой и ставились в читающее устройство ЭВМ. Машина, закончив очередные расчеты, сама вводила следующую программу, границы которой выделялись специальными картами, и приступала к ее решению. Операторам оставалось только контролировать работу ЭВМ и следить за тем, чтобы ее читающее устройство не пустовало.

Этот комплекс технических и организационных мер существенно повысил эффективность использования вычислительной техники.

Машины третьего поколения. Совершенствование технологии производства полупроводников привело к созданию микроэлементных устройств, получивших название интегральных схем. Интегральная схема представляет собой кристалл химически чистого материала, на котором в результате сложного многоэтапного процесса создаются полупроводниковые области, резисторы (сопротивления), соединения. На одном кристалле размещается небольшая схема, состоящая примерно из двух десятков элементов и заменяющая целый электронный блок.

Интегральные схемы стали элементной базой машин третьего поколения. Они открыли возможность для дальнейшего совершенствования логики ЭВМ, повышения быстродействия до 10^5—10^7 операций в секунду, расширения оперативной памяти до 10^4—10^5 машинных слов.

При разработке машин третьего поколения был учтен богатый опыт эксплуатации ЭВМ второго поколения со всеми их достоинствами и недостатками. Достоинства старались развить дальше, а недостатки — устранить. В результате новые машины отличались от своих предшественниц не только элементной базой, быстродействием, объемом памяти, они предоставляли пользователям гораздо больше возможностей и удобств в работе. Мы расскажем только о двух принципиально новых особенностях, характерных для третьего поколения ЭВМ, которых не было у машин второго поколения.

Первая из них связана с изменением формы общения человека с машиной. Выше уже говорилось о трудностях, возникших для пользователей в эпоху ЭВМ второго

поколения. Решение проблемы «контактов» было найдено в машинах третьего поколения за счет одновременного выполнения вычислений и операций ввода-вывода. Машину разбили на отдельные независимые модули: центральный процессор и специальные процессоры для управления устройствами ввода и вывода. Это позволило перейти на мультипрограммный режим, при котором ЭВМ работала сразу с несколькими программами.

В мультипрограммном режиме центральный процессор не простаивает без дела. Пока он проводит расчеты по одной из программ, внешние устройства печатают полученные ранее результаты, вводят и подготавливают следующую программу. Закончив очередной этап расчетов, центральный процессор сразу переключается на другую программу, передав выдачу сосчитанных результатов устройству вывода.

Организация работы ЭВМ в мультипрограммном режиме потребовала сложнейшего согласования действия всех модулей, создания специальных операционных систем, управляющих работой машины, решения проблемы прерывания работы программы с запоминанием ее состояния и т. д.

Эти усовершенствования позволили машинам третьего поколения одновременно обслуживать несколько пользователей. Каждый из них связывается с ЭВМ через выносной терминал *), который установлен либо непосредственно в его рабочей комнате, либо в специально оборудованном помещении. С выносного терминала пользователь может ввести в ЭВМ свою программу или (это бывает гораздо чаще) вызвать ее в оперативную память с внешнего носителя памяти и произвести нужные расчеты. Просмотрев результаты, он может внести изменения в программу, в исходные данные и продолжить вычисления. При таком режиме пользователь работает с ЭВМ в форме диалога. Если бы он ничего не знал о существовании других терминалов, за которыми одновременно с ним сидят другие пользователи, то у него была бы полная иллюзия того, что машина полностью находится в его личном распоряжении.

Итак, спираль развития вычислительной техники и ее использования человеком завершила очередной виток: ма-

*) Выносной терминал (от латинского terminus — конец, предел) — устройство ввода-вывода, расположенное вне машинного зала ЭВМ.

тематики-вычислители, попавшие «в опалу» в эпоху машин второго поколения, были реабилитированы. Они снова получили прямой доступ к ЭВМ, только теперь и сами машины, и форма связи с ними существенно изменились. Математику не нужно было больше идти в машинный зал, машина

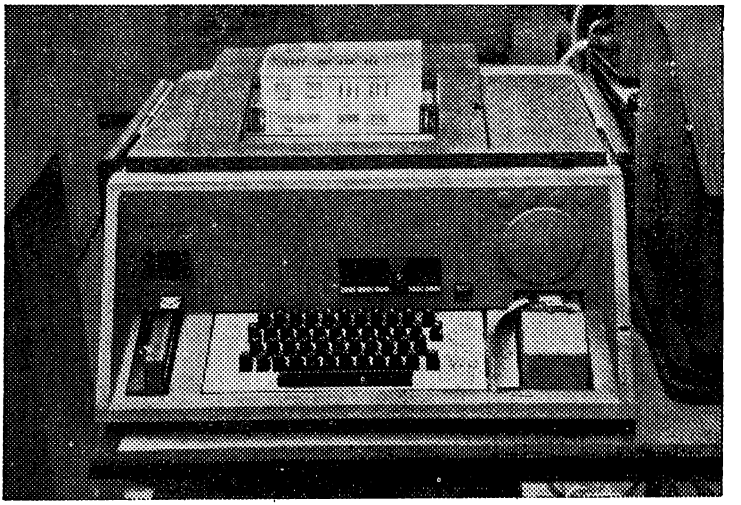

Рис. 16. Телетайп.

сама пришла в его рабочее помещение через выносной терминал. Существенное расширение контактов с машиной и изменение их формы потребовали создания нового периферийного оборудования.

Широкое распространение получили телетайпы, которыми начали пользоваться еще в эпоху машин второго поколения. Они применяются в машинных залах для связи операторов с ЭВМ и в качестве выносных терминалов для пользователей. Телетайп — это электромеханическое устройство, аналогичное телеграфному аппарату, который можно увидеть в любом почтовом отделении (см. рис. 16). Буквенный и цифровой текст набирают на клавиатуре пишущей машинки телетайпа. Он печатается на бумаге и одновременно в закодированном виде (как телеграмма) посылается по кабелю в ЭВМ. Машина может отвечать на вопросы, заданные ей через телетайп, выдавать результаты расчетов. Ее ответы будут восприняты

устройством, декодированы и отпечатаны пишущей машинкой.

Большие удобства пользователям предоставляет терминальное устройство с электронно-лучевой трубкой, получившее название дисплея *) (см. рис. 17). Дисплеи имеют

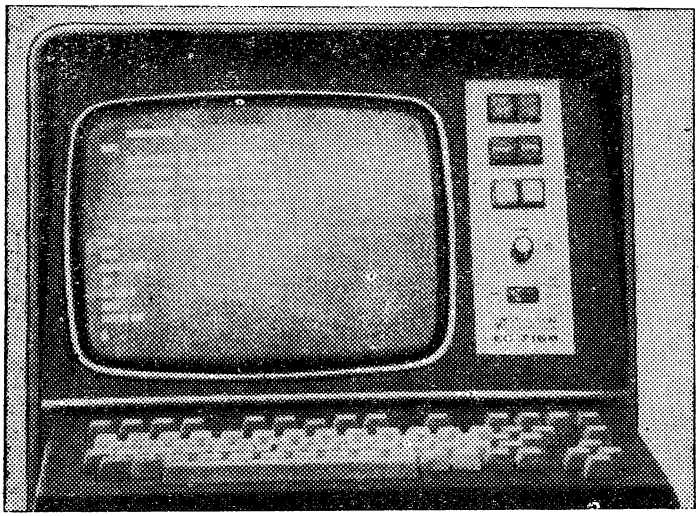

Рис. 17. Дисплей.

клавиатуру и с них, как и с телетайпов, можно вводить программу и данные в ЭВМ, только в этом случае контрольный текст не печатается, а высвечивается на экране. Информация, которая выводится из машины, также подается на экран. Она может иметь форму либо алфавитно-цифрового текста, либо графического материала. Это делает работу с дисплеем особенно удобной, поскольку позволяет представить результаты расчетов не только в виде таблиц, но и в виде графиков, наглядность которых существенно упрощает последующий анализ.

При моделировании на ЭВМ сложного объекта появляется возможность увидеть на экране дисплея его поведение и, в случае необходимости, снять с экрана на кинопленку. Такие кинофильмы стали одним их распространенных способов представления результатов расчета.

*) Дисплей (по-английски — display) — показ, выставление напоказ, выставка.

Комбинация фотоэлемента и электронно-лучевой трубки дисплея используется в качестве прямого устройства ввода в ЭВМ информации, зафиксированной на фотопленке. Пленка помещается между трубкой и фотоэлементом. Когда луч пробегает экран, фотоэлемент фиксирует точки экрана, которые оказались скрытыми непрозрачными участками пленки. Информация о почернении плепки преобразуется в числовую и вводится в ЭВМ. Такие устройства оказались спасением для физиков: они позволили автоматизировать обработку огромного числа (десятки, сотни тысяч) фотографий треков частиц в пузырьковых камерах.

Дисплеи предоставляют очень широкие возможпости для работы с машиной в форме диалога. Они непрерывно совершенствуются: появились цветные дисплеи, дисплеи, обладающие собственным процессором и памятью.

Переходя к обсуждению второй особенности вычислительной техники третьего поколения, отметим следующее. В эпоху ЭВМ второго поколения появилось очень много различных типов машин, которые имели близкие характеристики, но отличались системой команд и способом их кодировки. Для каждой из них приходилось разрабатывать свое математическое обеспечение: трансляторы с алгоритмических языков, библиотеку стандартных программ и т. д. Стоимость математического обеспечения быстро росла и нередко стала превышать стоимость самих машин. Это явилось одной из причин, по которой машины третьего поколения, обладающие еще более сложным математическим обеспечением, стали разрабатывать не поодиночке, а семействами. ЭВМ одного семейства могли отличаться быстродействием, объемом памяти, однако все они являлись конструктивно, программно, информационно совместимыми и обладали одинаковым математическим обеспечением. Это существенно снижало расходы на разработку математического обеспечения и предоставляло широкие возможности для наиболее экономного, эффективного использования вычислительной техники в зависимости от характера решаемых задач.

Одним из примеров такого семейства является единая система электронных вычислительных машин (ЕС ЭВМ) третьего поколения, которая создается и выпускается в рамках международного сотрудничества социалистических стран — участниц СЭВ по многостороннему соглашению, подписанному в декабре 1969 года. Выпуск машин

единой системы начался в 1972 году. Их вместе с периферийным оборудованием производили и поставляли друг другу страны — участницы соглашения. Осуществление в короткий срок такого крупного проекта явилось претворением в жизнь принципов экономической интеграции социалистических стран, лежащей в основе деятельности СЭВ.

В последующие годы начался выпуск модифицированных моделей ЕС ЭВМ (см. рис. 18). В настоящее время

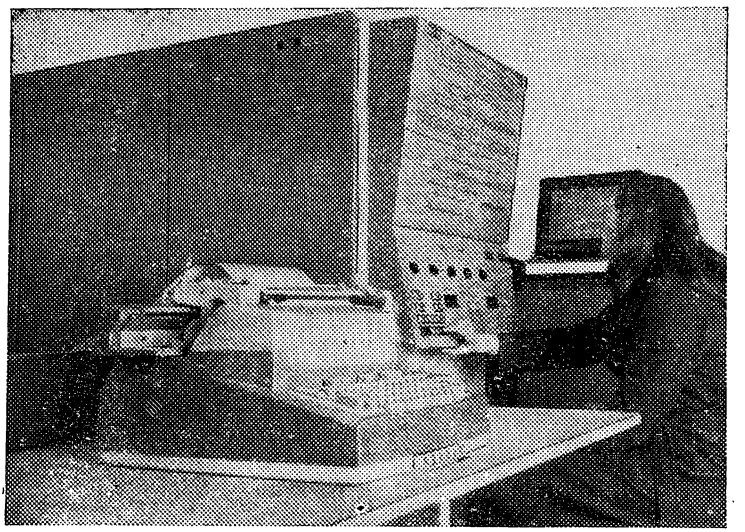

Рис. 18. Машина ЕС 1022 из серии ЕС ЭВМ. На переднем плане виден телетайп, с помощью которого оператор может вести диалог с машиной, на заднем плане — дисплей.

в рамках единой серии насчитывается 17 ЭВМ и более 150 внешних устройств, каждое из которых совместимо с любой из 17 машин. В единой серии приняты международные стандарты на технические характеристики всех устройств и узлов ЭВМ, на систему кодов, операций, средств программирования. ЭВМ единой серии совместимы с моделями ЭВМ ряда капиталистических стран, например, крупнейшей американской фирмы IBM. Это открывает возможности для широкого международного сотрудничества в области использования вычислительной техники и обмена программным обеспечением.

Машины четвертого поколения. ЭВМ четвертого поколения, которые призваны сделать следующий важный шаг в развитии вычислительной техники, в повышении ее производительности, представляют собой многопроцессорные вычислительные комплексы с общей памятью и системой ввода-вывода. Их элементной базой являются большие интегральные схемы (БИС). БИС также размещается на одном кристалле, число элементов в ней достигает нескольких тысяч. Одна БИС могла бы заменить все арифметическое устройство ламповой машины первого поколения.

Машины четвертого поколения одновременно параллельно выполняют большое число операций. Это дает возможность поднять их быстродействие до 10^8—10^9 операций в секунду. Новая элементная база позволяет уменьшить размеры машины, что также имеет существенное значение для быстродействия. (Сигнал, распространяющийся со скоростью света, проходит 3 м за время $t = 10^{-8}$ с. Если бы машина выполняла все операции последовательно и после каждой из них пересылала информацию на расстояние порядка 3-х метров, то поднять ее быстродействие выше 10^8 операций в секунду было бы невозможно.)

Машины четвертого поколения позволят поставить обслуживание вычислительной техникой на принципиально новую основу. Будут созданы мощные вычислительные центры для организации «коммунального» обслуживания пользователей. Исследователи, научные группы получат возможность связываться с вычислительным центром из любого места страны по совершенным линиям связи (телефонным, телеграфным линиям, через спутники связи), выдавать задание машине и получать через определенное время от нее ответ. Развитие такой системы коммунального обслуживания потребует не только совершенствования вычислительной техники, но и решения многих проблем на общегосударственном уровне. Среди них на первое место выдвигается развитие средств связи. Важное значение приобретает также выпуск совершенного терминального оборудования, способного выполнять сложные действия по редактированию вводимой и выводимой информации, по управлению вводными и выводными устройствами терминалов.

§ 4. Применение ЭВМ

Движение человечества к ЭВМ имело вполне определенную цель: с помощью полной автоматизации упростить и ускорить процесс вычислений. Такая задача была решена на основе принципа программного управления. Однако включение в систему команд наряду с арифметическими операциями логических операций, хранение программы в виде чисел в оперативной памяти машины и связанная с этим возможность изменять программу в ходе вычислений сразу позволили ЭВМ перешагнуть первоначальные рамки быстродействующего вычислительного автомата и превратили их в мощный, гибкий инструмент, который широко применяется в самых различных областях человеческой деятельности.

Мы уже говорили в первой главе о математических моделях как методе математического описания сложных процессов, систем, явлений, конструкций. ЭВМ благодаря своему огромному быстродействию и логическим возможностям позволяют провести всесторонний анализ этих моделей и получить детальную количественную информацию о свойствах изучаемого объекта. Данный метод исследования часто называют «вычислительным экспериментом». Он стал в настоящее время одним из наиболее эффективных и универсальных способов познания законов реального мира и их использования в практической деятельности людей.

Вычислительный эксперимент имеет целый ряд преимуществ перед экспериментом реальным. Он значительно дешевле и доступнее. Во многих случаях вычислительный эксперимент позволяет глубже понять результаты реального эксперимента, сопоставить их с теорией. Часто вычислительный эксперимент проводится для планирования будущих экспериментов и прогнозирования их результатов, для проектирования экспериментальных установок следующего поколения и определения оптимальных режимов их работы. Вычислительный эксперимент совершенно незаменим при изучении таких сложных объектов, как космос, человеческое общество, где постановка большого числа натуральных экспериментов либо затруднена, либо вообще невозможна.

Подчеркивая достоинства вычислительного эксперимента, нужно в то же время отметить его ограниченность. Мы знаем, что математическая модель всегда является

упрощенным отражением реального объекта, что полученные с ее помощью результаты носят для изучаемого объекта приближенный характер. Установить пределы применимости математической модели, степень ее соответствия объекту можно только с помощью настоящего эксперимента (критерий практики). Вот почему, какими бы совершенными ЭВМ мы ни обладали, вычислительный эксперимент никогда не вытеснит обычный эксперимент. Будущее за их разумным, гармоничным сочетанием.

ЭВМ — не только техническая база вычислительного эксперимента, одновременно они являются важным элементом реальных экспериментов. Высокий уровень автоматизации многих естественнонаучных экспериментов и способов регистрации их результатов позволяет получить в короткий срок весьма большой объем информации: десятки и сотни тысяч снимков, осциллограмм, показаний детекторов и т. д. Для интерпретации этой информации требуется сложная математическая обработка. Во многих случаях такую обработку необходимо проводить практически одновременно с экспериментом. Сделать это можно только с помощью ЭВМ.

Включение ЭВМ в общий экспериментальный комплекс потребовало создания эффективных численных методов решения математических задач, возникающих при обработке и интерпретации результатов эксперимента, и разработки реализующих эти методы программ. Пришлось также решать сложные технические проблемы сопряжения измерительной аппаратуры с устройствами ввода ЭВМ. Для полной автоматизации обработки важно, чтобы преобразование сигналов любой природы в систему чисел, понятную ЭВМ, осуществлялось автоматически, без участия человека.

Применение ЭВМ позволяет не только обрабатывать эксперимент, но с помощью системы обратных связей управлять им: поддерживать нужные значения параметров, определяющих его условия, менять параметры по заданному закону, вести поиск оптимальных режимов протекания процессов.

Очень быстро были оценены логические возможности ЭВМ, на которые стали обращать особое внимание при проектировании новых машин. ЭВМ стали применяться для логического анализа сложных объектов и ситуаций, для решения задач, связанных с получением выводов из некоторой системы исходных предпосылок. Это дало

новый толчок такой абстрактной науке, как математическая логика. Характерным примером решения сложной логической задачи на ЭВМ является создание трансляторов для перевода программ с алгоритмического языка на машинный язык. Логические возможности ЭВМ позволили создавать программы для игры в шахматы, перевода с одного языка на другой, стихосложения и т. д. Возникли дискуссии, может ли машина «мыслить», проводилось обсуждение вопроса, что такое интеллект, творческая деятельность. Споры часто носили схоластический характер, потому что спорящие стороны вкладывали в одни и те же слова разный смысл. Однако они стимулировали изучение ЭВМ как мощного средства повышения интеллектуальных возможностей человека. То, что человек, вооруженный ЭВМ, интеллектуально сильнее человека без вычислительной машины, то, что человек с помощью ЭВМ сумел решить существенно более широкий круг задач, чем до их появления, сомнений не вызывало и не оспаривалось.

ЭВМ стимулировали развитие кибернетики — науки об общих законах управления. Машины начали широко применяться для управления сложными системами самой различной природы: производственными процессами, полетом космических ракет, естественнонаучными экспериментами и т. д.

На ЭВМ возложили диспетчерские функции в задачах, связанных с переработкой большого объема информации. Вычислительные машины широко используются для оптимальной организации перевозок (мы познакомимся с постановкой транспортной задачи в шестой главе), в информационно-поисковых системах по продаже авиационных и железнодорожных билетов, в системе библиотечного поиска.

Широкое применение получили ЭВМ в задачах проектирования и конструирования объектов самой различной природы: инженерно-строительных сооружений, самолетов и космических аппаратов, нефтяных промыслов, радиоэлектронной аппаратуры. Они не только ускоряют разработку проекта, повышают его качество и тем самым дают большой экономический эффект, но и решают задачи проектирования, анализ которых без ЭВМ оказался бы невозможен.

В качестве характерного примера можно привести проектирование самих ЭВМ, которое в настоящее время

нельзя осуществить без применения ЭВМ. Рост логической сложности, внедрение новой элементной базы привели к тому, что даже целая группа высококвалифицированных специалистов не может удержать в памяти огромное количество связей внутри проектируемой машины. Поэтому работу по проведению связей поручают специальным программам трассировки, которые способны запомнить все существующие связи и по определенным правилам проводить новые. Когда проект готов, он проверяется с помощью моделирования на ЭВМ. Если проверка устанавливает, что проект работоспособен, то машина готовит необходимую техническую документацию для производства отдельных плат, узлов и связей между ними. Эта производственная документация представляет собой перфорированную ленту, которая вводится в программно-управляющее устройство, обеспечивающее технологический цикл производства новой вычислительной машины. В противном случае ЭВМ-«родитель» сообщает проектировщику об имеющемся дефекте. Проектировщик должен понять, в чем состоит ошибка задания, и исправить ее. Так диалог между проектировщиком и ЭВМ создает объединение удивительных творческих способностей человека с педантичностью, быстродействием и прекрасной памятью машины. Этим примером мы и закончим обсуждение темы, поскольку дальнейшее перечисление без анализа деталей и подробностей не сможет добавить существенно новой информации.

Широта и разнообразие областей применения ЭВМ, существенный прогресс, которого удается добиться всюду, где они начинают использоваться, существование проблем и проектов, которые вообще не могли бы развиваться без ЭВМ,— все это делает вычислительные машины одним из определяющих факторов научно-технического прогресса нашего времени.

* * *

В первых трех главах мы стремились дать общее представление о современной прикладной математике и ее методах, рассказать о важных народно-хозяйственных задачах, которые решаются с помощью ЭВМ. Этим и определялся описательный стиль изложения материала.

Последующие главы носят более специальный характер и посвящены обсуждению некоторых конкретных задач. Стиль изложения в этой части книги существенно

отличается от предыдущего: описание сменяется подробным анализом вопроса с полной математической подстановкой задачи, обоснованием методов ее решения, обсуждением результатов расчетов.

Это, естественно, накладывает определенные ограничения на круг отобранных задач: все они сравнительно простые и по своему характеру близко примыкают к программе средней школы. Однако материал подается таким образом, чтобы был виден переход от «школьных» задач к реальным большим задачам прикладной математики. Начинается эта часть книги с главы, посвященной хорошо известной вам задаче — решению уравнений.

Глава 4

ЧИСЛЕННЫЕ МЕТОДЫ РЕШЕНИЯ УРАВНЕНИЙ

§ 1. Решение уравнения в виде формулы не правило, а исключение

Уравнения сыграли важную роль в истории математики, в развитии ее идей и методов. В то же время они и сегодня представляют большой интерес, поскольку часто встречаются в теоретических и прикладных задачах.

Из школьного курса математики вы знакомы с линейными и квадратными уравнениями, корни которых могут быть найдены по известным формулам. Существуют также формулы для решения уравнений третьей и четвертой степеней, однако они очень сложны и неудобны для практического применения. Мы не будем приводить этих формул и останавливаться на их обсуждении, чтобы не отвлекаться от основной цели разговора.

Методы решения линейных и квадратных уравнений были известны еще древним грекам. Решение уравнений третьей и четвертой степеней было получено усилиями итальянских математиков дель Ферро, Тартальи, Кардано, Феррари в XV веке в эпоху Возрождения, когда началось пробуждение европейской математики после средневековой спячки. Затем наступила пора поиска формул для корней уравнений пятой и более высоких степеней. В них принимали участие многие крупнейшие математики. Настойчивые, но безрезультатные попытки продолжались около трехсот лет и завершились в двадцатых годах XIX века благодаря работам норвежского математика Абеля. Абель доказал, что общее уравнение пятой и более высоких степеней неразрешимо в радикалах: решения таких уравнений нельзя выразить через коэффициенты с помощью арифметических действий и извлечения корней. Таким образом, для корней общего уравнения n-й степени

$$a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_0, \quad a_0 \neq 0, \qquad (1)$$

при $n \geqslant 5$ формулы не существует.

Если мы будем рассматривать неалгебраические уравнения, то задача еще больше усложнится. В этом случае найти для корней явные выражения, за редким исключением, не удается.

Возьмем в качестве примера очень простое уравнение:

$$x = \cos x. \tag{2}$$

Построим графики функций, стоящих в левой и правой части. Как видно из рис. 19, они пересекаются при некотором $x = c$, $0 < c < 1$. Число c является корнем уравнения (2), однако получить для него формулу невозможно. «Хоть видит око, да зуб неймет».

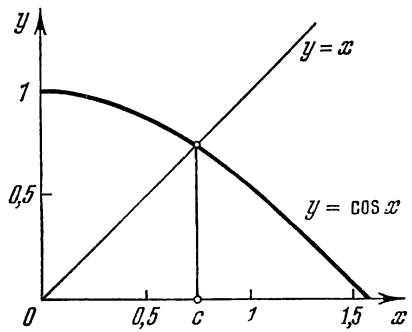

Рис. 19. Графическое решение уравнения $y = \cos x$.

В условиях, когда формулы «не работают», когда рассчитывать на них можно только в самых простейших случаях, важное значение приобретают универсальные вычислительные алгоритмы. Известен целый ряд алгоритмов решения рассматриваемой математической задачи. Если записать уравнение в виде

$$f(x) = 0, \tag{3}$$

то эти алгоритмы обычно не накладывают никаких ограничений на конкретный вид функции $f(x)$, а предполагают только, что она обладает некоторыми свойствами типа непрерывности, дифференцируемости и т. д.

В этой главе мы познакомимся с тремя из них. Выбранные нами для обсуждения алгоритмы основаны на различных идеях, каждый обладает определенными достоинствами, поэтому в конце главы будет интересно сравнить их между собой. Однако прежде, чем перейти к описанию и обоснованию алгоритмов, рассмотрим в следующем параграфе некоторые общие вопросы качественного исследования уравнений.

§ 2. Качественное исследование уравнений. Теорема о существовании корня у непрерывной функции

Часто при решении уравнений важно знать заранее, имеет ли оно корни и, если имеет, то где они, примерно, располагаются. Рассмотрим квадратное уравнение

$$ax^2 + bx + c = 0. \qquad (4)$$

Если подсчитать его дискриминант $\delta = b^2 - 4ac$ и убедиться, что он положителен, то можно сделать следующий вывод: уравнение (4) имеет два действительных корня, причем один из них лежит левее точки $x_0 = -b/(2a)$, другой — правее.

Этот случай тривиален — наш вывод основан на формуле для корней, т. е. на известном решении задачи. Гораздо важнее научиться проводить исследование уравнений, не имея под рукой готового ответа.

Посмотрите на рис. 20. На нем изображен график некоторой функции $f(x)$, непрерывной на отрезке $[0, 1]$ и принимающей на концах отрезка значения разных знаков: $f(0) < 0$, $f(1) > 0$. График является непрерывной линией, которую можно нарисовать, не отрывая карандаша от бумаги. Линия должна перейти из нижней полуплоскости $y < 0$ в верхнюю $y > 0$. При этом она не может «перепрыгнуть» через ось x, а должна ее обязательно пересечь в некоторой точке $x = c$. В этой точке функция $f(x)$ обращается в нуль, т. е. c является корнем уравнения (3).

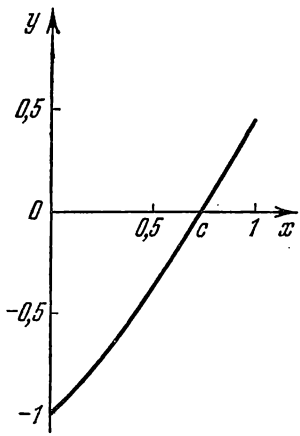

Рис. 20. Пример функции $f(x)$, непрерывной на отрезке $[0, 1]$ и принимающей на концах отрезка значения разных знаков: $f(0) < 0$, $f(1) > 0$.

Мы рассуждали на интуитивном уровне, теперь сформулируем результат в виде теоремы.

Теорема о существовании корня у непрерывной функции. *Если функция $f(x)$ непрерывна на отрезке $[a, b]$ и принимает на его концах значения разных знаков, то на этом отрезке существует по крайней мере один корень уравнения* (3).

Обратите внимание на то, что, гарантируя существование решения уравнения, теорема не позволяет опреде-

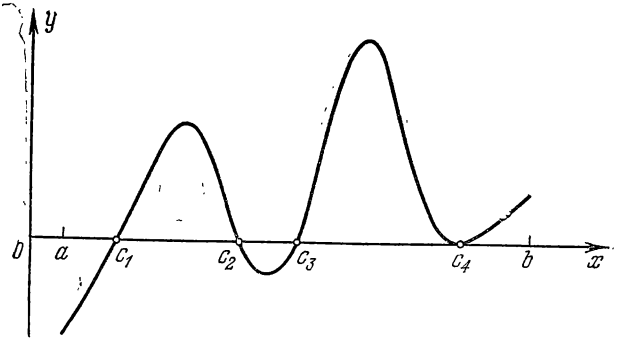

Рис. 21. Пример функции $f(x)$, удовлетворяющей условиям теоремы и имеющей на отрезке $[a, b]$ четыре корня.

лить точного числа его корней. На рис. 21 в качестве примера приведен график функции, удовлетворяющей условиям теоремы и имеющей на рассматриваемом отрезке четыре корня.

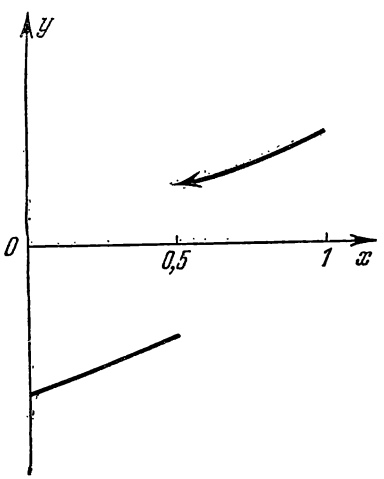

Рис. 22. Пример разрывной функции $f(x)$, принимающей на концах отрезка $[a, b]$ значения разных знаков, но не имеющей на этом отрезке корней.

Требование непрерывности функции $f(x)$ во всех точках отрезка $[a, b]$ существенно. При наличии хотя бы одной точки разрыва утверждение теоремы становится неверным. Соответствующий пример показан на рис. 22. Мы видим на нем график разрывной функции, которая принимает на концах отрезка $[a, b]$ значения разных знаков, но не имеет корней.

Сформулированная теорема относится к числу так называемых теорем существования. Таких теорем, устанавливающих условия разрешимости различных математических задач, очень много. С некоторыми из них мы познакомимся в этой книге.

Доказательства теорем существования можно разделить на два типа. Бывают конструктивные доказательства, их основу составляет метод фактического построения искомого решения. Они играют особенно важную роль в прикладной математике, в которой всегда требуется получить решение рассматриваемой задачи.

Наряду с этим весьма часто встречаются неконструктивные доказательства теорем существования. Как правило, они основаны на рассуждениях от противного: цепь логических заключений показывает, что решение обязано существовать, ибо в противном случае получится противоречие. В качестве характерного примера теоремы с неконструктивным доказательством можно привести основную теорему высшей алгебры. Она утверждает, что всякое алгебраическое уравнение (1) имеет по крайней мере один корень (вообще говоря, комплексный).

В следующем параграфе мы дадим конструктивное доказательство теоремы о корне непрерывной функции. Сейчас же приведем несколько примеров, показывающих, как она применяется при исследовании уравнений.

В качестве первого примера обратимся еще раз к уравнению (2), которое предварительно перепишем в виде (3):

$$f(x) = x - \cos x = 0.$$

Функция $f(x) = x - \cos x$ непрерывна на отрезке $[0, 1]$, а ее значения на концах отрезка имеют разные знаки:

$$f(0) = -1 < 0, \quad f(1) = 1 - \cos 1 > 0.$$

Отсюда сразу следует существование на отрезке $[0, 1]$ по крайней мере одного корня уравнения (2). Раньше мы пришли к этому выводу с помощью наглядных, но математически нестрогих геометрических соображений. Теперь этот вывод — прямое следствие сформулированной теоремы.

Мы уже говорили, что теорема не позволяет определить общего числа корней. Однако в данном случае это легко сделать с помощью дополнительного исследования. Вычислим производную функцию $f(x)$:

$$f'(x) = 1 + \sin x.$$

В интересующей нас области изменения переменной x: $x \in [0, 1]$ она положительна. Следовательно, функция $f(x)$ на отрезке $[0, 1]$ монотонно возрастает и может иметь только один корень.

В качестве второго примера рассмотрим алгебраическое уравнение (1) произвольной нечетной степени n. Обозначим многочлен, стоящий в левой части уравнения, через $P_n(x)$ и отметим, что функция $P_n(x)$ непрерывна на всей числовой прямой. Знак многочлена при достаточно больших по модулю значениях x совпадает со знаком его старшего члена a_0x^n. В силу нечетности n он различен для отрицательных и положительных x. Это позволяет утверждать, что всякое алгебраическое уравнение нечетной степени имеет по крайней мере один действительный корень.

На уравнения четной степени вывод не распространяется, в чем легко убедиться на примере простейшего уравнения:

$$x^2 + 1 = 0.$$

Однако для них с помощью данной теоремы можно установить другой результат: если в алгебраическом уравнении произвольной четной степени n знаки коэффициентов a_0 и a_n противоположны, то это уравнение имеет по крайней мере один отрицательный и один положительный корень.

Предположим, для определенности, что $a_0 > 0$, $a_n < 0$. Тогда при больших по модулю значениях x многочлен $P_n(x)$, как и его старший член a_0x^n, является положительным. В то же время при $x = 0$ он принимает отрицательное значение: $P_n(0) = a_n < 0$. Отсюда сразу следует нужное утверждение.

Познакомившись на этих примерах с методом предварительного качественного исследования уравнений, перейдем теперь к обсуждению вычислительных алгоритмов для нахождения их корней. Первый из них, который мы рассмотрим в следующем параграфе, будет одновременно методом доказательства теоремы о корне непрерывной функции.

§ 3. Метод вилки

В артиллерии существует следующий метод пристрелки: один снаряд посылают с недолетом, второй с перелетом и при этом говорят, что цель взята в «вилку». Послав следующий снаряд со средним значением прицела между двумя предыдущими, смотрят, как он упадет — с недолетом или перелетом. В результате «вилка» сужается. Такая

корректировка прицела продолжается до тех пор, пока снаряды не накроют цель.

Идея этого метода лежит в основе одного из самых простых и эффективных алгоритмов решения уравнений. Его основу составляет процесс построения по методу «артиллерийской вилки» последовательности вложенных друг в друга отрезков $[a_n, b_n]$. Их концы образуют две монотонные последовательности, одна из которых $\{a_n\}$ («недолеты») сходится к некоторой точке $x = c$ снизу ($a_n \leqslant c$), вторая $\{b_n\}$ («перелеты») — сверху ($b_n \geqslant c$). При выполнении условий теоремы, сформулированной в предыдущем параграфе, доказывается, что предельная точка $x = c$ является корнем уравнения (3). Тем самым оказывается установленным факт существования решения этого уравнения на отрезке $[a, b]$. Сам процесс построения последовательности вложенных отрезков $[a_n, b_n]$, содержащих искомый корень $x = c$, позволяет найти его приближенное значение с любой точностью ε подобно вычислению π по периметрам правильных вписанных и описанных многоугольников $\{p_n\}$ и $\{q_n\}$.

Прежде чем переходить к подробному описанию и обоснованию данного метода, докажем одно вспомогательное утверждение, которое нам понадобится в дальнейшем.

Лемма о переходе к пределу в неравенствах. *Пусть члены последовательности* $\{x_n\}$ *удовлетворяют неравенству*

$$x_n \leqslant b \tag{5}$$

и пусть последовательность $\{x_n\}$ *сходится к пределу* a:

$$\lim_{n\to\infty} x_n = a. \tag{6}$$

Тогда предел a *также удовлетворяет неравенству*

$$a \leqslant b. \tag{7}$$

Утверждение остается в силе, если неравенства (5) и (7) заменить на противоположные.

Смысл этой леммы очень прост: в соотношении (5) можно перейти к пределу при $n \to \infty$, заменяя x_n на a. При этом знак неравенства не меняется.

Доказательство. Допустим противное: несмотря на неравенство (5) для членов последовательности $\{x_n\}$, предел a удовлетворяет неравенству противоположного знака: $a > b$.

Положим $\varepsilon = a - b > 0$ и для данного ε укажем такой номер N, чтобы для всех номеров $n > N$ выполнялось соотношение

$$|a - x_n| < \varepsilon = a - b. \tag{8}$$

Отсюда, как следствие, вытекает неравенство

$$a - x_n < a - b$$

или $x_n > b$, что противоречит (5). Лемма доказана.

Теперь перейдем к описанию метода вилки и доказательству с его помощью теоремы о корне непрерывной функции. Предположим, для определенности, что функция $f(x)$ принимает на левом конце отрезка $[a, b]$ отрицательное значение, на правом — положительное:

$$f(a) < 0, \quad f(b) > 0.$$

Возьмем среднюю точку отрезка $[a, b]$ $\xi = (a + b)/2$ и вычислим в ней значение функции $f(x)$. Если $f(\xi) = 0$, то утверждение теоремы доказано: мы нашли на отрезке $[a, b]$ точку $c = \xi$, в которой наша функция обращается в нуль. В противном случае, когда $f(\xi) \neq 0$, поступим следующим образом: рассмотрим два отрезка $[a, \xi]$ и $[\xi, b]$ и выберем один из них, исходя из условия, чтобы функция $f(x)$ принимала на его концах значения разных знаков. Выбранный отрезок обозначим $[a_1, b_1]$. По построению

$$f(a_1) < 0, \quad f(b_1) > 0. \tag{9}$$

Читатель, наверное, уже догадался, что нужно делать дальше. Возьмем среднюю точку отрезка $[a_1, b_1]$ $\xi_1 = (a_1 + b_1)/2$ и опять вычислим в ней значение функции $f(\xi_1)$. Если $f(\xi_1) = 0$, то доказательство теоремы закончено. В противном случае, когда $f(\xi_1) \neq 0$, снова рассмотрим два отрезка $[a_1, \xi_1]$, $[\xi_1, b_1]$ и выберем тот из них, на концах которого функция $f(x)$ принимает значения разных знаков. Выбранный отрезок обозначим $[a_2, b_2]$. По построению

$$f(a_2) < 0, \quad f(b_2) > 0. \tag{10}$$

Будем продолжать этот процесс. В результате либо он оборвется на некотором шаге n благодаря тому, что $f(\xi_n) = 0$, либо будет продолжаться неограниченно. В первом случае вопрос о существовании корня уравнения (3) решен, поэтому нам нужно рассмотреть второй случай.

Неограниченное продолжение процесса дает последовательность отрезков $[a, b]$, $[a_1, b_1]$, $[a_2, b_2]$, ... Эти отрезки вложены друг в друга: каждый последующий отрезок принадлежит всем предыдущим:

$$a_n \leqslant a_{n+1} < b_{n+1} \leqslant b_n, \tag{11}$$

причем

$$f(a_n) < 0, \quad f(b_n) > 0. \tag{12}$$

Длины отрезков с возрастанием номера n стремятся к нулю:

$$\lim_{n\to\infty} (b_n - a_n) = \lim_{n\to\infty} \frac{b-a}{2^n} = 0. \tag{13}$$

Рассмотрим левые концы отрезков $\{a_n\}$. Согласно (11) они образуют монотонно неубывающую ограниченную последовательность. Такая последовательность имеет предел, который мы обозначим через c_1:

$$\lim_{n\to\infty} a_n = c_1. \tag{14}$$

Согласно (11) и лемме о переходе к пределу в неравенствах имеем

$$c_1 \leqslant b_n. \tag{15}$$

Теперь рассмотрим правые концы отрезков $\{b_n\}$. Они образуют монотонно невозрастающую ограниченную последовательность, которая тоже имеет предел. Обозначим этот предел через c_2:

$$\lim_{n\to\infty} b_n = c_2. \tag{16}$$

Согласно неравенству (15) и лемме эти пределы удовлетворяют неравенству $c_1 \leqslant c_2$.

Итак,

$$a_n \leqslant c_1 \leqslant c_2 \leqslant b_n \tag{17}$$

и, следовательно,

$$c_2 - c_1 \leqslant b_n - a_n = \frac{b-a}{2^n}. \tag{18}$$

Таким образом, разность $c_2 - c_1$ меньше любого наперед заданного положительного числа. Это означает, что $c_2 - c_1 = 0$, т. е.

$$c_1 = c_2 = c. \tag{19}$$

Найденная точка c интересна тем, что она является единственной общей точкой для всех отрезков построенной последовательности. Используя непрерывность функции $f(x)$, докажем, что она является корнем уравнения (3).

Мы знаем, что $f(a_n) \leqslant 0$. Согласно определению непрерывности и возможности предельного перехода в неравенствах имеем

$$f(c) = \lim_{n\to\infty} f(a_n) \leqslant 0. \tag{20}$$

Аналогично, учитывая, что $f(b_n) \geqslant 0$, получаем

$$f(c) = \lim_{n\to\infty} f(b_n) \geqslant 0. \tag{21}$$

Из (20) и (21) следует, что

$$f(c) = 0, \tag{22}$$

т. е. c — корень уравнения (3). Теорема доказана.

Процесс построения последовательности вложенных стягивающихся отрезков методом вилки является эффективным вычислительным алгоритмом решения уравнения (3). На n-м шаге процесса получаем

$$a_n \leqslant c \leqslant b_n. \tag{23}$$

Это двойное неравенство показывает, что число a_n определяет искомый корень c с недостатком, а число b_n — с избытком, с ошибкой, не превышающей длину отрезка $\Delta_n = b_n - a_n = (b-a)/2^n$. При увеличении n ошибка стремится к нулю по закону геометрической прогрессии со знаменателем $q = 1/2$. Если задана необходимая точность ε ($\varepsilon > 0$), то чтобы ее достигнуть, достаточно сделать число шагов N, удовлетворяющее условию:

$$N > \log_2\left(\frac{b-a}{\varepsilon}\right). \tag{24}$$

В качестве примера применим метод вилки к решению уравнения (2), записанному в виде (3):

$$f(x) = x - \cos x = 0.$$

Результаты расчетов, связанных с двенадцатикратным делением исходного отрезка [0, 1] пополам, даны в табл. 1. Они определяют корень c с точностью $\varepsilon < (1/2)^{12} < 0{,}00025$.

Итак, мы можем утверждать, что искомый корень c принадлежит отрезку [0,739 013 671 875, 0,739 257 812 500].

ТАБЛИЦА 1

n	a_n	b_n	$\xi_n = \frac{a_n + b_n}{2}$	$f(\xi_n)$
0	0, 000 000 000 000	1, 000 000 000 000	0, 500 000 000 000	—0, 377 582
1	0, 500 000 000 000	1, 000 000 000 000	0, 750 000 000 000	+0, 018 311
2	0, 500 000 000 000	0, 750 000 000 000	0, 625 000 000 000	—0, 185 963
3	0, 625 000 000 000	0, 750 000 000 000	0, 687 500 000 000	—0, 085 335
4	0, 687 500 000 000	0, 750 000 000 000	0, 718 750 000 000	—0, 033 879
5	0, 718 750 000 000	0, 750 000 000 000	0, 734 375 000 000	—0, 007 875
6	0, 734 375 000 000	0, 750 000 000 000	0, 742 187 500 000	+0, 005 196
7	0, 734 375 000 000	0, 742 187 500 000	0, 738 281 250 000	—0, 001 345
8	0, 738 281 250 000	0, 742 187 500 000	0, 740 234 375 000	+0, 001 924
9	0, 738 281 250 000	0, 740 234 375 000	0, 739 257 812 500	+0, 000 289
10	0, 738 281 250 000	0, 739 257 812 500	0, 738 769 531 250	—0, 000 528
11	0, 738 769 531 250	0, 739 257 812 500	0, 739 013 671 875	—0, 000 120
12	0, 739 013 671 875	0, 739 257 812 500		

Отбрасывая десятичные знаки, лежащие за пределами достигнутой точности, будем иметь:

$$0{,}73901 < c < 0{,}73926. \tag{25}$$

§ 4. Метод итераций (метод последовательных приближений)

В этом параграфе мы познакомимся еще с одним численным методом решения уравнений. Предположим, что наше уравнение можно записать в виде

$$x = \varphi(x). \tag{26}$$

Возьмем произвольное значение x_0 из области определения функции $\varphi(x)$ и будем строить последовательность чисел $\{x_n\}$, определенных с помощью рекуррентной формулы *):

$$x_{n+1} = \varphi(x_n), \qquad n = 0, 1, 2, 3, \ldots \tag{27}$$

Последовательность $\{x_n\}$ называется итерационной последовательностью. При ее изучении встают два вопроса:

*) Способ задания последовательности с помощью рекуррентной формулы подробно обсуждался во второй главе. Там же было введено понятие итераций как многократного повторения одной и той же математической операции (см. подстрочные примечания на стр. 38 и 39). Последовательность (27) является итерационной, потому что вычисление ее членов связано с многократным применением одной и той же рекуррентной формулы.

1. Можно ли процесс вычисления чисел x_n продолжать неограниченно, т. е. будут ли числа x_n принадлежать области определения функции $\varphi(x)$?

2. Если итерационный процесс (27) бесконечен, то как ведут себя числа x_n при $n \to \infty$?

Исследование этих вопросов показывает, что при определенных ограничениях на функцию $\varphi(x)$ итерационная последовательность является бесконечной и сходится к корню уравнения (26):

$$\lim_{n\to\infty} x_n = c, \quad c = \varphi(c). \tag{28}$$

Однако для того, чтобы провести это исследование, нам нужно ввести одно новое понятие.

Говорят, что функция $f(x)$ удовлетворяет на отрезке $[a, b]$ условию Липшица, если существует такая постоянная α, что для любых x_1 и x_2, принадлежащих отрезку $[a, b]$, имеет место неравенство

$$|f(x_1) - f(x_2)| \leqslant \alpha |x_1 - x_2|. \tag{29}$$

Величину α в этом случае называют постоянной Липшица.

Если функция $f(x)$ удовлетворяет на отрезке $[a, b]$ условию Липшица, то она непрерывна на этом отрезке. Действительно, пусть x_0 — произвольная точка отрезка. Рассмотрим приращение функции $f(x)$ в этой точке

$$\Delta f = f(x_0 + \Delta x) - f(x_0)$$

и оценим его с помощью неравенства (29):

$$|\Delta f| \leqslant \alpha |\Delta x|. \tag{30}$$

Таким образом, $\lim\limits_{\Delta x\to 0} \Delta f = 0$, что означает непрерывность функции.

Условие Липшица имеет простой геометрический смысл. Возьмем на графике функции $y = f(x)$ две произвольные точки: M_1 с координатами $(x_1, f(x_1))$ и M_2 с координатами $(x_2, f(x_2))$ (см. рис. 23). Напишем уравнение прямой линии, проходящей через эти точки. Оно имеет вид

$$y = f(x_1) + k(x - x_1),$$

где k — тангенс угла наклона прямой к оси x — определяется формулой

$$k = \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

Если функция $f(x)$ удовлетворяет на отрезке $[a, b]$ условию Липшица (29), то при произвольном выборе точек M_1 и M_2 будем иметь: $|k| \leqslant \alpha$. Таким образом, с геометрической точки зрения условие Липшица означает ограниченность тангенса угла наклона секущих, проведенных через всевозможные пары точек графика функции $y = f(x)$.

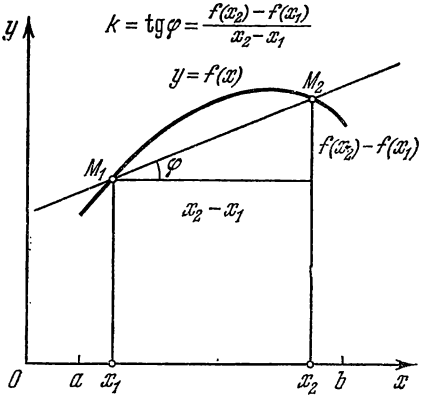

Рис. 23. Геометрическая иллюстрация условия Липшица.

Сделаем теперь следующий шаг — предположим, что функция $f(x)$ имеет на отрезке $[a, b]$ ограниченную производную: $|f'(x)| \leqslant m$ при $x \in [a, b]$. Можно доказать, что в этом случае она удовлетворяет условию Липшица с постоянной $\alpha = m$. Данное утверждение имеет простой геометрический смысл — каждой секущей графика функции $y = f(x)$ можно сопоставить параллельную ей касательную (см. рис. 24). Поэтому наибольший тангенс угла наклона секущих не превосходит наибольшего тангенса угла наклона касательных, и его можно оценить той же константой m: $|k| \leqslant m$. Таким образом, любая функция $f(x)$ с ограниченной производной обязательно удовлетворяет условию Липшица.

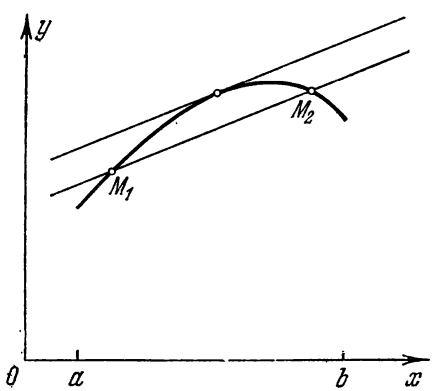

Рис. 24. Геометрическая иллюстрация связи условия Липшица с предположением о дифференцируемости функции $f(x)$.

Познакомившись с условием Липшица, перейдем к изучению итерационной последовательности при предположении, что уравнение (26) имеет корень $x = c$.

Существование этого корня можно установить с помощью предварительного качественного исследования уравнения с применением теоремы § 2 данной главы.

Теорема о сходимости итерационной последовательности. *Пусть c — корень уравнения* (26) *и пусть функция* $\varphi(x)$ *удовлетворяет на некотором отрезке* $[c-\delta, c+\delta]$ $(\delta > 0)$ *условию Липшица с постоянной* $\alpha < 1$. *Тогда при любом выборе* x_0 *на отрезке* $[c-\delta, c+\delta]$ *существует бесконечная итерационная последовательность* $\{x_n\}$ (27) *и эта последовательность сходится к корню* $x = c$, *который является единственным решением уравнения* (26) *на отрезке* $[c-\delta, c+\delta]$.

Сформулированная теорема имеет очень простой смысл. Будем говорить, что функция φ осуществляет отображение точки x на точку $y = \varphi(x)$. Тогда условие Липшица с постоянной $\alpha < 1$ означает, что отображение φ является сжимающим: расстояние между точками x_1 и x_2 больше, чем расстояние между их изображениями $y_1 = \varphi(x_1)$ и $y_2 = \varphi(x_2)$.

Корень c является неподвижной точкой отображения φ, он преобразуется сам в себя: $c = \varphi(c)$. Поэтому каждый шаг в итерационном процессе (27), сжимая расстояния, должен приближать члены последовательности $\{x_n\}$ к неподвижной точке c.

После этих соображений, поясняющих смысл теоремы, перейдем к ее доказательству. Возьмем произвольную точку x_0 на отрезке $[c-\delta, c+\delta]$, она отстоит от точки c не больше, чем на δ: $|c - x_0| \leqslant \delta$.

Вычислим x_1: $x_1 = \varphi(x_0)$, при этом $x_1 - c = \varphi(x_0) - \varphi(c)$. Разность $\varphi(x_0) - \varphi(c)$ можно оценить с помощью условия Липшица:

$$|x_1 - c| = |\varphi(x_0) - \varphi(c)| \leqslant \alpha |x_0 - c| \leqslant \alpha\delta. \quad (31)$$

Неравенство (31) показывает, что x_1 принадлежит отрезку $[c-\delta, c+\delta]$ и расположено ближе к точке c, чем x_0.

Продолжим построение итерационной последовательности. Вычислим x_2: $x_2 = \varphi(x_1)$, при этом $|x_2 - c| = |\varphi(x_2) - \varphi(c)| \leqslant \alpha |x_1 - c| \leqslant \alpha^2 |x_0 - c| \leqslant \alpha^2\delta$.

Точка x_2 опять принадлежит отрезку $[c-\delta, c+\delta]$ и расположена ближе к точке c, чем точка x_1, т. е. мы опять приблизились к c.

По индукции легко доказать, что последующие итерации также существуют и удовлетворяют неравенствам

$$| x_n - c | \leqslant \alpha^n | x_0 - c | \leqslant \alpha^n \delta. \qquad (32)$$

Отсюда следует, что

$$\lim_{n\to\infty} (x_n - c) = 0, \quad \text{т. е.} \quad \lim_{n\to\infty} x_n = c. \qquad (33)$$

Нам остается доказать, что корень $x = c$ является единственным решением уравнения (26) на отрезке $[c - \delta, c + \delta]$. Действительно, допустим, что существует еще один корень $x = c_1$:

$$c_1 = \varphi(c_1), \quad c_1 \in [c - \delta, c + \delta]. \qquad (34)$$

Примем c_1 за нулевое приближение и будем строить итерационную последовательность (27). Тогда с учетом равенства (34) получим: $x_n = c_1$, $n = 0, 1, 2, \ldots$ С другой стороны, по доказанному $\lim\limits_{n\to\infty} x_n = c$, т. е. $c_1 = c$. Никаких других решений уравнение (26) на отрезке $[c - \delta, c + \delta]$ иметь не может.

Сходимость итерационной последовательности к корню уравнения (26) может быть использована для приближенного определения этого корня с любой степенью точности. Для этого нужно только провести достаточное число итераций.

В качестве примера, иллюстрирующего данный метод, рассмотрим еще раз уравнение $x = \cos x$. Роль функции $\varphi(x)$ в нем играет $\cos x$. Это — дифференцируемая функция, производная которой равна $-\sin x$. На отрезке $[0, 1]$

$$| \varphi'(x) | = \sin x \leqslant \sin 1. \qquad (35)$$

Таким образом, функция $\varphi(x) = \cos x$ удовлетворяет на отрезке $[0, 1]$ условию Липшица с постоянной $\alpha = \sin 1 < 1$.

Результаты вычислений по рекуррентной формуле (27), которая в случае уравнения (2) принимает вид $x_{n+1} = \cos x_n$, даны в табл. 2. За нулевое приближение была выбрана средняя точка отрезка: $x_0 = 0{,}5$.

Для удобства анализа итерационной последовательности ее члены расположены по два в строке. В результате образовались столбцы членов с четными и нечетными номерами. Сравнивая их между собой, мы видим, что четные члены меньше нечетных: итерационная последовательность

ТАБЛИЦА 2

n	x_{2n}	x_{2n+1}
0	0, 500 000 000 000	0, 877 582 561 890
1	0, 639 012 494 166	0, 802 685 100 681
2	0, 694 778 026 789	0, 768 195 831 281
3	0, 719 165 445 942	0, 752 355 759 420
4	0, 730 081 063 138	0, 745 120 341 349
5	0, 735 006 309 016	0, 741 826 522 642
6	0, 737 235 725 443	0, 740 329 651 877
7	0, 738 246 238 333	0, 739 649 062 768
8	0, 738 704 539 357	0, 739 341 452 279
9	0, 738 912 449 332	0, 739 201 444 135

скачет то вверх, то вниз. С возрастанием номера четные члены возрастают, а нечетные — убывают, приближаясь друг к другу. Такое поведение последовательности означает, что корень уравнения (2) лежит между четными и нечетными итерациями, первые дают его значение с недостатком, вторые — с избытком. Это позволяет легко контролировать точность, достигнутую после любого числа итераций: погрешность не превышает разности между последними вычисленными нечетным и четным членами.

Например, мы остановили процесс вычислений на 19-й итерации и можем написать для корня c двойное неравенство:

$$x_{18} = 0{,}738\ 912\ 449\ 332 < c < x_{19} = 0{,}739\ 201\ 444\ 135, \quad (36)$$

т. е. члены итерационной последовательности x_{18} и x_{19} определяют c с недостатком и избытком с погрешностью, которая не превышает разности $x_{19} - x_{18}$:

$$\varepsilon < \Delta_{19} = x_{19} - x_{18} < 0{,}0003.$$

Точность, которую мы достигли после 19 итераций, примерно соответствует точности 12 шагов в методе вилки. Причина такого различия ясна. В обоих методах погрешность убывает по закону геометрической прогрессии. Для метода вилки знаменатель прогрессии равен 1/2, он не зависит от вида функции $f(x)$. Для метода итераций знаменатель равен α — постоянной Липшица функции $\varphi(x)$. В рассматриваемом примере $\alpha > 1/2$, поэтому сходимость итераций медленнее сходимости метода вилки. Это заме-

чание означает, что метод итераций имеет преимущество перед методом вилки с точки зрения скорости сходимости только в том случае, когда $\alpha < 1/2$.

§ 5. Метод касательных (метод Ньютона)

Метод касательных, связанный с именем Ньютона, является одним из наиболее эффективных численных методов решения уравнений. Идея метода очень проста. Предположим, что функция $f(x)$, имеющая корень c на отрезке

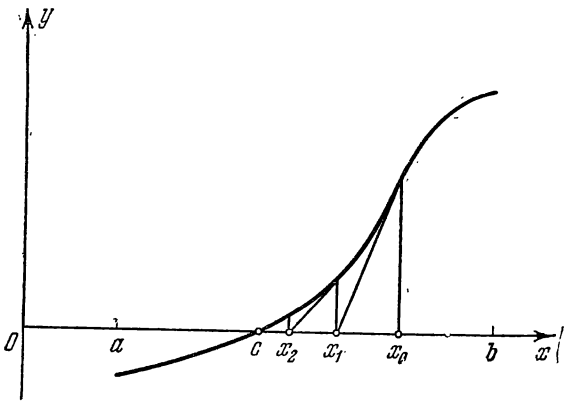

Рис. 25. Построение последовательности $\{x_n\}$ по методу касательных.

$[a, b]$, дифференцируема на этом отрезке и ее производная $f'(x)$ не обращается на нем в нуль. Возьмем произвольную точку x_0 и напишем в ней уравнение касательной к графику функции $f(x)$:

$$y = f(x_0) + f'(x_0)(x - x_0). \tag{37}$$

Графики функции $f(x)$ и ее касательной близки около точки касания, поэтому естественно ожидать, что точка x_1 пересечения касательной с осью x будет расположена недалеко от корня c (см. рис. 25).

Для определения точки x_1 имеем уравнение

$$f(x_0) + f'(x_0)(x_1 - x_0) = 0.$$

Таким образом,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}. \tag{38}$$

Повторим проделанную процедуру: напишем уравнение касательной к графику функции $f(x)$ при $x = x_1$ и найдем для нее точку пересечения x_2 с осью x (см. рис. 25):

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}.$$

Продолжая этот процесс, получим последовательность $\{x_n\}$, определенную с помощью рекуррентной формулы:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \qquad n = 0, 1, 2, \ldots \tag{39}$$

При исследовании этой последовательности, как и последовательности (27) метода итераций, встают два вопроса:

1. Можно ли процесс вычисления чисел x_n продолжать неограниченно, т. е. будут ли числа x_n принадлежать отрезку $[a, b]$?

2. Если процесс (39) бесконечен, то как ведет себя последовательность $\{x_n\}$ при $n \to \infty$?

При анализе этих вопросов предположим, что корень $x = c$ является внутренней точкой отрезка $[a, b]$ $(a < c < b)$, а функция $f(x)$ дважды дифференцируема на данном отрезке, причем ее производные удовлетворяют неравенствам:

$$|f'(x)| \geqslant m > 0, \quad |f''(x)| \leqslant M, \quad x \in [a, b], \tag{40}$$

и докажем следующую теорему.

Теорема о сходимости метода касательных. *Если функция $f(x)$ удовлетворяет сформулированным условиям, то найдется такое δ, $0 < \delta \leqslant \min(c - a, b - c)$, что при любом выборе начального приближения на отрезке $[c - \delta, c + \delta] \subset [a, b]$ существует бесконечная итерационная последовательность* (39) *и эта последовательность сходится к корню c.*

Доказательство. В силу предположения о дифференцируемости функции $f(x)$ и неравенстве нулю ее производной $f'(x)$ уравнение (3) эквивалентно на отрезке $[a, b]$ уравнению

$$x = \varphi(x), \quad \text{где} \quad \varphi(x) = x - \frac{f(x)}{f'(x)}, \tag{41}$$

так что корень $x = c$ исходного уравнения является одновременно корнем уравнения (41). Исследуем возможность отыскания этого корня с помощью метода итераций.

Вычислим производную функции $\varphi(x) = x - f(x)/f'(x)$:

$$\varphi'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2} \qquad (42)$$

и оценим полученное выражение. Согласно неравенствам (40) будем иметь

$$|\varphi'(x)| \leqslant \frac{M}{m^2}|f(x)|. \qquad (43)$$

Для дальнейшей оценки $|\varphi'(x)|$ воспользуемся непрерывностью и равенством нулю функции $f(x)$ в точке $x = c$:

$$\lim_{x \to c} f(x) = f(c) = 0. \qquad (44)$$

Положим $\varepsilon = m^2/(2M)$, тогда, в силу (44), для данного ε можно указать такое δ, $0 < \delta \leqslant \min(c - a, b - c)$, что для всех $x \in [c - \delta, c + \delta]$ выполняется неравенство

$$|f(x) - f(c)| = |f(x)| \leqslant \varepsilon = \frac{m^2}{2M}. \qquad (45)$$

Подставляя (45) в (43), получим

$$|\varphi'(x)| \leqslant \frac{M}{m^2} \cdot \frac{m^2}{2M} = \frac{1}{2}. \qquad (46)$$

Таким образом, функция $\varphi(x)$ удовлетворяет на отрезке $[c - \delta, c + \delta] \subset [a, b]$ условию Липшица с постоянной $\alpha = 1/2 < 1$. Это означает, что уравнение (41) можно решать методом итераций: при любом выборе нулевого приближения x_0 на отрезке $[c - \delta, c + \delta]$ существует бесконечная последовательность $\{x_n\}$ (27), сходящаяся к корню $x = c$.

Теперь нам остается заметить, что итерационной последовательностью для уравнения (41), сходимость которой мы только что установили, является последовательность (39) метода касательных. Теорема доказана.

Требование близости нулевого приближения x_0 к искомому корню c является существенным для метода касательных. Посмотрите на рис. 26. На нем нарисован график той же функции $f(x)$, что и на рис. 25, однако x_0 выбрано дальше от корня c, чем в первом случае. В результате после первого же шага получается точка x_1, которая не принадлежит исходному отрезку $[a, b]$, и на этом процесс построения рекуррентной последовательности метода касательных обрывается.

Таким образом, до начала расчетов по данному методу для выбора нулевого приближения x_0 нужно знать область локализации искомого корня $x = c$. Если известен в общих чертах график функции $f(x)$, то ее легко определить по этому графику. В случае необходимости можно сделать несколько шагов по методу вилки. Затруднения, связанные с предварительным исследованием уравнения, вполне окупаются высокой скоростью сходимости метода.

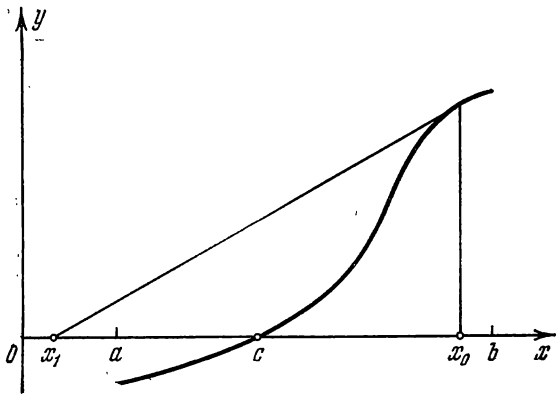

Рис. 26. Случай, когда процесс построения последовательности $\{x_n\}$ обрывается из-за плохого выбора нулевого приближения.

В качестве первого примера применения метода касательных рассмотрим задачу извлечения квадратного корня из произвольного положительного числа a, $a > 0$, который будем искать как решение уравнения

$$f(x) = x^2 - a = 0. \tag{47}$$

Рекуррентная формула метода Ньютона (39) в данном случае принимает вид:

$$x_{n+1} = x_n - \frac{x_n^2 - a}{2x_n} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right). \tag{48}$$

Она совпадает с формулой (18) главы 2. Таким образом, алгоритм вычисления $\sqrt{a}$, который обсуждался во второй главе, основан на решении уравнения (47) методом Ньютона. Мы уже знаем, что процесс сходится к $\sqrt{a}$ при любом выборе начального приближения $x_0 \in]0, +\infty[$, причем сходимость весьма быстрая.

Второй пример снова будет связан с уравнением $x = \cos x$ (мы уже пользовались этим уравнением для иллюстрации двух предыдущих методов). Запишем для него рекуррентную формулу (39) (помня, что $f(x) = x - \cos x$):

$$x_{n+1} = x_n - \frac{x_n - \cos x_n}{1 + \sin x_n}, \qquad n = 0, 1, 2, \ldots \tag{49}$$

Выберем, как и для метода итераций, в качестве нулевого приближения $x_0 = 0{,}5$ и вычислим несколько следующих приближений по формуле (49):

$$x_1 = 0,\ 755\ 222\ 320\ 557$$
$$x_2 = 0,\ 739\ 141\ 702\ 652$$
$$x_3 = 0,\ 739\ 085\ 197\ 449$$
$$x_4 = 0,\ 739\ 085\ 078\ 239$$
$$x_5 = 0,\ 739\ 085\ 078\ 239.$$

Мы видим, что, начиная с номера $n = 1$, последовательность $\{x_n\}$ убывает и приближается к корню $x = c$ сверху. После четвертого шага процесс «останавливается». Мы уже встречались с таким явлением во второй главе, когда обсуждали алгоритм вычисления квадратного корня. Остановка связана с тем, что расчеты ведутся с 12 знаками, и после достижения точности, превышающей 10^{-12}, становится невозможно уловить разницу между x_{n+1} и x_n, лежащую за пределами ошибки округления. Если есть необходимость повысить точность, то нужно перейти к расчетам с бóльшим числом знаков. ЭВМ допускают такую возможность.

Примеры с вычислением корня (47) и решением уравнения (2) показывают очень высокую скорость сходимости метода касательных. Для уравнения (2) после двух шагов была достигнута точность 10^{-4}, после четырех — 10^{-12}. Для сравнения укажем, что точность 10^{-4} метод вилки обеспечивает на 15 шаге, метод итераций — на 22.

§ 6. Заключительные замечания

Мы познакомились с тремя методами численного решения уравнений. Наряду с ними существуют еще несколько методов, на которых мы не останавливались. Ситуация, когда одну и ту же математическую задачу можно решать с помощью разных методов, является

довольно типичной. В таких случаях естественно возникает необходимость сравнить их между собой.

При оценке эффективности численных методов существенное значение имеют различные свойства:

1) универсальность;

2) простота организации вычислительного процесса и контроля за точностью;

3) скорость сходимости.

Посмотрим с этой точки зрения на разобранные методы решения уравнений.

1) Наиболее универсальным является метод вилки: он требует только непрерывности функции $f(x)$. Два других метода накладывают более сильные ограничения. В некоторых случаях это преимущество метода вилки может оказаться существенным.

2) С точки зрения организации вычислительного процесса все три метода очень просты. Однако и здесь метод вилки обладает определенным преимуществом. Вычисления можно начинать с любого отрезка $[a, b]$, на концах которого непрерывная функция $f(x)$ принимает значения разных знаков. Процесс будет сходиться к корню уравнения $f(x) = 0$, причем на каждом шаге он дает для корня двухстороннюю оценку, по которой легко определить достигнутую точность. Сходимость же метода итераций или касательных зависит от того, насколько удачно выбрано нулевое приближение.

3) Наибольшей скоростью сходимости обладает метод касательных. В случае, когда подсчет значений функции $f(x)$ сложен и требует больших затрат машинного времени, это преимущество становится определяющим.

Итак, мы видим, что ответ на вопрос о наилучшем численном методе решения уравнений не может быть однозначным. Он существенно зависит от того, какую дополнительную информацию о функции $f(x)$ мы имеем и, в соответствии с этим, каким свойствам метода придаем наибольшее значение.

При обосновании метода итераций и метода Ньютона на функции $\varphi(x)$ и $f(x)$, а также на выбор начального приближения x_0 накладывались определенные ограничения. Однако при решении конкретных задач проверить их выполнение часто бывает трудно и даже практически невозможно. Функция может не задаваться в виде простой формулы, а находиться в результате численного решения некоторой математической задачи, получаться из изме-

рений и т. д. В таких случаях применимость метода приходится проверять «экспериментально»: начинают расчет и следят за поведением первых членов последовательности $\{x_n\}$. Если по ним видно, что процесс сходится, то расчет продолжают, пока не достигнут нужной точности. В противном случае вычисления прекращают и анализируют полученные данные, пытаясь установить причину расходимости и, в соответствии с ней, выбрать другой метод решения задачи.

Такая организация работы не связана с конкретной темой нашего разговора — численными методами решения уравнений. Она носит общий характер. В прикладной математике существует много численных методов, особенно относящихся к сложным задачам, которые пока не получили строгого обоснования, но успешно применяются на практике. Именно этот факт является основным аргументом в их пользу.

В заключение отметим, что центральная идея метода итераций — сжимающие отображения — является весьма общей. При незначительной модификации она может быть использована для изучения гораздо более сложных математических задач, чем уравнение (3). Для многих типов нелинейных уравнений принцип сжимающих отображений является, по существу, единственным методом исследования и решения. Существенные обобщения допускает также метод Ньютона. Оба метода в своей общей форме играют важную роль в современной теоретической и вычислительной математике.

Глава 5

ЗАДАЧИ ОПТИМИЗАЦИИ

Эта глава посвящена математическим вопросам, связанным с проблемой оптимизации. Явно или неявно мы встречаемся с оптимизацией в любой сфере человеческой деятельности от самого высокого общегосударственного уровня (составление и выполнение пятилетних планов развития народного хозяйства) до сугубо личного (как лучше распределить месячный заработок семьи на питание, покупку необходимых вещей, проведение досуга). Экономическое планирование, управление, распределение ограниченных ресурсов, анализ производственных процессов, проектирование сложных объектов всегда должно быть направлено на поиск наилучшего варианта с точки зрения намеченной цели. Это — важнейшее условие научно-технического прогресса.

При небывалом разнообразии задач оптимизации только математика может дать общие методы их решения. Однако для того, чтобы воспользоваться математическим аппаратом, необходимо сначала сформулировать интересующую нас проблему как математическую задачу, придав количественные оценки возможным вариантам, количественный смысл словам «лучше», «хуже».

Многие задачи оптимизации сводятся к отысканию наименьшего (или наибольшего) значения некоторой функции, которую принято называть целевой функцией или критерием качества. Постановка задачи и методы исследования существенно зависят от свойств целевой функции и той информации о ней, которая может считаться доступной в процессе решения, а также которая известна априори (до опыта, заранее; здесь — до начала решения задачи).

Наиболее просты, с математической точки зрения, случаи, когда целевая функция задается явной формулой и является при этом дифференцируемой функцией. Вычислив производную, мы можем использовать ее для исследования самой функции. С этими идеями вы немного знакомы из школьного курса математики.

В последние десятилетия в условиях научно-технического прогресса круг задач оптимизации, поставленных практикой, резко расширился. Во многих из них целевая функция не задается формулой, ее значения могут получаться в результате сложных расчетов, браться из эксперимента и т. д. Такие задачи являются более сложными, потому что анализ целевой функции с помощью производной для них не работает. Пришлось уточнять их математическую постановку и разрабатывать специальные методы решения, рассчитанные на широкое применение ЭВМ. Следует также иметь в виду то, что сложность задачи существенно зависит от ее размерности, т. е. от числа аргументов целевой функции.

Первые три параграфа данной главы посвящены одномерным задачам оптимизации, в последнем параграфе рассматриваются многомерные задачи. Выделение и подробный разбор одномерных задач имеет определенный смысл. Эти задачи наиболее просты, на них легче понять постановку вопроса, методы решения и возникающие трудности. В ряде случаев, хотя и очень редко, одномерные задачи имеют самостоятельный практический интерес. Однако самое главное заключается в том, что алгоритмы решения многомерных задач оптимизации, как правило, сводятся к последовательному многократному решению одномерных задач и не могут быть поняты без умения решать такие задачи.

§ 1. Задача о наилучшей консервной банке

Перед вами поставили задачу: указать наилучший вариант консервной банки фиксированного объема V, имеющей обычную форму прямого кругового цилиндра. Получив такое задание, вы неизбежно должны спросить: «По какому признаку следует сравнивать банки между собой, какая банка считается наилучшей?» Иными словами, вы попросите указать цель оптимизации.

Рассмотрим два варианта этой задачи.

1. Наилучшая банка должна иметь наименьшую поверхность S. (На ее изготовление пойдет наименьшее количество жести.)

2. Наилучшая банка должна иметь наименьшую длину швов l. (Швы нужно сваривать, и мы хотим сделать эту работу минимальной.)

Для решения задачи напишем формулы для объема банки, площади ее поверхности и длины швов:

$$\begin{aligned} V &= \pi r^2 h, \\ S &= 2\pi r^2 + 2\pi r h, \\ l &= 4\pi r + h. \end{aligned} \tag{1}$$

Объем банки задан, это устанавливает связь между радиусом r и высотой h. Выразим высоту через радиус: $h = V/\pi r^2$ и подставим полученное выражение в формулы для поверхности и длины швов. В результате будем иметь:

$$S(r) = 2\pi r^2 + 2\frac{V}{r}, \qquad 0 < r < \infty, \tag{2}$$

$$l(r) = 4\pi r + \frac{V}{\pi r^2}, \qquad 0 < r < \infty. \tag{3}$$

Таким образом, с математической точки зрения, задача о наилучшей консервной банке сводится к определению такого значения r, при котором достигает своего наименьшего значения в одном случае функция $S(r)$, в другом — функция $l(r)$.

Вычислим производную функции $S(r)$:

$$S'(r) = 4\pi r - \frac{2V}{r^2} = \frac{2}{r^2}(2\pi r^3 - V) \tag{4}$$

и исследуем ее знак. При $0 < r < r_1 = \sqrt[3]{V/(2\pi)}$ производная отрицательна и функция $S(r)$ убывает, при $r_1 < r < \infty$ производная положительна и функция $S(r)$ возрастает. Следовательно, своего наименьшего значения эта функция достигает в точке $r = r_1$, в которой ее производная обращается в нуль. График функции $S(r)$, иллюстрирующий проведенный анализ, показан на рис. 27.

Итак, радиус и высота банки, наилучшей с точки зрения первого условия, определяются формулами:

$$r_1 = \sqrt[3]{\frac{V}{2\pi}}, \quad h_1 = 2r_1, \tag{5}$$

при этом

$$S(r_1) = 3\sqrt[3]{2\pi V^2} \leqslant S(r). \tag{6}$$

Рассмотрим теперь задачу во второй постановке. Продифференцируем функцию $l(r)$:

$$l'(r) = 4\pi - \frac{2V}{\pi r^3} = \frac{2}{\pi r^3}(2\pi^2 r^3 - V). \tag{7}$$

Как и в предыдущем случае, при $0 < r < r_2 = \sqrt[3]{V/(2\pi^2)}$ производная отрицательна и функция $l(r)$ убывает, при $r_2 < r < \infty$ производная положительна и функция $l(r)$ возрастает. Следовательно, своего наименьшего значения эта функция достигает в точке $r = r_2$, в которой ее производная обращается в нуль. График функции показан на рис. 28.

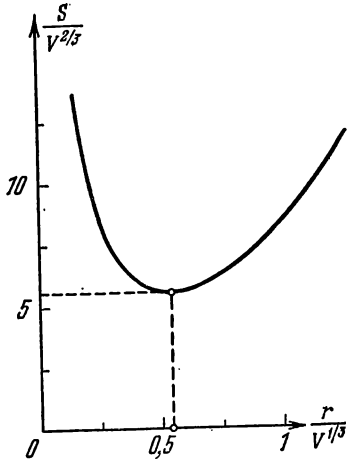

Рис. 27. График функции $S(r)$.

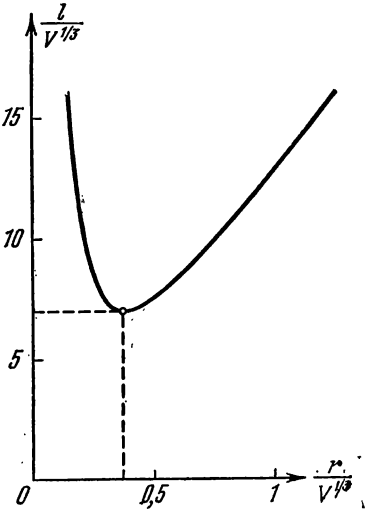

Рис. 28. График функции $l(r)$.

Итак, радиус и высота банки, наилучшей с точки зрения второго условия, определяются формулами:

$$r_2 = \sqrt[3]{\frac{V}{2\pi^2}}, \quad h_2 = 2\pi r_2, \tag{8}$$

при этом

$$l(r_2) = 3\sqrt[3]{4\pi V} \leqslant l(r). \tag{9}$$

Мы видим, что при разных критериях оптимизации получаются существенно разные ответы. В первом случае (5) высота «наилучшей» банки равна ее диаметру, во втором (8) она в π раз больше диаметра.

§ 2. Одномерные задачи оптимизации

После обсуждения простого, но характерного примера, рассмотрим общие вопросы постановки и методов решения одномерных задач оптимизации. С математической точки

зрения такую задачу можно сформулировать следующим образом.

Найти наименьшее (или наибольшее) значение целевой функции $f(x)$, заданной на множестве X. Определить значение переменной $x \in X$, при котором она принимает свое экстремальное значение.

Анализ поставленной задачи начнем с формулировки без доказательства одной теоремы.

Теорема Вейерштрасса. *Всякая функция $f(x)$, непрерывная на отрезке $[a, b]$, принимает на этом отрезке свое наименьшее и наибольшее значения, т. е. на отрезке $[a, b]$ существуют такие точки x_1 и x_2, что для любого $x \in [a, b]$ выполняются неравенства*

$$f(x_1) \leqslant f(x) \leqslant f(x_2). \tag{10}$$

Не исключается, в частности, возможность того, что наименьшее или наибольшее значение достигается сразу в нескольких точках. Вы легко можете убедиться в этом, рассмотрев в качестве примера функцию $y = \sin x$ на отрезке $[0, 4\pi]$. Она достигает своего наименьшего значения, равного -1, сразу в двух точках:

$$x = 3\pi/2 \text{ и } x = 7\pi/2.$$

Наибольшее значение, равное 1, достигается тоже в двух точках:

$$x = \pi/2 \text{ и } x = 5\pi/2.$$

Теорема Вейерштрасса играет в данном случае роль теоремы существования: согласно этой теореме задача оптимизации, в которой целевая функция $f(x)$ задана и непрерывна на отрезке, всегда имеет решение.

Теперь нам предстоит обсудить методы решения задач оптимизации. В этом параграфе мы рассмотрим наиболее простой класс задач, аналогичных задаче о наилучшей консервной банке. При их исследовании мы будем предполагать, что целевая функция $f(x)$ дифференцируема на отрезке $[a, b]$ и имеется возможность найти явное выражение для ее производной $f'(x)$.

Точки, в которых производная обращается в нуль, называются критическими или стационарными точками функции $f(x)$. Если интерпретировать производную как скорость изменения функции, то в критических точках

эта скорость равна нулю, изменение функции на мгновение «останавливается».

Функция $f(x)$ может достигать своего наименьшего (и наибольшего) значения либо в одной из двух граничных точек отрезка $[a, b]$, либо в какой-нибудь его внутренней точке. В последнем случае такая точка обязательно должна быть критической, это необходимое условие экстремума. Учитывая изложенные соображения, мы можем сформулировать следующее правило решения задачи оптимизации для рассматриваемого класса функций.

Для того чтобы определить наименьшее и наибольшее значения дифференцируемой функции $f(x)$ на отрезке $[a, b]$, нужно найти все ее критические точки на данном отрезке, присоединить к ним граничные точки a и b и во всех этих точках сравнить значения функции. Наименьшее и наибольшее из них дадут наименьшее и наибольшее значения функции для всего отрезка.

Поскольку граничные точки a и b искать не нужно, то с технической точки зрения все сводится к определению критических точек, которые являются корнями уравнения

$$f'(x) = 0. \tag{11}$$

Для иллюстрации изложенной выше схемы решения задачи рассмотрим на отрезке $[-2, 3]$ функцию

$$f(x) = 3x^4 - 4x^3 - 12x^2 + 2. \tag{12}$$

Вычислим ее производную:

$$f'(x) = 12x^3 - 12x^2 - 24x.$$

Таким образом, уравнение (11) для определения критических точек в данном случае принимает вид

$$x^3 - x^2 - 2x = 0. \tag{13}$$

Все корни этого уравнения: $x_1 = -1$, $x_2 = 0$, $x_3 = 2$ принадлежат исходному отрезку. Добавляя к ним граничные точки: $a = -2$ и $b = 3$, вычислим соответствующие значения функции (12):

$$f(-2) = 34, \quad f(-1) = -3, \quad f(0) = 2,$$
$$f(2) = -30, \quad f(3) = 29.$$

Из сравнения этих чисел следует, что наименьшее значение функции $f(x)$ достигается в одной из критических точек $x = 2$, а наибольшее — в граничной точке $x = -2$,

причем

$$f_{\min} = f(2) = -30, \quad f_{\max} = f(-2) = 34.$$

График функции (12), иллюстрирующий проведенное исследование, показан на рис. 29.

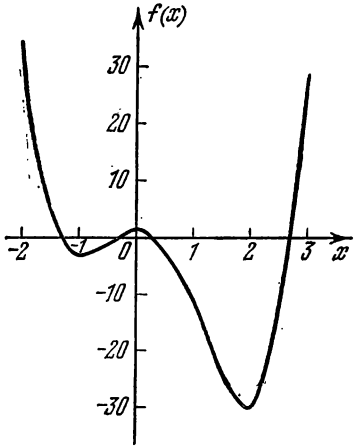

Рис. 29. График функции $(x) = 3x^4 - 4x^3 - 12x^2 + 2$.

В простейших случаях, как в задаче о консервной банке или в рассмотренном примере (12), нули производной удается найти аналитически. На это в первую очередь и рассчитан данный метод, хотя не исключается возможность численного решения уравнения (11). Однако при этом важно найти все критические точки, иначе мы рискуем допустить ошибку, пропустив истинное наименьшее или наибольшее значение функции.

§ 3. Одномерные задачи оптимизации. Продолжение

Рассмотрим следующий пример. Химический завод производит некоторое вещество. Выход интересующего нас продукта определяется температурой: $y = f(T)$. Температуру можно варьировать в определенных пределах: $T_1 \leqslant T \leqslant T_2$. Вид функции f заранее не известен, он зависит от используемого сырья. Получив очередную партию сырья, нужно найти температуру T, при которой наиболее выгодно вести производство, т. е. функция $f(T)$ достигает своего наибольшего значения.

С математической точки зрения мы имеем типичную одномерную задачу оптимизации, сформулированную в начале предыдущего параграфа. В то же время между этой задачей и задачей о консервной банке имеется существенное различие. В данном случае нет никакой формулы для целевой функции $f(T)$. Чтобы определить ее значение при некоторой температуре T, нужно провести опыт либо в лаборатории (если это возможно), либо прямо в производственных условиях. Совершенно ясно, что возможно лишь конечное число измерений и тем самым функция

$f(T)$ будет известна нам в конечном числе точек. Значений ее производной мы вообще определить не можем. Строго говоря, мы даже не знаем, существует ли у нее производная, хотя опыт таких исследований и здравый смысл говорят, что функция $f(T)$, по-видимому, дифференцируема. Нам остается добавить, что каждое измерение требует времени, а задерживать производство нельзя. Поэтому необходимо получить ответ на поставленный вопрос после небольшого числа измерений, т. е. по значениям функции $y = f(T)$ в нескольких точках.

Возможны также задачи оптимизации, в которых целевая функция $y = f(x)$ находится в результате численного решения некоторой математической задачи. Данный случай по своему характеру весьма близок к предыдущему: мы не имеем явной формулы для целевой функции, но можем определить ее значение для любого аргумента $x \in [a, b]$. Ясно, что при этом в ходе решения задачи нам окажется непосредственно доступной информация о целевой функции в конечном числе точек.

Итак, обсудим математические вопросы, связанные со следующей постановкой одномерной задачи оптимизации: определяя значения непрерывной функции $f(x)$ в некотором конечном числе точек отрезка $[a, b]$, нужно приближенно найти ее наименьшее (или наибольшее) значение на данном отрезке.

Возможны разные подходы к решению этой задачи. Рассмотрим сначала метод, идея которого наиболее проста и естественна.

1. Метод равномерного распределения точек по отрезку. Возьмем некоторое целое число n, вычислим шаг $h = (b - a)/n$ и определим значения функции $f(x)$ в точках $x_k = a + kh$, $k = 0, 1, \ldots, n$, $y_k = f(x_k)$. После этого найдем среди полученных чисел наименьшее:

$$m_n = \min(y_0, y_1, \ldots, y_n),$$
$$m_n \geqslant m = \min f(x), \quad x \in [a, b].$$

Число m_n можно приближенно принять за наименьшее значение функции $f(x)$ на отрезке $[a, b]$. Благодаря непрерывности функции $f(x)$ будем иметь:

$$\lim_{n \to \infty} m_n = m,$$

т. е. с увеличением числа точек n ошибка, которую мы допускаем, принимая m_n за m, стремится к нулю.

Какое же n нужно взять, чтобы погрешность в определении наименьшего значения функции $\delta_n = m_n - m$ не превышала заданной точности ε, т. е. чтобы $\delta_n \leqslant \varepsilon$? Это — обычный вопрос, который всегда встает при приближенном решении математических задач. Вспомним, что для метода вилки, благодаря двухсторонней оценке корня, мы легко получили условие достижения точности ε в виде неравенства (24) на стр. 92.

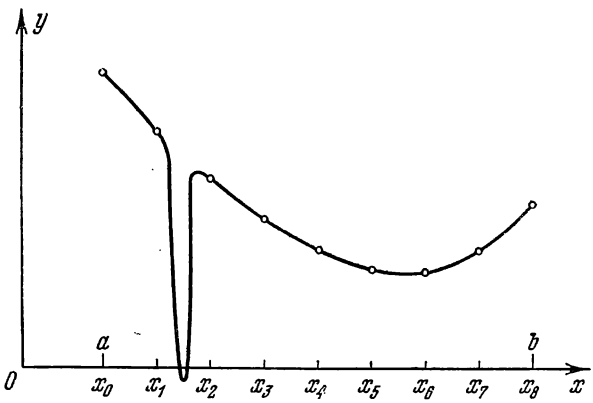

Рис. 30. Пример, иллюстрирующий трудности, которые могут возникнуть при приближенном определении наименьшего значения функции по ее значениям в нескольких точках.

Для данной задачи ситуация оказывается более сложной. Если нам известно только то, что функция $f(x)$ непрерывна на отрезке $[a, b]$, то ответить на поставленный вопрос нельзя. Эта трудность не связана с предложенным способом выбора точек x_k, она носит принципиальный характер. Какое бы n мы ни взяли и как бы ни выбирали n точек на отрезке $[a, b]$, всегда можно указать такую непрерывную функцию, что для нее m_n будет отличаться от m больше чем на ε.

Справедливость этого утверждения иллюстрирует рис. 30. На нем приведен график некоторой непрерывной функции. Допустим, что, желая найти ее наименьшее значение, мы взяли $n = 8$. Точки x_k, $k = 0, 1, \ldots, 8$, показаны на рисунке. Определяя в них значения функции $y_k = f(x_k)$ (они выделены на графике кружочками), получим

$$m_8 = \min(y_0, y_1, \ldots, y_8) = y_6 = f(x_6).$$

Величину $m_8 = y_6$, согласно описанному методу, следует приближенно принять за наименьшее значение функции m.

Представьте, что у нас нет перед глазами рис. 30 (как это предполагает постановка задачи), а известны только 9 чисел y_k. По ним невозможно установить, что функция $f(x)$ имеет между точками x_1 и x_2 узкий «язык», который опускается гораздо ниже $y_6 = m_8$. Из-за небольшого числа точек мы его пропустим. Если взять n побольше, то данный «язык» обнаружится, но может оказаться незамеченным какой-нибудь другой еще более тонкий «язык». При отсутствии дополнительной информации о свойствах функции $f(x)$, о том, насколько «резкими» могут быть ее изменения, сомнения останутся, какое бы большое число точек мы ни взяли.

Дать строго обоснованную оценку числа точек n, необходимого для решения задачи с точностью ε ($\delta_n \leqslant \varepsilon$), можно только, сужая класс рассматриваемых функций. Предположим, например, что функция $y = f(x)$ не просто непрерывна, а удовлетворяет условию Липшица с известной постоянной α, тогда легко получить следующее неравенство для нужного числа точек n в данном методе:

$$n \geqslant \frac{\alpha}{\varepsilon}(b - a). \qquad (14)$$

Однако неравенство (14) — это пиррова победа: при отсутствии явной формулы для функции $f(x)$ такой результат малоэффективен. Он может быть использован, если имеется априорная информация о величине постоянной Липшица α. В противном случае предположение о справедливости для целевой функции условия Липшица и оценка константы α, входящей в неравенство (14), носит характер гипотезы, которую, как правило, невозможно проверить. (Подумайте сами, как это сделать в задаче о химическом производстве, если каждое значение целевой функции $f(T)$ получается в результате экспериментальных измерений, а условие Липшица, с точки зрения его математического определения, нужно проверять для любой пары аргументов из рассматриваемой области.) Поэтому при решении вопроса о числе точек и точности важно максимально полно использовать всю дополнительную информацию о свойствах целевой функции, о степени ее гладкости, вытекающую из характера и особенностей задачи. Не последнюю роль играет и такой фактор, как опыт, интуиция исследователя.

2. Метод распределения точек по отрезку, учитывающий результаты вычисления целевой функции. Для метода, который был описан, характерно равномерное распределение «пробных» точек x_k по отрезку $[a, b]$. Их положение строго фиксировано заранее и никак не зависит от информации о функции $f(x)$, которую мы получаем по мере вычисления чисел $y_k = f(x_k)$. Такой подход дает возможность внимательно просмотреть весь отрезок. Он помогает наиболее надежно избежать пропуска какого-нибудь узкого, глубокого «языка» у целевой функции. В этом бесспорное достоинство метода.

Однако недаром говорят, что наши недостатки — это продолжение наших достоинств. Распределяя точки x_k равномерно, мы уделяем одинаковое внимание всем участкам отрезка: тем, где целевая функция велика, и тем, в направлении которых она убывает. Это существенно удлиняет расчеты и затягивает исследование.

Организацию вычислений по такой схеме можно сравнить с поведением в лесу неопытного грибника. В поисках грибов он ходит по всему лесу, не чувствуя разницы между грибными и негрибными местами. В результате ему приходится тратить много сил и времени понапрасну на осмотр негрибных мест. Совершенно иначе ведет себя опытный грибник. Он подолгу задерживается на грибных местах и осматривает их особенно внимательно, а через негрибные места проходит быстро, не тратя на них лишнего времени.

Чтобы организовать поиск наименьшего значения функции по методу «опытного грибника», нужно отказаться от предписанного заранее выбора пробных точек x_k, а выбирать очередную точку с учетом информации, которую мы уже получили о функции $f(x)$ в результате ее вычисления в предыдущих точках. При этом основное внимание следует уделять тем участкам отрезка $[a, b]$, где вычисления дают малые значения функции, просматривая другие участки более бегло. Реализовать эту идею можно, например, следующим образом.

Вычислим значения функции $f(x)$ в двух граничных точках $x_0 = a$ и $x_1 = b$: $y_0 = f(x_0)$, $y_1 = f(x_1)$. После этого следующую точку x_2 выберем ближе к тому концу отрезка, на котором функция принимает меньшее значение. Ее положение определим соотношением между числами y_0 и y_1: чем больше разница между ними, тем сильнее будет сдвиг точки x_2 в соответствующую сторону. Точку

x_3 выберем с учетом взаимного расположения точек x_0, x_1, x_2 и соотношения между числами y_0, y_1, y_2 и т. д. Мы не будем останавливаться на описании возможных алгоритмов выбора очередной точки x_k по информации, полученной в результате вычисления целевой функции на предыдущих шагах,— это специальный вопрос. Отметим лишь, что исследования в данной области продолжаются, алгоритмы совершенствуются, и пока рано говорить об окончательном решении задачи.

3. Специальные методы. Воспользуемся еще раз аналогией со сбором грибов, чтобы поставить новые вопросы о методах решения задачи оптимизации. Грибник может попасть в данный лес впервые и ничего не знать о нем заранее, кроме одного: грибы есть. Возможен и другой случай. Человек приходит в лес, который он уже немного знает. Его поведение в первом и втором случае будет различным. Причем, если он сумеет правильно воспользоваться известными ему особенностями леса, то наполнит корзину грибами гораздо быстрее.

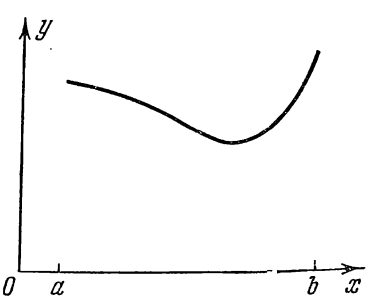

Рис. 31. Пример функции, имеющей один минимум.

До сих пор, обсуждая задачи оптимизации, мы говорили об универсальных методах их решения. Однако во многих случаях из характера задачи вытекает какая-то дополнительная информация о свойствах целевой функции. Это может быть использовано для разработки специальных алгоритмов. Такой подход позволяет существенно сократить объем вычислений и получить ответ наиболее эффективным способом.

В качестве примера рассмотрим случай, когда нам известно заранее, что целевая функция $y = f(x)$ имеет на отрезке $[a, b]$ только один минимум. График такой функции показан на рис. 31.

Для решения задачи в этом случае можно воспользоваться следующим методом. Возьмем некоторый шаг h и будем последовательно вычислять значения функции $f(x)$ в точках $x_0 = a$, $x_1 = a + h$, $x_2 = a + 2h$, ..., сравнивая получаемые числа $y_0, y_1, y_2, \ldots$. Сначала они будут убывать: $y_0 > y_1 > y_2 > \ldots$, однако в дальнейшем

найдется такая точка $x_k = a + kh$, что для значения функции в ней $y_k = f(x_k)$ справедливы неравенства: $y_{k-1} > y_k$, $y_{k+1} \geqslant y_k$. Это означает, что наименьшее значение функции достигается на отрезке $[x_{k-1}, x_{k+1}]$ и его приближенно можно принять равным $y_k = f(x_k)$. Если требуемая точность в решении задачи еще не обеспечена, то нужно уменьшить шаг h и повторить описанную процедуру для отрезка $[x_{k-1}, x_{k+1}]$.

Задача об оптимальной температуре для химического процесса часто относится к задачам подобного типа. График функции $f(T)$ для многих химических реакций имеет вид кривой на рис. 31, но только перевернутой: при увеличении температуры T функция сначала возрастает, а потом, пройдя через максимум, начинает убывать. Нам нужно найти этот максимум, для чего можно воспользоваться описанной выше процедурой. Она потребует небольшого числа измерений функции $f(T)$. То, что мы ищем максимум, а не минимум, не имеет принципиального значения с точки зрения метода, просто все неравенства изменят свои знаки на противоположные.

§ 4. Частные производные и градиент функции нескольких переменных

В следующем параграфе мы перейдем к обсуждению многомерных задач оптимизации. Для их анализа нам нужно познакомиться с основными понятиями дифференциального исчисления функций нескольких переменных.

Рассмотрим, для определенности, функцию двух переменных $u = f(x, y)$. Фиксируем значение переменной y, тогда наша функция будет зависеть только от x. Мы можем вычислить производную этой функции по известным вам правилам. Такую производную по x при фиксированном значении y называют частной производной и обозначают символом

$$\frac{\partial u}{\partial x} = \frac{\partial f}{\partial x}(x, y).$$

По определению

$$\frac{\partial u}{\partial x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x}. \qquad (15)$$

Подчеркнем еще раз, что в данном случае при вычислении приращения функции $\Delta f = f(x + \Delta x, y) - f(x, y)$ изменяется только переменная x, а переменная y фиксирована.

Аналогично вводится частная производная по y. В этом случае фиксируется переменная x, а дифференцирование проводится по переменной y. Таким образом,

$$\frac{\partial u}{\partial y} = \lim_{\Delta y \to 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y}. \qquad (16)$$

Частные производные характеризуют изменения функции $u = f(x, y)$ по каждой из независимых переменных в отдельности. Задача их вычисления с технической точки зрения не отличается от известной вам задачи дифференцирования функции одной переменной.

С помощью частных производных (15) и (16) можно построить вектор, который называют градиентом функции:

$$\operatorname{grad} f(x, y) = \frac{\partial f}{\partial x}(x, y)\, \boldsymbol{i} + \frac{\partial f}{\partial y}(x, y)\, \boldsymbol{j}, \qquad (17)$$

здесь $\boldsymbol{i}$ и $\boldsymbol{j}$ — единичные векторы, параллельные координатным осям x и y. Градиент дает общую характеристику поведения функции f в окрестности рассматриваемой точки (x, y). Его направление является направлением наибыстрейшего возрастания функции. Противоположное направление, которое часто называют антиградиентным, представляет собой направление наибыстрейшего убывания функции. Модуль градиента

$$|\operatorname{grad} f(x, y)| = \sqrt{\left(\frac{\partial f}{\partial x}(x, y)\right)^2 + \left(\frac{\partial f}{\partial y}(x, y)\right)^2} \qquad (18)$$

определяет скорость возрастания и убывания функции в точке (x, y) в направлении ее градиента и антиградиента. Для всех остальных направлений скорость изменения функции меньше модуля градиента $|\operatorname{grad} f(x, y)|$.

Заканчивая обсуждение понятия градиента функции, подчеркнем, что он сам является функцией точки, так что при переходе от одной точки к другой может менять свою величину и направление.

Для геометрической интерпретации функции двух переменных часто строят на плоскости x, y ее линии уровня. Так называют линии, вдоль которых функция сохраняет постоянное значение:

$$f(x, y) = C. \qquad (19)$$

Меняя C, мы получим систему линий уровня, дающую достаточно полное представление о функции $f(x, y)$. Вы

встречались с линиями уровня на географических картах и планах, где их проводят через точки, находящиеся на одинаковой высоте над уровнем моря. Отсюда и произошло название этих линий.

Пусть мы имеем функцию $u = f(x, y)$. Вычислим ее градиент в некоторой точке (x, y) и рассмотрим линию уровня, проходящую через эту точку. Они будут взаимно перпендикулярны: градиент определяет направление наибыстрейшего изменения функции, а вдоль линии уровня ее значения вообще не меняются.

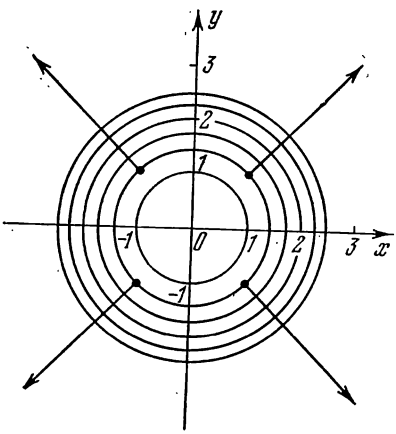

Рис. 32. Линии уровня и градиент функции $u = x^2 + y^2$.

Заканчивая обсуждение этих вопросов, возьмем в качестве примера функцию

$$u = x^2 + y^2. \qquad (20)$$

Вычислим ее градиент:

$$\operatorname{grad} u\,(x, y) = 2x\,\boldsymbol{i} + 2y\boldsymbol{j}. \qquad (21)$$

Эта формула определяет градиент в произвольной точке (x, y). Например, для точки $x = 1$, $y = -1$ будем иметь:

$$\operatorname{grad} u\,(1, -1) = 2\boldsymbol{i} - 2\boldsymbol{j}.$$

Модуль градиента (21) находится по формуле

$$|\operatorname{grad} u\,(x, y)| = 2\sqrt{x^2 + y^2}, \qquad (22)$$

он равен удвоенному расстоянию от начала координат до рассматриваемой точки (x, y). Направление градиента (21) в каждой точке совпадает с направлением радиуса-вектора, который проведен из начала координат в эту точку.

Уравнение линий уровня для функции (20) имеет вид

$$x^2 + y^2 = C, \qquad (23)$$

оно определяет семейство окружностей радиуса $r = \sqrt{C}$ $(C \geqslant 0)$.

На рис. 32 показаны линии уровня функции (20), соответствующие $C = 1, 2, 3, 4, 5, 6$. Там же в четырех точ-

ках: (1, 1), (—1, 1), (—1, —1), (1, —1) построены векторы grad u (x, y). Они направлены по радиусам и, тем самым, перпендикулярны окружности радиуса $\sqrt{2}$, которая является линией уровня, проходящей через эти точки. По мере удаления от начала координат модуль градиента (22) растет, т. е. изменение функции (20) вдоль радиуса (в направлении градиента) становится все более резким. На рисунке это проявляется в сгущении линий уровня при увеличении C.

Введенные выше понятия переносятся без всяких изменений на функции многих переменных, хотя их наглядная геометрическая интерпретация становится при этом затруднительной.

§ 5. Многомерные задачи оптимизации

До сих пор мы обсуждали одномерные задачи оптимизации, в которых целевая функция зависела только от одного аргумента. Однако подавляющее число реальных задач оптимизации, представляющих практический интерес, являются многомерными: в них целевая функция зависит от нескольких аргументов, причем иногда их число может быть весьма большим.

Вспомним, например, задачу о химическом производстве. Мы отметили, что в ней целевая функция зависит от температуры, и при определенном ее выборе производительность (выход интересующего нас продукта) оказывается максимальной. Однако, наряду с температурой, производительность зависит также от давления, соотношения между концентрациями вводимого сырья, катализаторов и ряда других факторов. Таким образом, задача выбора наилучших условий химического производства — это типичная многомерная задача оптимизации.

Математическая постановка таких задач аналогична их постановке в одномерном случае: ищется наименьшее (или наибольшее) значение целевой функции, заданной на некотором множестве E возможных значений ее аргументов. В случае, когда целевая функция непрерывна, а множество E является замкнутой ограниченной областью, остается справедливой теорема Вейерштрасса. Тем самым выделяется класс задач оптимизации, для которых гарантировано существование решения. В дальнейшем мы всегда будем предполагать, не оговаривая этого особо, что все рассматриваемые задачи принадлежат этому классу.

Как и в одномерном случае, характер задачи, и соответственно, возможные методы решения существенно зависят от той информации о целевой функции, которая нам доступна в процессе ее исследования. В одних случаях целевая функция задается аналитической формулой, являясь при этом дифференцируемой функцией. Тогда можно вычислить ее частные производные, получить явное выражение для градиента, определяющего в каждой точке направления возрастания и убывания функции, и использовать эту информацию для решения задачи. В других случаях никакой формулы для целевой функции нет, а имеется лишь возможность определить ее значение в любой точке рассматриваемой области (с помощью расчетов, в результате эксперимента и т. д.). В таких задачах в процессе решения мы фактически можем найти значения целевой функции лишь в конечном числе точек, и по этой информации требуется приближенно установить ее наименьшее значение для всей области.

Многомерные задачи, естественно, являются более сложными и трудоемкими, чем одномерные, причем обычно трудности при их решении возрастают при увеличении размерности. Для того чтобы вы лучше почувствовали это, возьмем самый простой по своей идее приближенный метод поиска наименьшего значения функции, который уже обсуждался для одномерных задач в третьем параграфе. Покроем рассматриваемую область сеткой с шагом h (см. рис. 33 для двумерного случая) и определим значения функции в ее узлах. Сравнивая полученные числа между собой, найдем среди них наименьшее и примем его приближенно за наименьшее значение функции для всей области.

Как мы уже говорили выше, данный метод используется для решения одномерных задач. Иногда он применяется также для решения двумерных, реже трехмерных задач. Однако для задач большей размерности он практически непригоден из-за слишком большого времени, необходимого для проведения расчетов.

Действительно, предположим, что целевая функция зависит от пяти переменных, а область определения является «пятимерным» кубом, каждую сторону которого при построении сетки мы делим на 40 частей. Тогда общее число узлов сетки будет равно $41^5 \approx 10^8$. Пусть вычисление значения функции в одной точке требует 1000 арифметических операций (это немного для функции пяти пере-

менных). В таком случае общее число операций составит 10^{11}. Если в нашем распоряжении имеется ЭВМ с быстродействием миллион операций в секунду, то для решения задачи с помощью данного метода потребуется 10^5 секунд, что превышает сутки непрерывной работы. Добавление еще одной независимой переменной увеличит это время в 40 раз.

Проведенная оценка показывает, что для больших задач оптимизации метод сплошного перебора не годится.

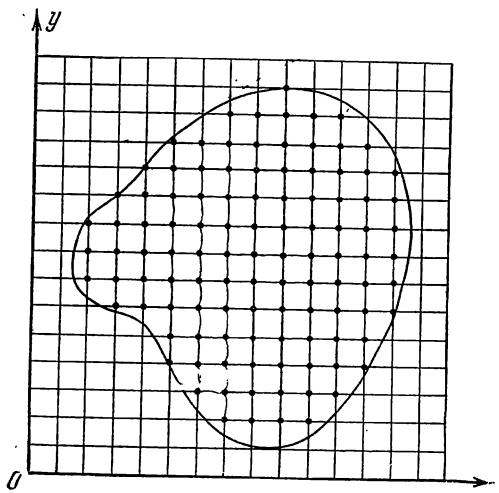

Рис. 33. Построение сетки с шагом h и выбор «пробных» точек в узлах сетки для приближенного определения наименьшего значения функции двух переменных.

Иногда сплошной перебор заменяют случайным поиском. В этом случае точки сетки просматриваются не подряд, а в случайном порядке. В результате поиск наименьшего значения целевой функции существенно ускоряется, но теряет свою надежность.

Перейдем теперь к обсуждению методов, позволяющих вести поиск наименьшего значения функции целенаправленно.

1. Метод покоординатного спуска. Пусть нужно найти наименьшее значение целевой функции $u = f(M) = f(x_1, x_2, x_3, \ldots, x_n)$. Здесь через M сокращенно обозначена точка n-мерного пространства с координатами $x_1, x_2, x_3, \ldots, x_n$: $M = (x_1, x_2, x_3, \ldots, x_n)$. Выберем

какую-нибудь начальную точку $M_0 = (x_{1,0}, x_{2,0}, x_{3,0}, \ldots, x_{n,0})$ и рассмотрим функцию f при фиксированных значениях всех переменных, кроме первой: $f(x_1, x_{2,0}, x_{3,0}, \ldots, x_{n,0})$. Тогда она превратится в функцию одной переменной x_1. Изменяя эту переменную, будем двигаться от начальной точки $x_1 = x_{1,0}$ в сторону убывания функции, пока не дойдем до ее минимума при $x_1 = x_{1,1}$, после которого она начинает возрастать. Точку с координатами $(x_{1,1}, x_{2,0}, x_{3,0}, \ldots, x_{n,0})$ обозначим через M_1, при этом $f(M_0) \geqslant f(M_1)$.

Фиксируем теперь переменные: $x_1 = x_{1,1}, x_3 = x_{3,0}, \ldots, x_n = x_{n,0}$ и рассмотрим функцию f как функцию одной переменной x_2: $f(x_{1,1}, x_2, x_{3,0}, \ldots, x_{n,0})$. Изменяя x_2, будем опять двигаться от начального значения $x_2 = x_{2,0}$ в сторону убывания функции, пока не дойдем до минимума при $x_2 = x_{2,1}$. Точку с координатами $(x_{1,1}, x_{2,1}, x_{3,0}, \ldots, x_{n,0})$ обозначим через M_2, при этом $f(M_1) \geqslant f(M_2)$.

Проведем такую же минимизацию целевой функции по переменным $x_3, x_4, \ldots, x_n$. Дойдя до конца, снова вернемся к переменной x_1 и продолжим процесс. Эта процедура вполне оправдывает название метода. С ее помощью мы построим последовательность точек $M_0, M_1, M_2, \ldots$, которой соответствует монотонная последовательность значений функции $f(M_0) \geqslant f(M_1) \geqslant f(M_2) \geqslant \ldots$ Обрывая ее на некотором шаге k, можно приближенно принять значение функции $f(M_k)$ за ее наименьшее значение в рассматриваемой области.

Отметим, что данный метод сводит задачу поиска наименьшего значения функции нескольких переменных к многократному решению одномерных задач оптимизации. Если целевая функция $f(x_1, x_2, \ldots, x_n)$ задана явной формулой и является дифференцируемой, то мы можем вычислить ее частные производные и использовать их для определения направления убывания функции по каждой переменной и поиска соответствующих одномерных минимумов. В противном случае, когда явной формулы для целевой функции нет, одномерные задачи следует решать с помощью методов, которые описаны в § 3.

На рис. 34 нарисованы линии уровня некоторой функции двух переменных $u = f(x, y)$ и показана траектория поиска ее наименьшего значения, которое достигается в точке O, с помощью метода покоординатного спуска. При этом вы должны ясно понимать, что рисунок служит только для иллюстрации метода. Когда мы приступаем к ре-

шению реальной задачи оптимизации, такой картинки, содержащей в себе готовый ответ, у нас, конечно, нет.

2. Метод градиентного спуска. Направления, параллельные координатным осям, по которым мы двигались в предыдущем методе, не являются, как правило, направлениями наиболее быстрого убывания функции. Таким направлением является, как мы знаем, направление антиградиента. Это учитывает метод градиентного спуска, который заключается в следующем.

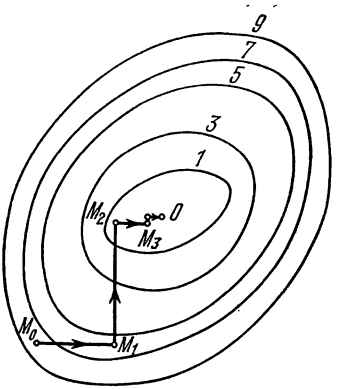

Рис. 34. Поиск наименьшего значения функции методом покоординатного спуска.

Выберем каким-либо способом начальную точку, вычислим в ней градиент рассматриваемой функции и сделаем небольшой шаг в обратном антиградиентном направлении. В результате мы придем в точку, в которой значение функции будет меньше первоначального. В новой точке повторим процедуру: снова вычислим градиент функции и сделаем шаг в обратном направлении. Продолжая этот процесс, мы будем двигаться в сторону убывания функции. Специальный выбор направления движения на каждом шаге позволяет надеяться на то, что в данном случае приближение к наименьшему значению функции будет более быстрым, чем в методе покоординатного спуска.

Метод градиентного спуска требует вычисления градиента целевой функции на каждом шаге. Если она задана аналитически, то это, как правило, не проблема: для частных производных, определяющих градиент, можно получить явные формулы. В противном случае частные производные в нужных точках приходится вычислять приближенно, заменяя их соответствующими разностными отношениями:

$$\frac{\partial f}{\partial x_i} \approx \frac{f(x_1, \ldots, x_i + \Delta x_i, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{\Delta x_i}. \tag{24}$$

Отметим, что при таких расчетах Δx_i нельзя брать слишком малым, а значения функции нужно вычислять с достаточно высокой степенью точности, иначе при

вычислении разности

$$\Delta f = f(x_1, \ldots, x_i + \Delta x_i, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)$$

будет допущена большая ошибка.

На рис. 35 нарисованы линии уровня той же функции двух переменных $u = f(x, y)$, что и на рис. 34, и приведена траектория поиска ее минимума с помощью метода градиентного спуска. Сравнение рис. 34 и 35 показывает,

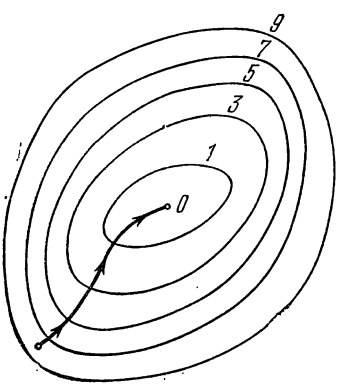

Рис. 35. Поиск наименьшего значения функции методом градиентного спуска.

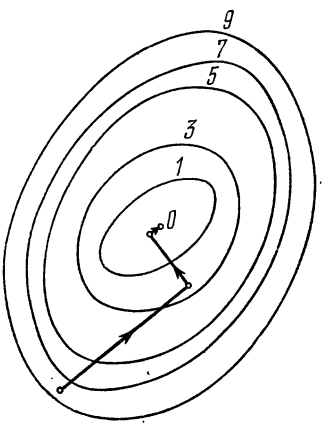

Рис. 36. Поиск наименьшего значения функции методом наибыстрейшего спуска.

насколько более эффективным является метод градиентного спуска.

3. Метод наибыстрейшего спуска. Мы уже приводили в этой главе известный афоризм: «Наши недостатки — это продолжение наших достоинств». Вычисление градиента на каждом шаге, позволяющее все время двигаться в направлении наибыстрейшего убывания целевой функции, может в то же время замедлить вычислительный процесс. Дело в том, что подсчет градиента — обычно гораздо более сложная операция, чем подсчет самой функции. Поэтому часто пользуются модификацией градиентного метода, получившей название метода наибыстрейшего спуска.

Согласно этому методу после вычисления в начальной точке градиента функции делают в направлении антиградиента не маленький шаг, а двигаются до тех пор, пока

функция убывает. Достигнув точки минимума на выбранном направлении, снова вычисляют градиент функции и повторяют описанную процедуру. При этом градиент вычисляется гораздо реже, только при смене направлений движения.

На рис. 36 показаны траектория поиска наименьшего значения целевой функции по методу наибыстрейшего спуска. Функция выбрана та же, что и на рис. 34, 35. Хотя траектория ведет к цели не так быстро, как на рис. 35, экономия машинного времени за счет более редкого вычисления градиента может быть весьма существенной.

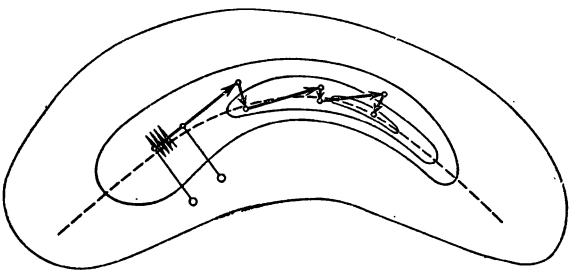

Рис. 37. Поиск наименьшего значения функции в случае «оврага».

4. Проблема «оврагов». Мы рассказали о трех вариантах методов спуска и показали на рис. 34—36, как хорошо они работают. В результате у вас могло сложиться впечатление, что проблема решена. На самом деле это не так. Все было хорошо потому, что был выбран «удобный» пример. Но посмотрите на рис. 37. На нем также показаны линии уровня некоторой функции, однако их конфигурация отличается от рис. 34—36. Линии уровня сильно вытянуты в одном направлении и сплющены в другом. Они напоминают рельеф местности с оврагом. Этот случай крайне неудобен для описанных выше методов.

Действительно, попытаемся найти наименьшее значение такой функции с помощью градиентного спуска. Двигаясь все время в направлении антиградиента, мы быстро спустимся на дно «оврага», и, поскольку движение идет хотя и маленькими, но конечными шагами, проскочим его. Оказавшись на противоположной стороне «оврага» и вычислив там градиент функции, мы будем вынуждены развернуться почти на 180° и сделать один или несколько шагов в обратном направлении. При этом мы снова проскочим

дно «оврага» и вернемся на его первоначальную сторону. Продолжая этот процесс, мы вместо того, чтобы двигаться по дну оврага в сторону его понижения, будем совершать зигзагообразные скачки поперек оврага, почти не приближаясь к цели. Таким образом, в случае «оврага» (этот нематематический термин прочно закрепился в литературе) описанные выше методы спуска оказываются неэффективными.

Для борьбы с «оврагами» был предложен ряд специальных приемов. Один из них заключается в следующем. Из двух близких точек совершают градиентный спуск на дно «оврага». Потом соединяют найденные точки прямой и делают вдоль нее большой (овражный) шаг. Из найденной точки снова спускаются на дно «оврага» и делают второй овражный шаг. В результате, двигаясь достаточно быстро вдоль «оврага» приближаемся к искомому наименьшему значению целевой функции (см. рис. 37). Такой метод достаточно хорошо работает для функций двух переменных, однако при большем числе переменных могут возникнуть трудности.

Все описанные выше методы приспособлены к случаю, когда наименьшее значение функции достигается внутри рассматриваемой области, и становятся малоэффективными, если наименьшее значение достигается на границе или вблизи нее. Для решения таких задач приходится разрабатывать специальные методы. Мы не будем на них останавливаться. Вам должно быть и без того ясно, что большое число специальных методов — это признак слабости, а не силы. Ведь, приступая к решению практической задачи, мы, как правило, не знаем всех ее особенностей и не можем сразу выбрать наиболее эффективный метод.

5. Проблема многоэкстремальности. Посмотрите на рис. 34—37 и сравните их с рис. 38. Первые четыре рисунка относятся к функциям, имеющим только один минимум. Поэтому, откуда бы ни начинался поиск, мы придем в конце концов к нужной точке. На рис. 38 приведены линии уровня функции с двумя «локальными» минимумами в точках O_1 и O_2. Такие функции принято называть многоэкстремальными. Сравнивая между собой значения функции в точках O_1 и O_2: $f_1 = 3$, $f_2 = 1$, находим, что наименьшее значение функция достигает в точке O_2.

Представьте себе теперь, что, не имея перед глазами рис. 38 и не зная о многоэкстремальности функции, мы начали поиск наименьшего значения с помощью метода

градиентного спуска из точки A_1. Поиск приведет нас в точку O_1, которую ошибочно можно принять за искомый ответ. С другой стороны, если мы начнем поиск с точки A_2, то окажемся на правильном пути и быстро придем в точку O_2.

Как бороться с многоэкстремальностью? Универсального ответа на этот вопрос нет. Самый простой прием состоит в том, что проводят поиск несколько раз, начиная

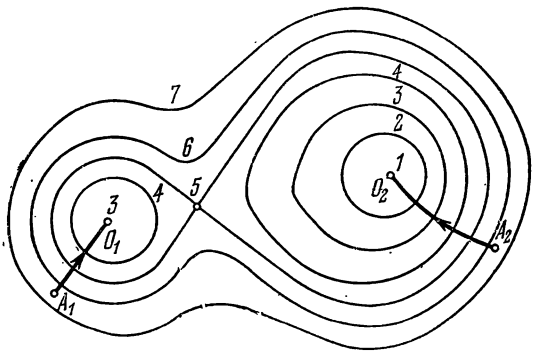

Рис. 38. Пример функции с двумя «локальными» минимумами в точках O_1 и O_2.

сго с разных точек. Если при этом получаются разные ответы, то сравнивают в них значения целевой функции и выбирают наименьшее. Расчеты останавливают после того, как несколько новых поисков не меняют полученного ранее результата. Выбор начальных точек поиска, обоснованность прекращения расчетов в значительной степени зависят от опыта и интуиции специалистов, решающих задачу.

Нарисованная картина может показаться слишком мрачной: сплошное гадание на кофейной гуще. На самом деле во многих случаях имеется различная дополнительная информация о характере задачи, которая существенно помогает при выборе метода, начальной точки поиска и т. д. Кроме того, пока мы не делали никаких предположений о специальных свойствах целевой функции и о характере рассматриваемой области. Это затрудняет анализ. Конкретизация задачи, выделение определенных классов функций и областей позволяет провести более глубокое иссле-

дование и разработать специальные методы, которые решают задачу исчерпывающим образом.

Важнейшим классом таких «специальных» задач оптимизации являются задачи линейного программирования. Вы немного знакомы с ними из школьной программы по математике. Постановка и методы решения этих задач разбираются в следующей главе. Существуют и другие типы задач оптимизации, конкретные специфические особенности которых позволяют провести их детальный анализ и разработать эффективные методы решения. Однако на них мы останавливаться не будем.

Глава 6
ЛИНЕЙНОЕ ПРОГРАММИРОВАНИЕ

В этой главе мы познакомимся с линейным программированием. Так называются задачи оптимизации, в которых целевая функция является линейной функцией независимых переменных, а условия, определяющие допустимые значения этих переменных, имеют вид линейных уравнений и неравенств. Конкретизация характера задачи позволяет провести ее полный анализ и разработать эффективные методы решения.

Линейное программирование развилось в первую очередь в связи с задачами экономики, с поиском способов оптимального распределения и использования ограниченных ресурсов. Оно послужило основой широкого использования математических методов в экономике. Следует подчеркнуть, что в реальных экономических задачах число независимых переменных обычно бывает очень большим (тысячи, десятки тысяч аргументов). Поэтому практическая реализация алгоритмов решения таких задач принципиально невозможна без использования современной вычислительной техники.

§ 1. Если бы директором был я...

Мы приглашаем вас занять директорский кабинет и рассмотреть некоторые производственно-экономические задачи.

Транспортная задача. В городе имеются два склада муки и два хлебозавода. Ежедневно с первого склада вывозится 50 тонн муки, а со второго — 70 тонн. Эта мука доставляется на хлебозаводы, причем первый завод получает 40 тонн, второй 80 тонн.

Вы являетесь директором автобазы, которой поручено выполнять эти перевозки. Вы можете, в рамках установленных норм, возить муку с любого склада на любой завод. Допустим, что перевозка одной тонны муки с первого

склада на первый завод стоит 1 рубль 20 копеек, с первого склада на второй завод — 1 рубль 60 копеек, со второго склада на первый завод — 80 копеек и со второго склада на второй завод — 1 рубль. Как нужно спланировать перевозки, чтобы их стоимость была минимальной?

Для того, чтобы ответить на этот вопрос, придадим задаче математическую формулировку. Обозначим через x_1 и x_2 количество муки, которую следует перевезти с первого склада на первый и второй заводы соответственно, а через x_3 и x_4 — количество муки, которую нужно перевезти со второго склада на первый и второй заводы. Числа x_i ($i = 1, 2, 3, 4$) должны удовлетворять следующим условиям:

$$\begin{aligned} x_1 + x_2 &= 50, \\ x_3 + x_4 &= 70, \\ x_1 + x_3 &= 40, \\ x_2 + x_4 &= 80, \end{aligned} \tag{1}$$

$$x_i \geqslant 0, \qquad i = 1, 2, 3, 4. \tag{2}$$

Первые два уравнения системы (1) определяют, сколько муки нужно вывести с каждого склада, два других уравнения показывают, сколько муки нужно привести на каждый завод. Неравенства (2) означают, что в обратном направлении с заводов на склады муку не возят.

Общая стоимость всех перевозок определяется формулой

$$f = 1{,}2x_1 + 1{,}6x_2 + 0{,}8x_3 + x_4. \tag{3}$$

С математической точки зрения задача заключается в том, чтобы найти четыре числа x_i, $i = 1, 2, 3, 4$, удовлетворяющие условиям (1) и (2) и минимизирующие стоимость перевозок (3).

Сформулированная задача является типичной задачей линейного программирования. В этих задачах ищется наименьшее или наибольшее значение некоторой линейной функции (в нашем примере это функция (3)) при условии, что входящие в нее переменные подчинены системе ограничений в виде линейных уравнений и неравенств (у нас — это условия (1) и (2)). В следующем разделе мы обсудим общую постановку и методы решения таких задач, а пока обратимся к задаче, которая стоит перед вами, как директором автобазы.

Рассмотрим систему (1). Это система четырех уравнений с четырьмя неизвестными. Однако независимыми в ней являются только первые три уравнения, четвертое — их следствие (сложите уравнения (1) и (2), а затем вычтите (3), и вы получите уравнение (4)). Таким образом, фактически нужно рассмотреть следующую систему, эквивалентную (1):

$$\begin{aligned} x_1 + x_2 &= 50, \\ x_3 + x_4 &= 70, \\ x_1 + x_3 &= 40. \end{aligned} \tag{4}$$

В ней число уравнений на единицу меньше числа неизвестных, так что мы можем выбрать какую-нибудь неизвестную, например, x_1, и выразить через нее с помощью уравнений (4) три остальных. Соответствующие формулы имеют вид

$$\begin{aligned} x_2 &= 50 - x_1, \\ x_3 &= 40 - x_1, \\ x_4 &= 30 + x_1. \end{aligned} \tag{5}$$

Согласно (2) все x_i, $i = 1, 2, 3, 4$, должны быть неотрицательны. Это дает систему неравенств:

$$\begin{aligned} x_1 &\geqslant 0, \\ 0 - x_1 &\geqslant 0, \\ 40 - x_1 &\geqslant 0, \\ 30 + x_1 &\geqslant 0, \end{aligned} \tag{6}$$

из которой получаем

$$0 \leqslant x_1 \leqslant 40. \tag{7}$$

Таким образом, задавая любое x_1, удовлетворяющее условию (7), и вычисляя x_2, x_3, x_4 по формулам (5), мы получим один из возможных планов перевозки. При реализации этого плана с каждого склада будет вывезено и на каждый завод доставлено нужное количество муки.

Вычислим стоимость перевозок. Для этого подставим выражения (5) в формулу (3). В результате будем иметь:

$$f = 142 - 0{,}2\, x_1. \tag{8}$$

Эта формула определяет величину f как функцию одной переменной x_1, которую можно выбирать произвольно в пределах условий (7). Стоимость окажется минимальной, если мы придадим величине x_1 наибольшее возможное

значение: $x_1 = 40$. Значения остальных величин x_i находятся при этом по формулам (5).

Итак, оптимальный по стоимости план перевозок имеет вид

$$\begin{aligned} x_1 &= 40, \\ x_2 &= 10, \\ x_3 &= 0, \\ x_4 &= 70. \end{aligned} \tag{9}$$

Стоимость перевозок в этом случае составляет 134 рубля. При любом другом допустимом плане перевозок она окажется выше: $f > f_{\min} = 134$ руб.

Задача об использовании ресурсов. Теперь мы попросим вас оставить кресло директора автобазы и стать директором мебельной фабрики. Возглавляемая вами фабрика выпускает стулья двух типов. На изготовление одного стула первого типа, стоящего 8 рублей, расходуется 2 метра досок стандартного сечения, 0,5 квадратного метра обивочной ткани и 2 человеко-часа рабочего времени. Аналогичные данные для стульев второго типа даются цифрами: 12 рублей, 4 метра, 0,25 квадратного метра и 2,5 человеко-часа.

Допустим, что в вашем распоряжении имеются: 440 метров досок, 65 квадратных метров обивочной ткани, 320 человеко-часов рабочего времени. Какие стулья и в каком количестве вы прикажете выпускать, чтобы стоимость продукции была максимальной?

Для ответа на этот вопрос постараемся опять сформулировать задачу как математическую. Обозначим через x_1 и x_2 запланированное к производству число стульев первого и второго типа. Ограниченный запас сырья и трудовых ресурсов означает, что числа x_1 и x_2 должны удовлетворять неравенствам:

$$\begin{aligned} 2x_1 + 4x_2 &\leqslant 440, \\ \frac{1}{2}x_1 + \frac{1}{4}x_2 &\leqslant 65, \\ 2x_1 + \frac{5}{2}x_2 &\leqslant 320. \end{aligned} \tag{10}$$

Кроме того, по смыслу задачи они должны быть неотрицательными:

$$x_1 \geqslant 0, \quad x_2 \geqslant 0. \tag{11}$$

Стоимость запланированной к производству продукции определяется формулой

$$f(x_1, x_2) = 8x_1 + 12x_2. \qquad (12)$$

Итак, с математической точки зрения задача составления оптимального по стоимости выпущенной продукции плана фабрики сводится к определению пары целых чисел x_1 и x_2, удовлетворяющих линейным неравенствам (10), (11) и дающих наибольшее значение линейной функции (12). Мы опять получили типичную задачу линейного программирования. По своей постановке она несколько отличается от транспортной задачи, однако, как мы увидим в дальнейшем, это различие не принципиально.

Для анализа сформулированной задачи рассмотрим плоскость и введем на ней декартову систему координат x_1, x_2. Найдем на этой плоскости множество точек, координаты которых удовлетворяют неравенствам (10) и (11). Неравенства (11) означают, что это множество лежит в первой четверти. Выясним теперь смысл ограничений, которые задаются неравенствами (10).

Проведем на нашей плоскости прямую, определяемую уравнением

$$2x_1 + 4x_2 = 440 \qquad (13)$$

(см. рис. 39). Она делит плоскость на две полуплоскости. На одной из них, расположенной ниже прямой (13), функция $F_1(x_1, x_2) = 2x_1 + 4x_2 - 440$ принимает отрицательные значения, на другой, расположенной выше прямой (13),— положительные. Таким образом, первое из неравенств (10) выполняется на множестве точек, которое включает в себя прямую (13) и полуплоскость, расположенную ниже этой прямой. На рис. 39 соответствующая часть плоскости заштрихована.

Совершенно аналогично можно найти множества точек, удовлетворяющих второму и третьему неравенствам из системы (10). Они показаны на рис. 40 и 41.

Возьмем пересечение трех найденных множеств и выделим его часть, расположенную в первой четверти. В результате мы получим множество точек, удовлетворяющих всей совокупности ограничений (10) и (11). Данное множество имеет вид пятиугольника, показанного на рис. 42. Его вершинами являются точки пересечения прямых, на которых неравенства (10) и (11) переходят в точные равенства. Координаты вершин указаны на рис. 42.

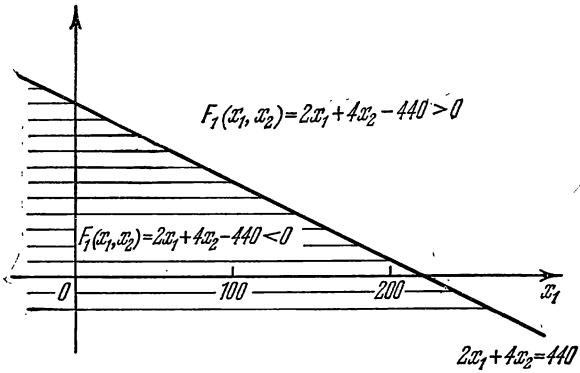

Рис. 39. Решение неравенства $2x_1 + 4x_2 \leqslant 440$. Часть плоскости, точки которой удовлетворяют неравенству, заштрихована.

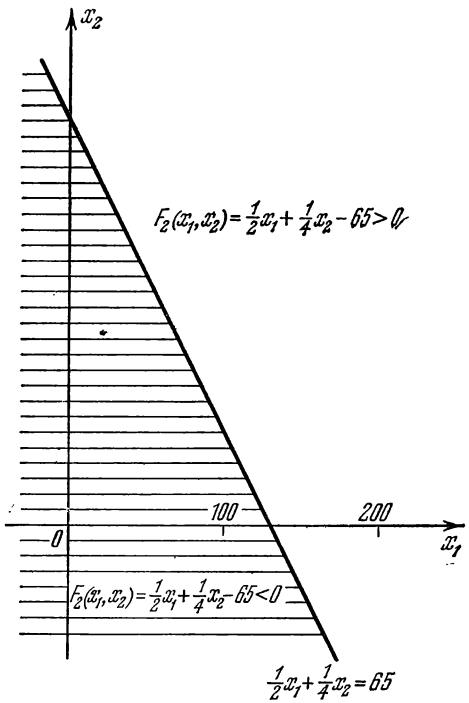

Рис. 40. Решение неравенства $1/2x_1 + 1/4x_2 \leqslant 65$. Часть плоскости, точки которой удовлетворяют неравенству, заштрихована.

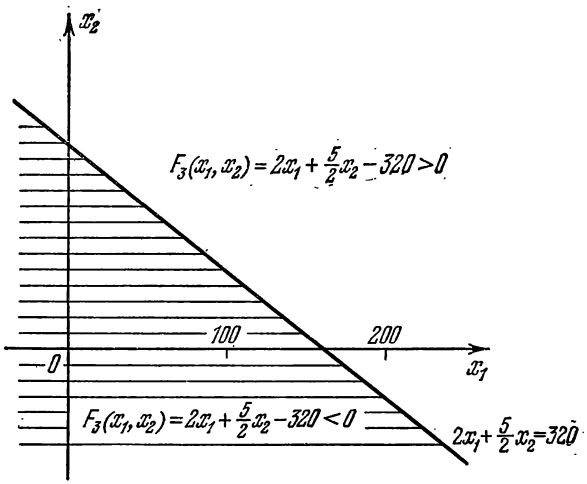

Рис. 41. Решение неравенства $2x_1 + 5/2x_2 \leqslant 320$. Часть плоскости, точки которой удовлетворяют неравенству, заштрихована.

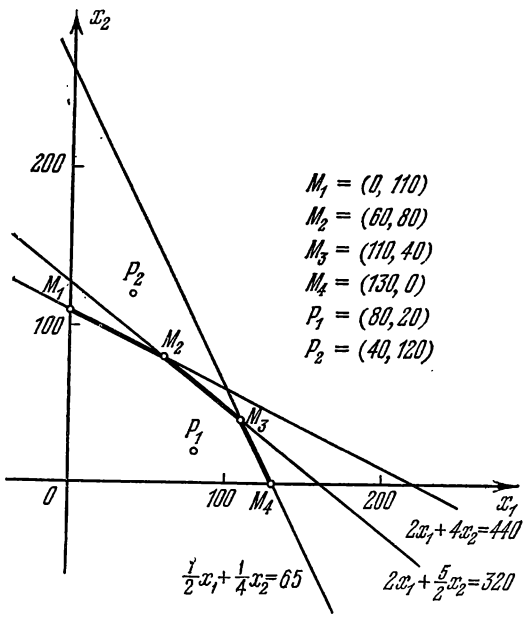

Рис. 42. Пятиугольник $OM_1M_2M_3M_4$, точки которого удовлетворяют системе неравенств (10), (11).

Любая точка P с целочисленными координатами (x_1, x_2) принадлежащая данному пятиугольнику, определяет план выпуска стульев, который может быть выполнен при имеющихся запасах сырья и трудовых ресурсов (реализуемый план). Наоборот, если точка P не принадлежит пятиугольнику, то соответствующий план выполнен быть не может (нереализуемый план).

Для того чтобы лучше понять изложенное, сделайте несколько простых упражнений. Рассмотрите точку P_1 с координатами (80, 20) и убедитесь в том, что они принадлежат пятиугольнику. Проверьте затем «в лоб» с помощью неравенств (10), что план выпуска 80 стульев первого типа и 20 стульев второго типа действительно реализуем.

Возьмем теперь точку P_2 с координатами (40, 120). Она не принадлежит пятиугольнику, т. е. план выпуска 40 и 120 стульев нереализуем. С помощью рис. 42 ответьте на вопрос, недостаток каких ресурсов не позволяет выполнить этот план?

Последнее задание: с помощью рис. 42 укажите такой план P_3, чтобы его выполнение согласовывалось с запасом досок и трудовыми ресурсами, но было бы невозможно из-за нехватки обивочного материала.

После того, как вы сделали упражнения, вернемся к нашей задаче. Мы определили на плоскости x_1, x_2 область реализуемых планов. Теперь нам нужно найти оптимальный план, соответствующий наибольшей стоимости продукции. Иными словами, нужно найти точку пятиугольника с целочисленными координатами, в которой функция $f(x_1, x_2)$ (12) достигает своего наибольшего значения.

Рассмотрим на плоскости x_1, x_2 линии уровня целевой функции (12)

$$8x_1 + 12x_2 = C. \tag{14}$$

Это уравнение описывает семейство прямых линий, параллельных прямой

$$8x_1 + 12x_2 = 0. \tag{15}$$

При параллельном переносе этой прямой вправо параметр C возрастает, влево — убывает.

Свойства функции (12) тесно связаны с прямыми (14). Вдоль каждой из них она сохраняет постоянное значение, равное C, а при переходе с одной прямой на другую ее значение меняется.

Будем рассматривать только первую четверть. Предположим, что мы перешли из точки P_1, расположенной на одной прямой, в точку P_2, расположенную на другой прямой (см. рис. 43). Если вторая прямая расположена дальше от начала координат, чем первая, то функция f при

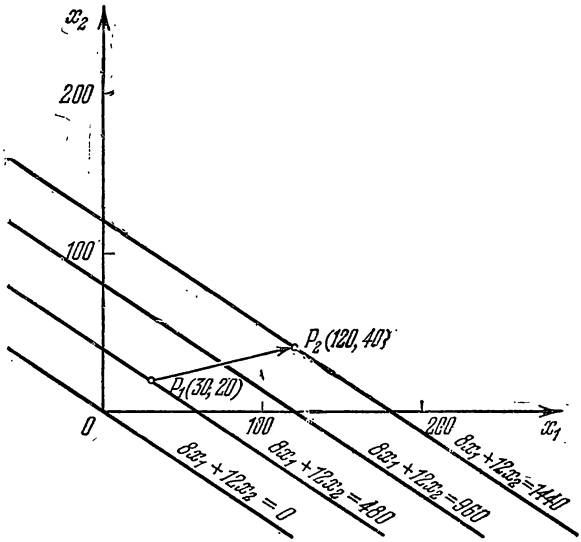

Рис. 43. Линии уровня функции $f(x_1, x_2) = 8x_1 + 12x_2$. Возрастание функции при переходе из точки P_1 в точку P_2.

этом переходе возрастет. Отсюда следует важный вывод: оптимальный план должен располагаться на прямой семейства (14), наиболее удаленной от начала координат.

Этот вывод легко позволяет закончить решение задачи. Посмотрите на рис. 44. На нем воспроизведен пятиугольник реализуемых планов и нарисована прямая семейства (14), проходящая через угловую точку M_2 с координатами (60, 80). Она является предельной прямой семейства, имеющей общую точку с пятиугольником. Если мы попытаемся с помощью параллельного переноса отодвинуть ее дальше от начала координат, то получим прямую, не имеющую общих точек с пятиугольником, т. е. соответствующие планы нереализуемы.

Итак, оптимальный план найден,— он предписывает производство 60 стульев первого типа и 80 стульев второго типа. Стоимость этой продукции 1440 рублей. На выпол-

нение плана нужно затратить: 440 метров досок, 50 квадратных метров обивочной ткани, 320 человеко-часов рабочего времени.

Вы видите, что оптимальный план требует полного использования запаса досок и трудовых ресурсов, в то время как обивочная ткань будет израсходована неполностью — останется 15 квадратных метров.

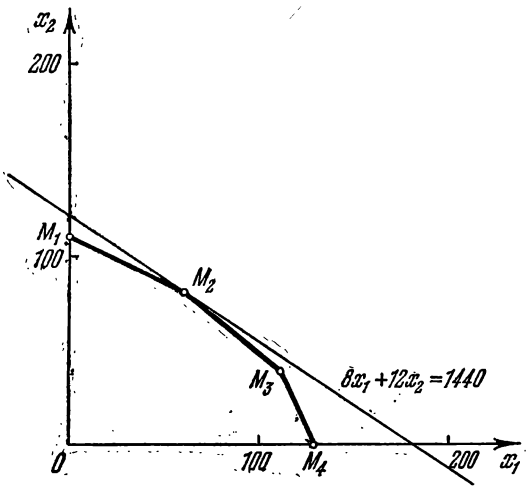

Рис. 44. Определение оптимального плана производства стульев.

Смысл этого результата ясен из рис. 44. Точка M_2, определяющая оптимальный план, является вершиной пятиугольника. Она лежит на пересечении прямых:

$$2x_1 + 4x_2 = 440,$$

$$2x_1 + \frac{5}{2}x_2 = 320.$$

Уравнения этих прямых получаются из первого и третьего условий системы (10) при замене их на строгие равенства. Это означает полный расход досок и трудовых ресурсов. Однако точка M_2 не принадлежит прямой

$$\frac{1}{2}x_1 + \frac{1}{4}x_2 = 65,$$

так что второе условие (10), связанное с ограниченным запасом обивочной ткани, имеет в ней форму неравенства:

$$50 < 65.$$

Проведенный анализ показывает, что дальнейшее увеличение стоимости продукции регламентируется запасом досок и трудовыми ресурсами.

Мы рассмотрели две задачи линейного программирования. По своей постановке они носят достаточно общий характер, несмотря на «учебный» вариант конкретных условий. Небольшое число переменных (4 маршрута, 2 вида продукции) позволило просто и наглядно получить их решение.

Представьте себе теперь, что вам нужно решить не учебную, а настоящую транспортную задачу с десятками пунктов отправления и приема грузов, с сотнями, тысячами маршрутов. Постановка задачи при этом не изменится, однако поиск наилучшего плана существенно усложнится. Точно так же обстоит дело с задачей об оптимальном использовании ограниченных ресурсов. В реальной постановке такие задачи могут быть решены только численно с применением ЭВМ.

§ 2. Математическая постановка задачи линейного программирования

Стандартная или, как принято говорить, каноническая постановка задачи линейного программирования заключается в следующем: найти значения переменных $x_1, x_2, \ldots, x_n$, которые удовлетворяют системе уравнений

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m, \end{aligned} \tag{16}$$

являются неотрицательными:

$$x_1 \geqslant 0, \quad x_2 \geqslant 0, \quad \ldots, \quad x_n \geqslant 0 \tag{17}$$

и дают наименьшее значение функции

$$f(x_1, x_2, \ldots, x_n) = c_1x_1 + c_2x_2 + \ldots + c_nx_n. \tag{18}$$

Примером задачи в такой канонической постановке является транспортная задача, рассмотренная в предыдущем разделе. После исключения зависимого уравнения ее условия были записаны в виде системы (4) трех уравнений ($m = 3$) с четырьмя неизвестными ($n = 4$) и неравенств (2).

Задача об оптимальном распределении ресурсов в своей первоначальной постановке не является канонической: условия (10) имеют вид неравенств, а не уравнений. Однако такую задачу можно свести к канонической. Делается это следующим образом.

Введем в дополнение к имеющимся переменным x_1, x_2 три вспомогательные переменные: x_3, x_4, x_5 и вместо неравенств (10) рассмотрим уравнения:

$$\begin{aligned} 2x_1 + 4x_2 + x_3 &= 440, \\ \frac{1}{2}x_1 + \frac{1}{4}x_2 + x_4 &= 65, \\ 2x_1 + \frac{5}{2}x_2 + x_5 &= 320. \end{aligned} \tag{19}$$

Перепишем первое из них в виде

$$x_3 = 440 - 2x_1 - 4x_2.$$

Отсюда и из (10) следует, что $x_3 \geqslant 0$. Аналогично устанавливается неотрицательность x_4 и x_5. Вместе с (11) это дает:

$$x_1 \geqslant 0, \quad x_2 \geqslant 0, \quad x_3 \geqslant 0, \quad x_4 \geqslant 0, \quad x_5 \geqslant 0. \tag{20}$$

В результате ценой увеличения числа переменных мы пришли к канонической задаче со стандартными условиями в виде уравнений (19) ($m = 3$, $n = 5$) и неравенств (20).

Отметим, что дополнительные переменные x_3, x_4, x_5 вошли в условия задачи, но не в выражение (12) для целевой функции. Она не меняет своего вида и остается функцией двух переменных x_1, x_2 ($c_3 = 0$, $c_4 = 0$, $c_5 = 0$).

И последнее замечание. То, что в задаче об оптимальном использовании ресурсов ищется наибольшее, а не наименьшее значение целевой функции,— не принципиально. Мы всегда можем рассмотреть вспомогательную функцию $F = -f$, тогда задача о наибольшем значении функции f окажется эквивалентной задаче о наименьшем значении функции F.

Итак, мы убедились в том, что «неканоническую» задачу линейного программирования можно свести к канонической.

Перейдем теперь к обсуждению задачи линейного программирования в ее канонической постановке. Всякое решение системы уравнений (16), удовлетворяющее неравенствам (17), будем называть допустимым; допустимое

решение, дающее наименьшее значение целевой функции (18), назовем оптимальным.

При исследовании данной задачи могут встретиться следующие пять случаев:

1) система (16) не имеет решений, т. е. ее уравнения несовместны;

2) система (16) не имеет допустимых решений, т. е. ни одно решение не удовлетворяет неравенствам (17);

3) система (16) имеет бесчисленное множество допустимых решений, однако целевая функция (18) на этом множестве не ограничена снизу, т. е. среди допустимых решений нет оптимального;

4) система (16) имеет единственное допустимое решение;

5) система (16) имеет бесчисленное множество допустимых решений, и целевая функция (18) ограничена на этом множестве снизу.

Легко привести примеры, реализующие каждый из этих случаев. Предположим, например, что (16) состоит из одного уравнения с двумя неизвестными

$$x_1 - x_2 = 1, \tag{21}$$

$$x_1 \geqslant 0, \quad x_2 \geqslant 0, \tag{22}$$

а целевая функция (18) имеет вид

$$f(x_1, x_2) = -x_1 - x_2. \tag{23}$$

Выразим из уравнения (21) x_1 через x_2:

$$x_1 = x_2 + 1. \tag{24}$$

Если переменная x_2 неотрицательна ($x_2 \geqslant 0$), то, согласно (24), переменная x_1 также будет неотрицательной.

Подставим (24) в выражение (23) для целевой функции. В результате получим

$$f = -2x_2 - 1 \quad (x_2 \geqslant 0). \tag{25}$$

Мы видим, что при допустимых значениях x_2 функция (25) может принимать значения, которые будут меньше любого наперед заданного числа. Иными словами, она не ограничена снизу. Этот пример соответствует случаю 3).

Транспортная задача, разобранная в предыдущем разделе, соответствует случаю 5). Примеры задач, реализующих случаи 1), 2) и 4), попробуйте составить сами. Это сделать нетрудно.

Очевидно, что в случаях 1), 2), 3) задача линейного программирования (16) — (18) неразрешима. В случае 4) единственное допустимое решение является одновременно и оптимальным. У него просто нет «конкурентов». И только в случае 5) мы сталкиваемся с проблемой выбора: как среди бесчисленного множества допустимых решений найти оптимальное. Именно этот случай представляет наибольший интерес, поскольку к нему, как правило, сводятся реальные задачи оптимизации. Остановимся подробнее на его исследовании.

Будем считать, что система (16) совместна и среди ее уравнений нет линейно зависимых, т. е. ни одно из них не является следствием остальных уравнений. Для того чтобы такая система имела бесчисленное множество решений, необходимо и достаточно, чтобы число уравнений в ней было меньше числа неизвестных.

Положим $k = n - m$ $(k > 0)$. При сделанных предположениях в системе (16) найдется m переменных, которые можно выразить через остальные k. Их называют, соответственно, базисными и свободными переменными. Придавая свободным переменным произвольные значения и вычисляя по ним базисные, мы будем получать различные решения системы (16).

Деление переменных на базисные и свободные не является однозначным. Например, для системы (4) любую тройку переменных можно принять за базисную, оставшаяся четвертая переменная будет при этом свободной. Таким образом, в зависимости от нашего выбора каждая переменная может играть роль как базисной, так и свободной переменной. То же самое можно сказать о системе (19). Любую тройку из пяти переменных, входящих в эту систему, можно принять за базисную, а две оставшиеся переменные считать свободными.

Пусть в системе (16) проведено одно из возможных разделений переменных на базисные и свободные. Положим все свободные переменные равными нулю и определим из системы соответствующие значения базисных переменных. Если ни одно из них не будет отрицательным, то мы получим допустимое решение, у которого k переменных равны нулю. Такое решение называется опорным. В теории линейного программирования доказывается, что оптимальное решение (если оно существует) является опорным.

Этот результат приводит к следующему возможному

алгоритму решения задачи (при предположении, что оно существует). Рассматривая различные разделения переменных x_i на базисные и свободные, нужно найти все опорные решения системы (16), вычислить в них значения целевой функции и найти среди них наименьшее. Тогда можно утверждать, что опорное решение, которому оно соответствует, является оптимальным.

В качестве примера рассмотрим еще раз транспортную задачу (4), (2), (3) ($n = 4$, $m = 3$, $k = 1$). Принимая по очереди x_1, x_2, x_3, x_4 за свободную переменную и полагая ее равной нулю, получим четыре решения:

$$\begin{array}{r} (0,\ 50,\ \ 40,\ 30), \\ (50,\ \ 0,\ -10,\ 80), \\ (40,\ 10,\ \ \ 0,\ 70), \\ (-30,\ 80,\ \ 70,\ \ 0). \end{array} \tag{26}$$

Два из них — первое и третье — допустимы, т. е. они являются опорными решениями, два других — нет. Вычисляя для опорных решений значения целевой функции (3), получим:

$$f(0,\ 50,\ 40,\ 30) = 142,$$
$$f(40,\ 10,\ 0,\ 70) = 134.$$

Из сравнения этих чисел делаем вывод, что оптимальным решением, обеспечивающим наименьшую стоимость перевозок, является решение (40, 10, 0, 70). Этот результат, конечно, совпадает с результатом § 1.

Однако поиск всех опорных решений прямым перебором годится только для такой учебной задачи с минимальным числом переменных. В практических задачах с большим числом переменных число вариантов настолько велико, что прямой перебор окажется не под силу даже сверхмощной ЭВМ. В следующем параграфе мы опишем метод, который позволяет вести поиск не вслепую, а целенаправленно, приближаясь с каждым шагом к оптимальному решению.

§ 3. Симплекс-метод

Поиск оптимального решения задачи (16), (17), (18) начинается с построения какого-нибудь опорного решения или с доказательства отсутствия таких решений у системы (16). (Последнее означает неразрешимость задачи.) Сущест-

вуют специальные методы, позволяющие сделать этот первый шаг. Однако мы, чтобы не усложнять изложение, не будем их описывать, а предположим, что первый шаг уже сделан, т. е. найдено нужное разделение переменных на базисные и свободные, которое при нулевых значениях свободных переменных дает опорное решение.

Пусть, для определенности, первые m переменных x_i $(1 \leqslant i \leqslant m)$ являются базисными, а остальные $(m + 1 \leqslant i \leqslant n)$ свободными. Решая систему (16) относительно базисных переменных, получим формулы, связывающие их со свободными:

$$\begin{aligned} x_1 &= p_1 + q_{1,m+1} x_{m+1} + q_{1,m+2} x_{m+2} + \ldots + q_{1,n} x_n, \\ x_2 &= p_2 + q_{2,m+1} x_{m+1} + q_{2,m+2} x_{m+2} + \ldots + q_{2,n} x_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_m &= p_m + q_{m,m+1} x_{m+1} + q_{m,m+2} x_{m+2} + \ldots + q_{m,n} x_n. \end{aligned} \tag{27}$$

При нулевых значениях свободных переменных они дают опорное решение:

$$x_1 = p_1,\ x_2 = p_2,\ \ldots,\ x_m = p_m,\ x_{m+1} = 0,\ \ldots,\ x_n = 0, \qquad p_i \geqslant 0,\quad 1 \leqslant i \leqslant m. \tag{28}$$

Чтобы исключить из рассмотрения один тонкий частный случай, предположим, что все числа p_i строго положительны.

Для дальнейшего анализа запишем целевую функцию (18) с помощью формул (17) как функцию только свободных переменных:

$$f = r_0 + r_{m+1} x_{m+1} + r_{m+2} x_{m+2} + \ldots + r_n x_n. \tag{29}$$

Здесь

$$r_0 = p_1 c_1 + p_2 c_2 + \ldots + p_m c_m, \tag{30}$$

$$r_i = c_i + c_1 q_{1,i} + c_2 q_{2,i} + \ldots + c_m q_{m,i}, \quad m + 1 \leqslant i \leqslant n. \tag{31}$$

В частности, на опорном решении (28) целевая функция принимает значение r_0.

Выясним теперь, является ли опорное решение (28) оптимальным или нет. Для этого рассмотрим, как будет меняться целевая функция при увеличении каждой из свободных переменных. (Уменьшить свободные переменные мы не можем, так как при этом они станут отрицательными.)

Если все коэффициенты r_i $(m+1 \leqslant i \leqslant n)$ неотрицательны, то увеличение любой свободной переменной не может привести к уменьшению целевой функции. Следовательно, в этом случае нам повезло: мы сразу получили оптимальное решение.

Предположим теперь, что среди коэффициентов r_i $(m+1 \leqslant i \leqslant n)$ имеется хотя бы один отрицательный. Пусть, например, $r_n < 0$. Теперь, увеличивая x_n при нулевых значениях остальных свободных переменных x_i $(m+1 \leqslant i \leqslant n-1)$, мы будем уменьшать целевую функцию (29) по сравнению со значением r_0, которое достигается на решении (28). Исследуем, насколько можно увеличивать x_n.

Если все коэффициенты $q_{1,n}, q_{2,n}, \ldots, q_{m,n}$ при переменной x_n в формулах (29) неотрицательны, то переменную x_n можно увеличивать неограниченно. При этом переменные $x_1, x_2, \ldots, x_n$ будут оставаться неотрицательными, и мы не выйдем из области допустимых решений. В то же время целевая функция (29) будет при этом неограниченно уменьшаться, т. е. она не ограничена снизу и оптимального решения у задачи не существует. Это третий случай, который практически обычно не встречается.

Предположим теперь, что среди коэффициентов $q_{1,n}, q_{2,n}, \ldots, q_{m,n}$ имеется хотя бы один отрицательный. Тогда мы уже не можем неограниченно увеличивать x_n из-за угрозы сделать отрицательными x_i в (27).

Рассмотрим числа $-p_i/q_{i,n}$ $(1 \leqslant i \leqslant m)$, выберем из них положительные, а среди положительных найдем минимальное. Пусть, например, оно соответствует номеру $i = m$. В этом случае мы можем увеличивать переменную x_n вплоть до значения $x_n = -p_m/q_{m,n}$, при котором x_m в формулах (27) обратится в нуль, а остальные x_i $(1 \leqslant i \leqslant m-1)$ будут оставаться положительными. В результате от (28) перейдем к новому опорному решению

$$
\begin{aligned}
x_1 &= p_1 - p_m \frac{q_{1,n}}{q_{m,n}}, \\
x_2 &= p_2 - p_m \frac{q_{2,n}}{q_{m,n}}, \\
&\cdots\cdots\cdots\cdots \\
x_{m-1} &= p_{m-1} - p_m \frac{q_{m-1,n}}{q_{m,n}}, \\
x_n &= -\frac{p_m}{q_{m,n}}, \\
x_m &= 0, \quad x_{m+1} = 0, \ldots, x_{n-1} = 0.
\end{aligned}
\tag{32}
$$

Целевая функция на этом решении принимает значение, меньшее чем r_0:

$$f = r_0 - r_n \frac{p_m}{q_{m,n}} < r_0.$$

Теперь нужно повторить всю эту процедуру заново, выбирая в качестве базисных переменных $x_1, x_2, \ldots, x_{m-1}$ и x_n. В результате опять мы либо убедимся в том, что опорное решение (32) является оптимальным, либо найдем другое опорное решение с меньшим значением целевой функции. Поскольку общее число опорных решений системы (16) конечно, то после некоторого конечного числа шагов мы придем к оптимальному решению.

Описанный метод поиска оптимального решения носит в теории линейного программирования название симплекс-метода. Подчеркнем, что, в отличие от «слепого» перебора, симплекс-метод позволяет вести поиск целенаправленно. Каждый шаг уменьшает значение целевой функции, приближая нас к оптимальному решению.

§ 4. Снова задача о стульях

Рассмотрим еще раз задачу об оптимальном распределении ресурсов мебельной фабрики, используя на этот раз симплекс-метод. Запись условий задачи в канонической форме сводится к системе уравнений (19) ($m = 3$, $n = 5$, $k = n - m = 2$) и неравенств (20). Целевую функцию (12) возьмем со знаком минус:

$$f = -8x_1 - 12x_2, \tag{33}$$

чтобы искать ее наименьшее, а не наибольшее значение.

Поиск оптимального решения задачи начинается с построения какого-нибудь опорного решения системы (19). Наиболее просто найти опорное решение, соответствующее нулевым значениям переменных x_1 и x_2:

$$x_1^{(0)} = 0, \quad x_2^{(0)} = 0, \quad x_3^{(0)} = 440, \quad x_4^{(0)} = 65, \quad x_5^{(0)} = 320. \tag{34}$$

Ему соответствует нулевое значение целевой функции:

$$f_0 = 0. \tag{35}$$

Для исследования решения (34) мы должны ненулевые переменные x_3, x_4, x_5 принять за базисные, переменные x_1, x_2 — за свободные и выразить из системы (19) базисные переменные через свободные. Соответствующие

формулы имеют вид

$$x_3 = 440 - 2x_1 - 4x_2,$$
$$x_4 = 65 - \frac{1}{2}x_1 - \frac{1}{4}x_2, \qquad (36)$$
$$x_5 = 320 - 2x_1 - \frac{5}{2}x_2.$$

Решение (34) не является оптимальным. Формула (33) показывает, что увеличение любой из двух свободных переменных уменьшает целевую функцию.

Будем, например, увеличивать переменную x_1, оставляя переменную x_2 равной нулю. Максимальное допустимое значение x_1 определится вторым соотношением (36): $x_1 = 130$. При этом x_4 обратится в нуль, а переменные x_3 и x_5 останутся положительными. Дальнейшее увеличение x_1 сделало бы x_4 отрицательным, что противоречит условию допустимости.

Итак, полагая в формулах (36) $x_1 = 130$, $x_2 = 0$ и вычисляя x_3, x_4, x_5, мы перейдем от решения (34) к новому опорному допустимому решению:

$$x_1^{(1)} = 130, \quad x_2^{(1)} = 0, \quad x_3^{(1)} = 180, \quad x_4^{(1)} = 0, \quad x_5^{(1)} = 60. \quad (37)$$

Вычислим значение целевой функции, соответствующее этому решению:

$$f_1 = -1040 < f_0 = 0. \qquad (38)$$

Мы видим, что решение (37) «лучше» решения (34).

Для исследования решения (37) проделаем ту же процедуру. Примем ненулевые переменные x_1, x_3, x_5 за базисные, переменные x_2, x_4 — за свободные и выразим из системы (19) первые через вторые:

$$x_1 = 130 - \frac{1}{2}x_2 - 2x_4,$$
$$x_3 = 180 - 3x_2 + 4x_4, \qquad (39)$$
$$x_5 = 60 - \frac{3}{2}x_2 + 4x_4.$$

Подставим выражение для x_1 в формулу (33) и запишем целевую функцию как функцию свободных переменных:

$$f = -1040 - 8x_2 + 16x_4. \qquad (40)$$

Переменная x_2 входит в (40) со знаком минус, а переменная x_4 — со знаком плюс. Таким образом, для уменьшения целевой функции нужно увеличивать переменную x_2, оставляя переменную x_4 равной нулю. Максимальное

допустимое значение переменной x_2 определится третьим соотношением (38): $x_2 = 40$. Полагая в формулах (39) $x_2 = 40$, $x_4 = 0$, мы придем к новому опорному решению:

$$x_1^{(2)} = 110, \quad x_2^{(2)} = 40, \quad x_3^{(2)} = 60, \quad x_4^{(2)} = 0, \quad x_5^{(2)} = 0. \tag{41}$$

Этот переход позволяет нам снова уменьшить значение целевой функции:

$$f_2 = -1360 < f_1 = -1040. \tag{42}$$

Примем опять ненулевые переменные x_1, x_2, x_3 за базисные, переменные x_4, x_5 — за свободные и выразим из системы (19) первые через вторые:

$$\begin{aligned} x_1 &= 110 - \frac{10}{3} x_4 + \frac{1}{3} x_5, \\ x_2 &= 40 + \frac{8}{3} x_4 - \frac{2}{3} x_5, \\ x_3 &= 60 - 4x_4 + 2x_5. \end{aligned} \tag{43}$$

Подставляя (43) в формулу (33), запишем целевую функцию как функцию свободных переменных x_4, x_5:

$$f = -1360 - \frac{16}{3} x_4 + \frac{16}{3} x_5. \tag{44}$$

Переменная x_4 входит в формулу (44) со знаком минус, а переменная x_5 — со знаком плюс. Следовательно, теперь нужно увеличивать переменную x_4, оставляя переменную x_5 равной нулю. Максимальное допустимое значение x_4 определится третьим соотношением (43): $x_4 = 15$. Полагая в формулах (43) $x_4 = 15$, $x_5 = 0$, придем к очередному опорному решению:

$$x_1^{(3)} = 60, \quad x_2^{(3)} = 80, \quad x_3^{(3)} = 0, \quad x_4^{(3)} = 15, \quad x_5^{(3)} = 0, \tag{45}$$

которое снова уменьшает значение целевой функции:

$$f_3 = -1440 < f_2 = -1360. \tag{46}$$

Решение (45) показывает, что для выполнения очередного шага за базисные переменные нужно принять x_1, x_2, x_4, а переменные x_3, x_5 считать свободными. Выражение базисных переменных и целевой функции через свободные переменные дается формулами:

$$\begin{aligned} x_1 &= 60 + \frac{5}{6} x_3 - \frac{4}{3} x_5, \\ x_2 &= 80 - \frac{2}{3} x_3 + \frac{2}{3} x_5, \\ x_4 &= 15 - \frac{1}{4} x_3 + \frac{1}{2} x_5, \end{aligned} \tag{47}$$

$$f = -1440 + \frac{4}{3} x_3 + \frac{8}{3} x_5. \tag{48}$$

Мы видим, что x_3 и x_5 входят в выражение для целевой функции с положительными коэффициентами, так что их увеличение может только увеличить целевую функцию. Это означает, что решение (45) является оптимальным, а число —1440 — наименьшим значением целевой функции на классе допустимых решений.

Внимательный читатель, наверное, заметил, что мы начали поиск оптимального плана работы фабрики с очень плохого решения. Действительно, исходное решение (34) предлагает не делать ни одного стула, оставив неиспользованными запасы материалов и трудовые ресурсы. Стоимость продукции при этом, естественно, равна нулю (35).

Затем в результате ряда коррекций мы пришли к решению (45), которому соответствует оптимальный план: нужно произвести 60 стульев первого типа и 80 стульев второго типа на общую сумму 1440 рублей. Этот план уже известен нам из первого параграфа, где он был найден другим методом.

Если нанести решения (34), (37), (41) и (45) на рис. 44, то они попадут в вершины O, M_4, M_3 и M_2 пятиугольника допустимых планов. Таким образом, симплекс-метод вел нас из начала координат O к оптимальному плану M_2 по периметру этого пятиугольника.

Мы попали на первом шаге из точки O в точку M_4 потому, что начали увеличивать x_1, оставляя x_2 равным нулю. Если бы мы поступили наоборот, то попали бы из точки O в точку M_1. В этом случае мы достигли бы оптимального решения M_2 уже на втором шаге. Попробуйте сами проделать необходимые вычисления, это будет хорошим упражнением.

Правила, по которым «работает» симплекс-метод, легко запрограммировать и поручить проведение всех вычислений машине. Тогда появится возможность перейти от учебных задач с четырьмя — пятью переменными к реальным задачам оптимизации с тысячами переменных. Применение ЭВМ к решению задач линейного программирования, начавшееся в пятидесятые годы, послужило основой широкого использования математических методов в экономике. В настоящее время симплекс-метод реализован в виде стандартных программ, которые позволяют получить ответ наиболее эффективным путем с минимальными затратами машинного времени.

Глава 7

ОПРЕДЕЛЕННЫЙ ИНТЕГРАЛ. ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

§ 1. Как подсчитать путь при неравномерном движении или работу переменной силы

Пусть материальная точка движется со скоростью $v(t)$. Требуется определить путь, который она пройдет за время $T_1 \leqslant t \leqslant T_2$. Если точка движется равномерно, т. е. $v(t) = v = \text{const}$, то ответ дается простой формулой:

$$S = v(T_2 - T_1).$$

Подсчитаем пройденный путь в случае неравномерного движения. Для этого фиксируем между начальным и конечным моментами времени ряд промежуточных моментов: $T_1 < t_1 < t_2 < \ldots < t_{n-1} < T_2$. Положим также, для единообразия обозначений, $t_0 = T_1$ и $t_n = T_2$. В результате исходный временной промежуток $[T_1, T_2]$ окажется разбитым на более мелкие промежутки $[t_{k-1}, t_k]$, $k = 1, 2, \ldots, n$.

Рассмотрим один из этих промежутков. Выберем на нем какой-нибудь момент времени τ_k, $t_{k-1} \leqslant \tau_k \leqslant t_k$, и найдем соответствующее ему значение скорости: $v_k = v(\tau_k)$. Если данный временной промежуток является достаточно коротким, то скорость $v(t)$ изменяется в течение него мало. Пренебрегая этими изменениями, будем приближенно считать движение равномерным со скоростью $v_k = v(\tau_k)$, $t_{k-1} \leqslant t \leqslant t_k$. В результате для пройденного за данный промежуток времени пути Δs_k можно написать

$$\Delta s_k \approx v(\tau_k)\,\Delta t_k, \quad \Delta t_k = t_k - t_{k-1}.$$

Чтобы подсчитать весь путь, нужно проделать данную процедуру для каждого временного промежутка и результаты сложить:

$$s \approx v(\tau_1)\Delta t_1 + v(\tau_2)\Delta t_2 + \ldots + v(\tau_n)\Delta t_n. \quad (1)$$

Полученное приближенное выражение зависит от того, как мы разбили временной промежуток $[T_1, T_2]$ на части и как для каждой из них выбрали момент времени τ_k, с помощью которого определяется скорость равномерного движения $v_k = v(\tau_k)$, заменяющего заданное неравномерное движение при $t_{k-1} \leqslant t \leqslant t_k$. Разным разбиениям и разному выбору моментов τ_k соответствуют разные значения суммы (1).

Будем делать разбиение временного промежутка $[T_1, T_2]$ все более мелким. При этом отклонение движения от равномерного для каждого частичного временного промежутка $[t_{k-1}, t_k]$ должно уменьшаться, и естественно ожидать, что в пределе при неограниченном измельчении разбиения приближенное равенство (1) перейдет в точную формулу для пройденного пути. Предел выражения, стоящего в правой части равенства (если он существует), называется определенным интегралом от функции $v(t)$ по отрезку $[T_1, T_2]$ и обозначается символом

$$s = \int_{T_1}^{T_2} v(t)\,dt. \tag{2}$$

Рассмотрим еще одну задачу. Пусть материальная точка движется вдоль оси x под действием некоторой силы F, величина которой зависит от положения точки: $F = F(x)$. Требуется подсчитать работу, совершенную при перемещении точки из положения $x = a$ в положение $x = b$. Эта задача с математической точки зрения совершенно аналогична предыдущей, и для ее решения можно использовать тот же подход.

Выберем на отрезке $[a, b]$ ряд промежуточных точек $a < x_1 < x_2 < \ldots < x_{n-1} < b$ и положим $x_0 = a$, $x_n = b$. В результате исходный отрезок $[a, b]$ окажется разбитым на n отрезков $[x_{k-1}, x_k]$, $k = 1, 2, \ldots, n$. Возьмем на каждом из них по произвольной точке ξ_k: $x_{k-1} \leqslant \xi_k \leqslant x_k$ и подсчитаем величину силы в этой точке: $F_k = F(\xi_k)$. Если отрезок $[x_{k-1}, x_k]$ достаточно мал, то сила изменяется на нем мало и мы приближенно можем считать ее постоянной: $F(x) \approx F(\xi_k)$, $x \in [x_{k-1}, x_k]$. При этом для работы по перемещению материальной точки из положения x_{k-1} в положение x_k получим

$$\Delta A_k \approx F(\xi_k)\,\Delta x_k, \quad \Delta x_k = x_k - x_{k-1}.$$

Складывая эти величины, подсчитаем полную работу:

$$A \approx F(\xi_1)\Delta x_1 + F(\xi_2)\Delta x_2 + \ldots + F(\xi_n)\Delta x_n. \quad (3)$$

Эта формула аналогична формуле (1). Чем мельче разбиение, тем она точнее. В пределе при неограниченном из-

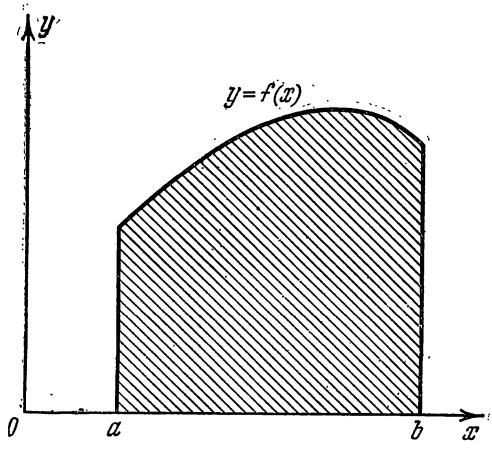

Рис. 45. Криволинейная трапеция, ограниченная графиком функции $f(x)$, осью x и двумя вертикальными прямыми $x = a$ и $x = b$.

мельчении разбиения она дает точное выражение для работы в виде интеграла от функции $F(x)$:

$$A = \int_a^b F(x)\,dx. \quad (4)$$

Совершенно аналогично можно подсчитать заряд Q, который переносится через поперечное сечение проводника током $I(t)$ за время $T_1 \leqslant t \leqslant T_2$:

$$Q = \int_{T_1}^{T_2} I(t)\,dt. \quad (5)$$

Рассмотрим последний пример. На рис. 45 показана фигура, ограниченная графиком некоторой функции $y = f(x)$, осью x и двумя вертикальными прямыми $x = a$ и $x = b$. Такую фигуру принято называть криволинейной трапецией. Требуется подсчитать ее площадь.

Для решения этой задачи, как и трех предыдущих, разобьем отрезок $[a, b]$ точками x_k $(k = 0, 1, 2, \ldots, n,$

$x_0 = a$, $x_n = b$), на частичные отрезки $[x_{k-1}, x_k]$. На каждом из них выберем произвольную точку $\xi_k \in [x_{k-1}x_k]$, $k = 1, 2, \ldots, n$, вычислим в ней значение функции $f(x)$ и построим прямоугольник высотой $f(\xi_k)$ (см. рис. 46).

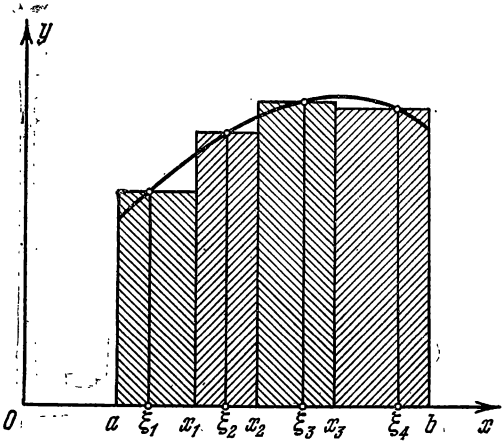

Рис. 46. Построение ступенчатого многоугольника, аппроксимирующего криволинейную трапецию.

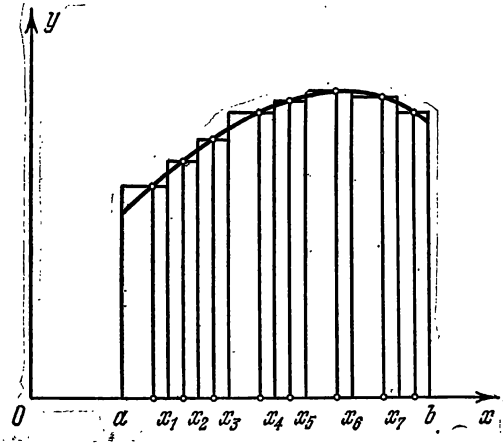

Рис. 47. Улучшение аппроксимации криволинейной трапеции ступенчатым многоугольником при измельчении разбиения.

В результате мы получим ступенчатый многоугольник, который можно рассматривать как некоторое приближение к более сложной фигуре — заданной криволинейной трапеции.

Легко подсчитать площадь этого многоугольника S, являющуюся суммой площадей отдельных прямоугольников:

$$S = f(\xi_1)\Delta x_1 + f(\xi_2)\Delta x_2 + \ldots + f(\xi_n)\Delta x_n. \qquad (6)$$

При измельчении разбиения точность аппроксимации криволинейной трапеции ступенчатым многоугольником повышается (сравните рис. 46 и 47). Поэтому за площадь криволинейной трапеции естественно принять предел площадей ступенчатых многоугольников при неограниченном измельчении разбиения, т. е. соответствующий определенный интеграл:

$$S = \int_a^b f(x)\,dx. \qquad (7)$$

Можно было бы предложить множество других задач, решение которых также записывается через интеграл, но в этом нет необходимости: характер таких задач достаточно ясен из разобранных примеров. Нам остается подчеркнуть, что понятие интеграла является одним из наиболее важных математических понятий.

§ 2. Формула Ньютона — Лейбница

Для того чтобы воспользоваться формулами типа (2), (4), (5), (7) для подсчета соответствующих величин, нужно уметь вычислять определенные интегралы. Наиболее простой способ их вычисления основан на формуле Ньютона — Лейбница

$$\int_a^b f(x)\,dx = F(b) - F(a), \qquad (8)$$

где $F(x)$ — какая-нибудь первообразная подынтегральной функции $f(x)$.

Формула (8) обсуждается в учебнике по алгебре и началам анализа для 10 класса, там же даются примеры ее применения. Формула играет важнейшую роль в математическом анализе. Устанавливая связь задачи определенного интегрирования с другой математической задачей, с отысканием первообразной, она позволяет сравнительно легко вычислять широкий круг интегралов. Однако формула (8) не дает общего правила нахождения интеграла от произвольной функции $f(x)$ по ее значениям на

отрезке $[a, b]$, т. е. она не является алгоритмом решения рассматриваемой задачи. Дело в том, что отыскание первообразной — это достаточно сложная математическая задача, для решения которой в явном виде не существует универсальных методов.

Рассмотрим в качестве примера два интеграла: $\int_1^2 (1/x)\,dx$ и $\int_1^2 (e^{-x}/x)\,dx$. В первом случае первообразная подынтегральной функции $f(x) = 1/x$ есть $F(x) = \ln x$, в результате получаем:

$$\int_1^2 \frac{1}{x}\,dx = \ln 2 - \ln 1 = \ln 2. \qquad (9)$$

Однако для вычисления второго интеграла такой подход не применим, потому что первообразная функции $f(x) = e^{-x}/x$ не является элементарной функцией.

Отметим также, что формула Ньютона — Лейбница совершенно не работает как метод вычисления определенных интегралов в случае, когда подынтегральная функция задана графиком или таблицей.

Мы не будем больше останавливаться на обсуждении формулы Ньютона — Лейбница. У нас совсем другая цель — познакомить читателей с универсальными алгоритмами решения этой задачи, которые позволяют подсчитывать интегралы прямо по значениям подынтегральной функции $f(x)$. Соответствующие формулы обычно называют формулами численного интегрирования или квадратурными формулами (буквально — формулами вычисления площади).

Однако, прежде чем начинать обсуждение формул численного интегрирования, нам необходимо более глубоко проанализировать круг вопросов, связанных с понятием определенного интеграла.

§ 3. Понятие определенного интеграла

В первом параграфе мы на интуитивном уровне рассмотрели несколько примеров, приводящих к идее определенного интеграла. Обсудим теперь это понятие на уровне строгих математических определений и теорем.

Пусть на отрезке $[a, b]$ задана некоторая функция $f(x)$. Разобьем этот отрезок точками x_k: $a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$ на частичные отрезки $[x_{k-1}, x_k]$, $k = 1, 2, \ldots, n$. Такую операцию условимся называть разбиением и обозначать символом $T(x_k)$. Вычислим длины полученных отрезков $\Delta x_k = x_k - x_{k-1}$ и положим $\Delta = \max\{\Delta x_k\}$. Величина Δ является длиной наибольшего отрезка при разбиении $T(x_k)$.

Выберем на каждом отрезке $[x_{k-1}, x_k]$ какую-нибудь точку ξ_k: $x_{k-1} \leqslant \xi_k \leqslant x_k$ и вычислим в них значение функции $f(x)$. Из найденных величин образуем сумму, которую называют интегральной суммой:

$$I(x_k, \xi_k) = f(\xi_1)\Delta x_1 + f(\xi_2)\Delta x_2 + \ldots + f(\xi_n)\Delta x_n. \quad (10)$$

Значение интегральной суммы определяется выбором точек разбиения x_k и точек ξ_k. Выражения (1), (3), (6), которые появлялись при обсуждении примеров первого параграфа, представляют собой интегральные суммы для соответствующих функций.

Нас будет интересовать поведение интегральных сумм при неограниченном измельчении разбиения, т. е. при $\Delta \to 0$. Сформулируем два определения.

Определение. Число $\mathcal{J}$ называется *пределом интегральных сумм* $I(x_k, \xi_k)$ при $\Delta \to 0$, если для всякой точности $\varepsilon > 0$ можно указать такое δ, что для любого разбиения $T(x_k)$, удовлетворяющего условию $\Delta < \delta$, при произвольном выборе точек $\xi_k \in [x_{k-1}, x_k]$ выполняется неравенство:

$$|\mathcal{J} - I(x_k, \xi_k)| < \varepsilon. \quad (11)$$

Определение. Если для функции $f(x)$, заданной на отрезке $[a, b]$, существует предел $\mathcal{J}$ интегральных сумм при $\Delta \to 0$, то она называется *интегрируемой* на отрезке $[a, b]$, а число $\mathcal{J}$ называется *определенным интегралом* от этой функции по отрезку $[a, b]$ и обозначается символом

$$\mathcal{J} = \int_a^b f(x)\,dx.$$

Рассмотрим в качестве примера функцию, равную постоянной на отрезке $[a, b]$: $f(x) = C$. Проведем какое-нибудь разбиение отрезка $T(x_k)$, выберем точки ξ_k и соста-

вим интегральную сумму

$$I(x_k, \xi_k) = C(\Delta x_1 + \Delta x_2 + \ldots + \Delta x_n) = C(b - a).$$

Мы видим, что она не зависит ни от разбиения, ни от выбора точек ξ_k. Следовательно, существует предел интегральных сумм при $\Delta \to 0$, равный той же величине:

$$\lim_{\Delta \to 0} I(x_k, \xi_k) = \int_a^b C\,dx = C(b - a).$$

Мы доказали, что функция, равная постоянной, интегрируема, и вычислили соответствующий интеграл.

Неправильно думать, что интуитивное представление о существовании площади криволинейной трапеции является доказательством интегрируемости. На самом деле вопрос об интегрируемости — достаточно сложный вопрос. В следующем разделе мы докажем интегрируемость любой функции, монотонной на отрезке $[a, b]$. Это будет одновременно доказательством важного геометрического факта — квадрируемости криволинейной трапеции, ограниченной графиком монотонной функции.

§ 4. Интегрируемость монотонных функций

Пусть на отрезке $[a, b]$ задана монотонная функция (x). Для определенности будем считать ее неубывающей:

$$f(x_1) \leqslant f(x_2) \quad \text{при} \quad a \leqslant x_1 \leqslant x_2 \leqslant b. \tag{12}$$

В этом разделе мы исследуем вопрос об интегрируемости таких функций.

Проведем какое-нибудь разбиение $T(x_k)$ рассматриваемого отрезка. В силу предположения о монотонности (12) для каждого частичного отрезка $[x_{k-1}, x_k]$ будем иметь

$$f(x_{k-1}) \leqslant f(x) \leqslant f(x_k), \quad x \in [x_{k-1}, x_k]. \tag{13}$$

Это позволяет образовать две специальные интегральные суммы:

$$\begin{aligned} s(x_k) &= I(x_k, x_{k-1}) = \\ &= f(x_0)\Delta x_1 + f(x_1)\Delta x_2 + \ldots + f(x_{n-1})\Delta x_n, \\ S(x_k) &= I(x_k, x_k) = \\ &= f(x_1)\Delta x_1 + f(x_2)\Delta x_2 + \ldots + f(x_n)\Delta x_n. \end{aligned} \tag{14}$$

Для первой из них в качестве точек ξ_k выбраны левые концы отрезков $[x_{k-1}, x_k]$ ($\xi_k = x_{k-1}$), для вторых — правые ($\xi_k = x_k$). В силу (13) любая интегральная сумма

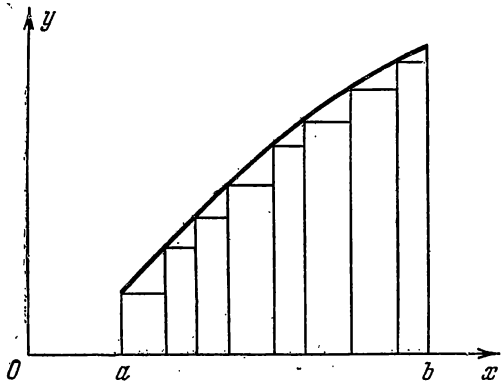

Рис. 48. Геометрическая интерпретация нижней суммы Дарбу как площади ступенчатого многоугольника, содержащегося в криволинейной трапеции.

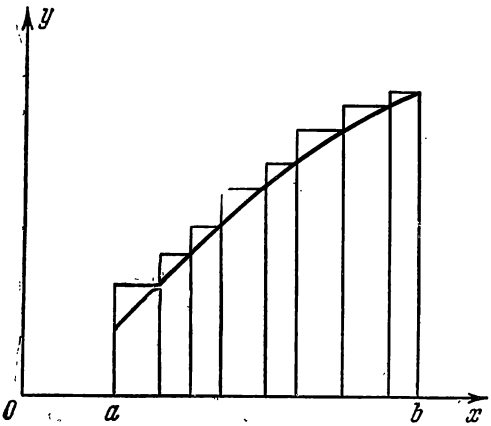

Рис. 49. Геометрическая интерпретация верхней суммы Дарбу как площади ступенчатого многоугольника, содержащего криволинейную трапецию внутри себя.

$I(x_k, \xi_k)$ (10) для того же разбиения удовлетворяет неравенству

$$s(x_k) \leqslant I(x_k, \xi_k) \leqslant S(x_k). \qquad (15)$$

Суммы (14) называют соответственно нижней и верхней суммами Дарбу.

С геометрической точки зрения сумма $s(x_k)$ равна площади ступенчатого многоугольника, содержащегося в криволинейной трапеции, а сумма $S(x_k)$ — площади ступенчатого многоугольника, который содержит криволинейную трапецию внутри себя (см. рис. 48 и 49).

Составим разность между верхней и нижней суммами:

$$S(x_k) - s(x_k) = (f(x_1)) - f(x_0))\Delta x_1 + (f(x_2) - f(x_1))\Delta x_2 + \ldots + (f(x_n) - f(x_{n-1}))\Delta x_n. \quad (16)$$

Она равна сумме площадей прямоугольников, которые

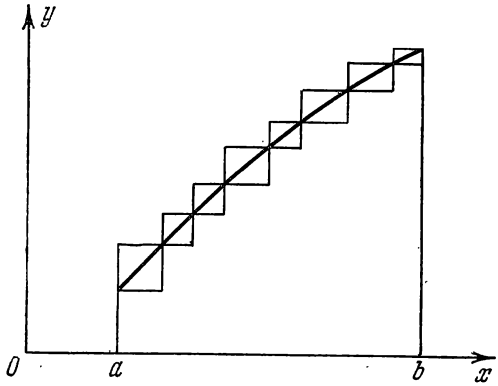

Рис. 50. Прямоугольники, сумма площадей которых равна разности между верхней и нижней суммами Дарбу.

образуются при вычитании первого ступенчатого многоугольника из второго. Эти прямоугольники, как бусы, «нанизаны» на график функции $f(x)$ (см. рис. 50).

Докажем две леммы.

Лемма 1. *Пусть разбиение T_1 получается из разбиения T добавлением нескольких новых точек. Тогда нижняя и верхняя суммы для функции $f(x)$, соответствующие этим разбиениям, удовлетворяют неравенствам*

$$s(x_k) \leqslant s_1(x_k) \leqslant S_1(x_k) \leqslant S(x_k). \quad (17)$$

Неравенства (17) показывают, что измельчение разбиения с помощью добавления новых точек не может уменьшить нижнюю сумму или увеличить верхнюю. В результате при таком измельчении верхние и нижние суммы будут сближаться:

$$S(x_k) - s(x_k) \geqslant S_1(x_k) - s_1(x_k).$$

Доказательство. Пусть разбиение T_1 получается из разбиения T добавлением одной точки x^*, принадлежащей отрезку $[x_{k_0-1}, x_{k_0}]$: $x_{k_0-1} < x^* < x_{k_0}$. Рассмотрим верхнюю сумму $S(x_k)$, соответствующую разбиению $T(x_k)$, и выделим в ней член с номером k_0:

$$f(x_{k_0})\Delta x_{k_0} = f(x_{k_0})(x_{k_0} - x_{k_0-1}). \tag{18}$$

Переход от разбиения T к разбиению T_1 приведет к тому, что вместо члена (18) в сумме $S_1(x_k)$ появятся два новых слагаемых:

$$f(x^*)(x^* - x_{k_0-1}) + f(x_{k_0})(x_{k_0} - x^*). \tag{19}$$

Остальные члены в суммах $S(x_k)$ и $S_1(x_k)$ одинаковы.

Для сравнения выражений (18) и (19) заметим, что в силу монотонности функции $f(x)$ (12) справедливо неравенство: $f(x^*) \leqslant f(x_{k_0})$. Поэтому

$$f(x^*)(x^* - x_{k_0-1}) + f(x_{k_0})(x_{k_0} - x^*) \leqslant \\ \leqslant f(x_{k_0})(x_{k_0} - x_{k_0-1}),$$

т. е. замена слагаемого (18) в сумме $S(x_k)$ на два слагаемых (19) в сумме $S_1(x_k)$ не может увеличить этой сумм по сравнению с $S(x_k)$.

Доказательство для нижних сумм проводится аналогично. Если разбиение T_1 отличается от разбиения T не одной, а несколькими точками, то мы можем добавлять их по одной. Каждое такое добавление не уменьшает нижнюю сумму и не увеличивает верхнюю сумму. Выполнив конечное число шагов, мы перейдем от разбиения T к разбиению T_1 и получим для его нижней и верхней сумм неравенства (17). Лемма доказана.

Лемма 2. *Пусть T_1 и T_2 — два произвольных разбиения отрезка $[a, b]$ и пусть $s_1(x_k)$ и $S_2(x_k)$ — соответственно нижняя сумма для разбиения T_1 и верхняя сумма для разбиения T_2, соответствующие функции $f(x)$. Тогда имеет место неравенство*

$$s_1(x_k) \leqslant S_2(x_k). \tag{20}$$

Неравенство (20) показывает, что для данной функции $f(x)$ нижняя сумма не может превзойти верхнюю. Для одного и того же разбиения это утверждение тривиально, оно непосредственно следует из неравенств (13). В утверждении леммы 2 важно то, что неравенство (20) справедливо также для сумм, соответствующих разным разбиениям.

Доказательство. Рассмотрим разбиение T, которое объединяет все точки разбиений T_1 и T_2. Это означает, что разбиение T является одновременно измельчением разбиений T_1 и T_2 с помощью добавления новых точек. Обозначим через $s(x_k)$ и $S(x_k)$ соответственно нижнюю и верхнюю суммы для разбиения T. Согласно лемме 1 будем иметь:

$$s_1(x_k) \leqslant s(x_k) \leqslant S(x_k) \leqslant S_2(x_k).$$

Опуская два вспомогательных внутренних неравенства, получим (20). Лемма доказана.

Перейдем теперь к формулировке основной теоремы данного параграфа.

Теорема. *Функция $f(x)$, монотонная на отрезке $[a, b]$, интегрируема на этом отрезке.*

Доказательство этой теоремы достаточно сложно. Мы проведем его в три этапа.

1. Рассмотрим последовательность разбиений T_n, построенную следующим образом. Первое разбиение T_1 заключается в делении отрезка $[a, b]$ пополам, второе — в делении на четыре равные части, третье — на восемь равных частей и т. д. Для длин отрезков разбиения T_n получим:

$$\Delta_n = \frac{b-a}{2^n} \to 0 \quad \text{при} \quad n \to \infty. \tag{21}$$

Благодаря равномерности разбиений T_n выражения для верхних и нижних сумм имеют особенно простой вид:

$$S_n(x_k) = (f(x_1) + f(x_2) + \ldots + f(x_n))\,\Delta_n,$$
$$s_n(x_k) = (f(x_0) + f(x_1) + \ldots + f(x_{n-1}))\,\Delta_n.$$

При этом

$$S_n(x_k) - s_n(x_k) = (f(x_n) - f(x_0))\,\Delta_n = (f(b) - f(a))\,\Delta_n,$$

так что в силу (21) получаем:

$$\lim_{n\to\infty} (S_n(x_k) - s_n(x_k)) = 0. \tag{22}$$

Последовательность разбиений T_n строится таким образом, что каждое из них содержит все точки предыдущего разбиения. Согласно леммам 1 и 2 это означает, что последовательности верхних и нижних сумм будут монотонными ограниченными последовательностями:

$$S_n(x_k) \geqslant S_{n+1}(x_k), \quad S_n(x_k) \geqslant s_1,$$
$$s_n(x_k) \leqslant s_{n+1}(x_k), \quad s_n(x_k) \leqslant S_1.$$

По теореме о монотонных ограниченных числовых последовательностях они должны иметь пределы. Согласно (22) их пределы совпадают:

$$\lim_{n\to\infty} S_n(x_k) = \lim_{n\to\infty} s_n(x_k) = \mathcal{J}. \tag{23}$$

При этом последовательность $S_n(x_k)$ приближается к пределу $\mathcal{J}$ сверху, последовательность $s_n(x_k)$ — снизу:

$$S_n(x_k) \geqslant \mathcal{J} \geqslant s_n(x_k). \tag{24}$$

Установлением соотношений (22), (23), (24) заканчивается первый этап доказательства теоремы.

Полученный результат имеет простой геометрический смысл. Мы показали, что при неограниченном измельчении равномерных разбиений отрезка $[a, b]$ последовательности площадей ступенчатых многоугольников, заключенных в криволинейной трапеции и охватывающих ее, стремятся к одному и тому же пределу $\mathcal{J}$. Этот предел принимается за площадь самой криволинейной трапеции. Интегральные суммы, заключенные между нижними и верхними суммами (15), стремятся к тому же пределу $\mathcal{J}$. На следующих этапах доказательства теоремы нам предстоит обобщить этот результат на случай произвольных, не обязательно равномерных разбиений.

2. Докажем теперь следующее утверждение: нижняя и верхняя суммы, соответствующие любому разбиению T, удовлетворяют неравенствам

$$S(x_k) \geqslant \mathcal{J} \geqslant s(x_k). \tag{25}$$

Допустим противное: существует разбиение T^*, для которого $S^*(x_k) < \mathcal{J}$. Положим $\varepsilon = \mathcal{J} - S^*(x_k) > 0$. В силу (23) и (24) для данного ε можно указать такой номер N, что член последовательности нижних сумм s_N, построенной на первом этапе доказательства, удовлетворяет неравенству

$$0 \leqslant \mathcal{J} - s_N(x_k) < \varepsilon = \mathcal{J} - S^*(x_k).$$

Отсюда получаем: $S^*(x_k) < s_N(x_k)$, что противоречит лемме 2. Полученное противоречие доказывает справедливость утверждения для верхних сумм. Для нижних сумм доказательство проводится аналогично.

Геометрический смысл неравенств (25) очевиден: площадь криволинейной трапеции $\mathcal{J}$ больше площади любого заключенного в ней ступенчатого многоугольника $s(x_k)$,

но меньше площади любого охватывающего ступенчатого многоугольника $S(x_k)$. Причем, в отличие от (24), теперь этот результат не связан с предположением о равномерном разбиении отрезка $[a, b]$.

3. Завершая доказательство теоремы, убедимся в том, что число $\mathcal{J}$ является пределом интегральных сумм $I(x_k, \xi_k)$ (10) при $\Delta \to 0$.

Рассмотрим какое-нибудь разбиение $T(x_k)$. Для него интегральные суммы $I(x_k, \xi_k)$, верхняя и нижняя суммы $S(x_k)$, $s(x_k)$ (14) и число $\mathcal{J}$ связаны неравенствами (15) и (25). Вычитая первое из второго, получим:

$$-(S(x_k) - s(x_k)) \leqslant \mathcal{J} - I(x_k, \xi_k) \leqslant S(x_k) - s(x_k).$$

Это двойное неравенство можно заменить одним неравенством с модулем:

$$|\mathcal{J} - I(x_k, \xi_k)| \leqslant S(x_k) - s(x_k). \tag{26}$$

Таким образом, для оценки разности между числом $\mathcal{J}$ и произвольной интегральной суммой $I(x_k, \xi_k)$ для разбиения $T(x_k)$ достаточно оценить разность между верхней и нижней суммами (16) для этого разбиения.

Зададим произвольную точность $\varepsilon > 0$ и сопоставим ей δ, определенное с помощью соотношения

$$\delta = \frac{\varepsilon}{f(b) - f(a)}. \tag{27}$$

Будем считать, что рассматриваемое разбиение $T(x_k)$ удовлетворяет условию $\Delta = \max\{\Delta x_k\} < \delta$, т. е. что оно является достаточно мелким. Оценим разность (16), заменяя величины Δx_k на δ:

$$\begin{aligned} S(x_k) - s(x_k) &< (f(x_1) - f(x_0) + f(x_2) - f(x_1) + \\ &+ f(x_3) - f(x_2) + \ldots + f(x_n) - f(x_{n-1}))\,\delta = \\ &= (f(x_n) - f(x_0))\,\delta = (f(b) - f(a))\frac{\varepsilon}{f(b) - f(a)} = \varepsilon, \end{aligned}$$

т. е.

$$S(x_k) - s(x_k) < \varepsilon. \tag{28}$$

Это неравенство означает, что для монотонных функций $f(x)$ разность между верхней и нижней суммами стремится к нулю при $\Delta \to 0$. Данный результат обобщает вывод (22), полученный на первом этапе доказательства теоремы для специальной последовательности разбиений T_n.

Подставляя (28) в (26), будем иметь

$$|\mathcal{J} - I(x_k, \xi_k)| < \varepsilon. \qquad (29)$$

Таким образом, для произвольной точности ε мы указали такое значение δ (27), что для любого разбиения $T(x_k)$, удовлетворяющего условию $\Delta < \delta$, выполняется неравенство (29). По определению это означает, что число $\mathcal{J}$ является пределом интегральных сумм при $\Delta \to 0$, т. е. что функция $f(x)$, монотонная на отрезке $[a, b]$, интегрируема на этом отрезке. Теорема доказана.

В заключение подчеркнем, что требование монотонности — это достаточное условие интегрируемости, которое, однако, не является необходимым. Например, установленная теорема легко обобщается на класс функций, обладающих на отрезке $[a, b]$ следующим свойством: отрезок $[a, b]$ можно разбить на ряд отрезков $[a, c_1]$, $[c_1, c_2]$, ..., $[c_{n-1}, b]$, на каждом из которых функция $f(x)$ монотонна. Такая функция не будет монотонной на отрезке $[a, b]$ в целом, потому что в точках c_i, $i = 1, 2, \ldots, n-1$, возрастание сменяется на убывание или наоборот.

Этот класс функций достаточно широк. В качестве примера можно привести функцию $y = \sin x$. Любой отрезок $[a, b]$ разбивается для нее на участки возрастания и убывания. Например, если в качестве исходного отрезка мы возьмем отрезок $[0, 2\pi]$, то его можно разделить на три части: на отрезок $[0, \pi/2]$, где $\sin x$ возрастает, отрезок $[\pi/2, 3\pi/2]$, где он убывает, и отрезок $[3\pi/2, 2\pi]$, где $\sin x$ снова возрастает.

Попробуйте сами, опираясь на доказанную теорему и использованные в ней методы, установить интегрируемость таких «кусочно» монотонных функций.

Мы не останавливаемся на обсуждении других предположений, при которых имеет место интегрируемость. Например, можно доказать, что любая функция, непрерывная на отрезке $[a, b]$, интегрируема на этом отрезке. Доказательство данной теоремы вообще никак не связано с понятием монотонности, однако оно опирается на свойства непрерывных функций, которых вы не знаете. Поэтому мы его приводить не будем.

§ 5. Алгоритмы численного интегрирования

Наиболее просто к идее численного интегрирования можно подойти, принимая во внимание определение интеграла как предела интегральных сумм. Если взять ка-

кое-нибудь достаточно мелкое разбиение отрезка $[a, b]$ и построить для него интегральную сумму (10), то ее значение можно приближенно принять за значение соответствующего интеграла.

Пусть, например, отрезок $[a, b]$ разбит на n равных частей длины $h = (b - a)/n$ и в качестве точек ξ_k выбраны средние точки соответствующих отрезков: $\xi_k = a + h(k - 1/2)$, $k = 1, 2, \ldots, n$. В этом случае выражение для интегральной суммы примет вид:

$$I_n = (f(\xi_1) + f(\xi_2) + \ldots + f(\xi_n)) \frac{b-a}{n}.$$

Если функция $f(x)$ интегрируема на отрезке $[a, b]$, то

$$\lim_{n\to\infty} I_n = \mathcal{J} = \int_a^b f(x)\,dx. \tag{30}$$

Согласно (30) выражение для интеграла $\mathcal{J}$ можно записать в виде

$$\mathcal{J} = I_n + \alpha_n, \quad \text{причем} \quad \lim_{n\to\infty} \alpha_n = 0. \tag{31}$$

Пренебрегая величиной α_n, получим приближенную формулу для вычисления интеграла $\mathcal{J}$, которую обычно называют формулой прямоугольников:

$$\mathcal{J} \approx I_n = (f(\xi_1) + f(\xi_2) + \ldots + f(\xi_n)) \frac{b-a}{n}. \tag{32}$$

Причина такого названия связана с геометрической интерпретацией формулы (32). Интеграл $\mathcal{J}$ — это площадь криволинейной трапеции, а интегральная сумма, которой мы приближенно аппроксимируем интеграл,— площадь фигуры, составленной из прямоугольников (см. рис. 51). Все прямоугольники имеют одинаковое основание $h = (b - a)/n$, а их высоты определяются значениями функции $f(x)$ в средних точках ξ_k отрезков разбиения. Разность между площадями этих двух фигур равна α_n, с возрастанием n она стремится к нулю.

Предположим, что функция $f(x)$ имеет на отрезке $[a, b]$ непрерывную вторую производную. В этом случае справедливо следующее утверждение: на отрезке $[a, b]$ существует такая точка x_n^*, что погрешность формулы (32) можно записать в виде

$$\alpha_n = \frac{(b-a)^3}{24n^2} f''(x_n^*). \tag{33}$$

Важная особенность формулы (33) состоит в том, что нам гарантировано существование соответствующей точки x_n^*, но ничего не известно о ее положении. Поэтому формула (33) не позволяет вычислить α_n, но дает возможность ее оценить:

$$|\alpha_n| \leqslant \frac{(b-a)^3}{24n^2} \max |f''(x)|. \qquad (34)$$

Неравенство (34) показывает, что с возрастанием n ошибка в формуле (31) убывает не медленнее, чем $1/n^2$.

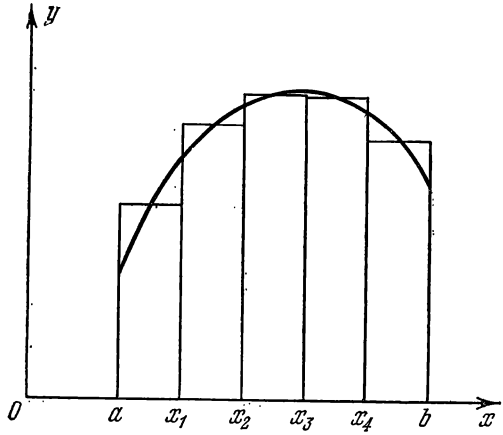

Рис. 51. Геометрическая интерпретация формулы прямоугольников.

Вывод формулы прямоугольников был основан на замене интеграла интегральной суммой (32). Наряду с этим для построения формул численного интегрирования можно использовать другой подход: построить вспомогательную функцию, близкую к подынтегральной функции $f(x)$, и приближенно заменить интеграл от функции $f(x)$ интегралом от вспомогательной функции.

Рассмотрим простейшую реализацию этой идеи. Пусть отрезок $[a, b]$ разбит на n равных частей длины $h = (b - a)/n$ точками $x_k = a + kh$, $k = 0, 1, \ldots, n$; $x_0 = a$, $x_n = b$. Построим функцию $g_n(x)$, определенную следующим образом. На каждом из отрезков $[x_{k-1}, x_k]$ она является линейной, а в граничных точках x_{k-1}, x_k принимает те же значения, что и функция $f(x)$: $f(x_{k-1})$ и $f(x_k)$. Анали-

тически эта функция задается с помощью формул

$$g_n(x) = f(x_{k-1}) + \frac{f(x_k) - f(x_{k-1})}{h}(x - x_{k-1}), \qquad (35)$$
$$x \in [x_{k-1},\ x_k], \quad k = 1,\ 2,\ \ldots,\ n.$$

Ее график представляет собой ломаную линию, начальная, конечная и угловые точки которой принадлежат также графику функции $f(x)$ (см. рис. 52). С увеличением n число общих точек растет и ломаная $y = g_n(x)$ приближается к линии $y = f(x)$.

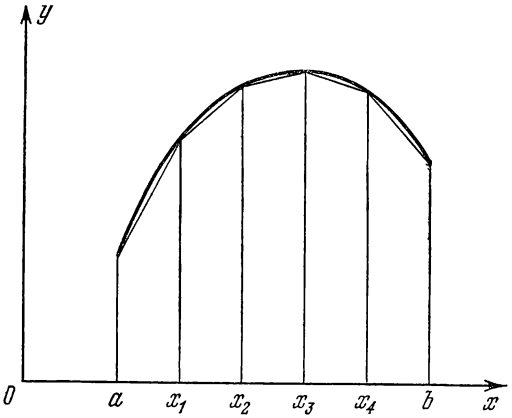

Рис. 52. Геометрическая интерпретация формулы трапеций.

Вычислим с помощью формулы Ньютона — Лейбница интеграл от функции $g_n(x)$ по отрезку $[x_{k-1},\ x_k]$:

$$\int_{x_{k-1}}^{x_k} g_n(x)\,dx =$$
$$= f(x_{k-1}) \int_{x_{k-1}}^{x_k} dx + \frac{f(x_k) - f(x_{k-1})}{h} \int_{x_{k-1}}^{x_k} (x - x_{k-1})\,dx =$$
$$= f(x_{k-1})\,h + \frac{1}{2}(f(x_k) - f(x_{k-1}))\,h =$$
$$= \frac{1}{2}(f(x_{k-1}) + f(x_k))\,h. \qquad (36)$$

Полученный ответ имеет простой геометрический смысл. Интеграл равен площади фигуры, ограниченной графиком

функции $g_n(x)$ (35), осью x и вертикальными линиями $x = x_{k-1}$, $x = x_k$. В данном случае эта фигура является трапецией (см. рис. 51), и ее площадь равна произведению полусуммы оснований на высоту. Однако мы получили этот результат прямым интегрированием, а не ссылкой на формулы геометрии.

Подсчитаем теперь интеграл от функции $g_n(x)$ по всему отрезку $[a, b]$:

$$T_n = \int_a^b g_n(x)\,dx = \int_a^{x_1} g_n(x)\,dx + \int_{x_1}^{x_2} g_n(x)\,dx + \ldots + \int_{x_{n-1}}^{b} g_n(x)\,dx.$$

Подставляя вместо интегралов по отдельным отрезкам $[x_{k-1}, x_k]$ их значения (36), будем иметь:

$$T_n = \left(\frac{1}{2}f(a) + f(x_1) + f(x_2) + \ldots + f(x_{n-1}) + \frac{1}{2}f(b)\right)\frac{b-a}{n}.$$

Полученное выражение можно представить в виде

$$T_n = (f(x_1) + f(x_2) + \ldots + f(x_{n-1}) + f(b))\frac{b-a}{n} + (f(a) - f(b))\frac{b-a}{2n}.$$

Первое слагаемое в этой формуле представляет собой интегральную сумму для функции $f(x)$ в случае, когда в качестве точек ξ_k на отрезках $[x_{k-1}, x_k]$ выбираются правые концы этих отрезков: $\xi_k = x_k$, $k = 1, 2, \ldots, n$. При $n \to \infty$ интегральная сумма стремится к интегралу $\mathcal{J} = \int_a^b f(x)\,dx$. Второе слагаемое при $n \to \infty$ стремится к нулю. Таким образом,

$$\lim_{n\to\infty} T_n = \lim_{n\to\infty} \int_a^b g_n(x)\,dx = \mathcal{J} = \int_a^b f(x)\,dx. \qquad (37)$$

Согласно (37) выражение для интеграла $\mathcal{J}$ можно записать в виде

$$\mathcal{J} = T_n + \beta_n, \quad \text{причем} \quad \lim_{n\to\infty} \beta_n = 0. \qquad (38)$$

Пренебрегая величиной β_n, получим приближенную формулу для вычисления интеграла $\mathcal{J}$, которую обычно на-

зывают формулой трапеций:

$$\mathcal{J} \approx T_n =$$

$$= \left(\frac{1}{2} f(a) + f(x_1) + \ldots + f(x_{n-1}) + \frac{1}{2} f(b)\right) \frac{b-a}{n}. \quad (39)$$

Причина такого названия, как и в предыдущем случае, связана с геометрической интерпретацией формулы. Она приближенно представляет площадь криволинейной трапеции, соответствующей интегралу $\mathcal{J}$, в виде суммы площадей обычных трапеций, которые образуются графиком функции $g_n(x)$ (см. рис. 51). Разность между площадями этих двух фигур равна β_n, с возрастанием n она стремится к нулю.

Предположим, как и при анализе формулы прямоугольников (32), что функция $f(x)$ имеет на отрезке $[a, b]$ непрерывную вторую производную. В этом случае справедливо следующее утверждение: на отрезке $[a, b]$ существует такая точка x_n^{**}, что погрешность формулы (39) можно записать в виде

$$\beta_n = -\frac{(b-a)^3}{12n^2} f''(x_n^{**}). \quad (40)$$

Эта формула совершенно аналогична формуле (33). Она не позволяет вычислить величину β_n, поскольку не известно положение точки x_n^{**}, но дает возможность ее оценить:

$$|\beta_n| \leqslant \frac{(b-a)^3}{12n^2} \max |f''(x)|. \quad (41)$$

Данная оценка отличается от оценки погрешности формулы прямоугольников (34) только числовым множителем. Таким образом, формулы прямоугольников и трапеций характеризуются примерно одинаковой точностью.

Идею, которая была использована при построении формулы трапеций, можно развивать дальше и использовать для получения более точных формул численного интегрирования. Например, если воспользоваться для аппроксимации подынтегральной функции на отдельных отрезках $[x_{k-1}, x_k]$ не линейной функцией $g_n(x)$ (35), а полиномом второй степени, то мы придем к формуле Симпсона:

$$\mathcal{J} = \int_a^b f(x)\,dx \approx S_n =$$

$$= (f(a) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \ldots + 2f(x_{n-2}) +$$

$$+ 4f(x_{n-1}) + f(b)) \frac{b-a}{3n}. \quad (42)$$

В этой формуле, как и в формуле трапеций, предполагается, что отрезок $[a, b]$ разбит на n равных частей длины $h = (b - a)/n$ точками $x_k = a + kh$, $k = 0, 1, \ldots, n$, причем число n должно быть обязательно четным. Значение функции $f(x)$ в нечетных точках разбиения $x_1, x_3, \ldots, x_{n-1}$ входит с коэффициентом 4, в четных $x_2, x_4, \ldots, x_{n-2}$ — с коэффициентом 2 и в двух граничных $x_0 = a$, $x_n = b$ — с коэффициентом 1.

Вывод формулы Симпсона не содержит новых идей по сравнению с выводом формулы трапеций, но является более громоздким. Мы на нем останавливаться не будем.

Обозначим через γ_n погрешность формулы Симпсона. Предположим, что функция $f(x)$ имеет на отрезке $[a, b]$ непрерывную четвертую производную, тогда на данном отрезке найдется такая точка $\bar{x}_n$, что величину γ_n можно записать в виде

$$\gamma_n = -\frac{(b-a)^5}{180n^4} f^{(4)}(\bar{x}_n). \tag{43}$$

Отсюда получаем оценку:

$$|\gamma_n| \leqslant \frac{(b-a)^5}{180n^4} \max |f^{(4)}(x)|. \tag{44}$$

Отметим, что с возрастанием n ошибка убывает как $1/n^4$, т. е. быстрее, чем в формулах прямоугольников и трапеций.

В заключение для иллюстрации изложенного материала снова обратимся к интегралам, рассмотренным в § 2. Первый из них, $\int_1^2 (1/x)\,dx$, был вычислен по формуле Ньютона — Лейбница (9). Подсчитаем его теперь по формулам прямоугольников (32), трапеций (39) и Симпсона (42), выбирая $n = 10$.

В табл. 1 приведены значения подынтегральной функции $f(x) = 1/x$ в точках $\xi_k = 1 + h(k - 1/2)$, $h = (b - a)/n = 0{,}1$; $k = 1, 2, \ldots, 10$, нужные для формулы прямоугольников. Складывая значения функции, приведенные в третьем столбце таблицы, и умножая сумму на $h = 0{,}1$, получим приближенное значение интеграла по формуле прямоугольников:

$$\mathcal{J} \approx I_{10} = 0,\ 692\ 835\ 360.$$

В табл. 2 приведены значения подынтегральной функции $f(x) = 1/x$ в точках $x_k = 1 + kh$, $k = 0, 1, \ldots, n$,

ТАБЛИЦА 1

k	ξ_k	$f(\xi_k)$
1	1,05	0, 952 380 952
2	1,15	0, 869 565 217
3	1,25	0, 800 000 000
4	1,35	0, 740 740 740
5	1,45	0, 689 655 172
6	1,55	0, 645 161 290
7	1,65	0, 606 060 606
8	1,75	0, 571 428 571
9	1,85	0, 540 540 540
10	1,95	0, 512 820 512

ТАБЛИЦА 2

k	x_k	$f(x_k)$
0	1,0	1, 000 000 000
1	1,1	0, 909 090 909
2	1,2	0, 833 333 333
3	1,3	0, 769 230 769
4	1,4	0, 714 285 714
5	1,5	0, 666 666 666
6	1,6	0, 625 000 000
7	1,7	0, 588 235 294
8	1,8	0, 555 555 555
9	1,9	0, 526 315 789
10	2,0	0, 500 000 000

нужные для формулы трапеций и Симпсона. Сложим значения функции в третьем столбце, умножив предварительно нулевую и десятую строчки на 1/2, и умножим полученную сумму на $h = 0{,}1$. В результате получим значение интеграла по формуле трапеций

$$\mathcal{J} \approx T_{10} = 0,\ 693\ 771\ 403.$$

Наконец, воспользовавшись формулой Симпсона (42), будем иметь:

$$\mathcal{J} \approx S_{10} = 0,\ 693\ 150\ 230.$$

В данном случае мы знаем точное значение интеграла:

$$\mathcal{J} = \ln 2 = 0,\ 693\ 147\ 180.$$

Интересно подсчитать ошибки всех трех формул:

$$\alpha_{10} = \mathcal{J} - I_{10} = 0,\ 000\ 311\ 820,$$
$$\beta_{10} = \mathcal{J} - T_{10} = -\ 0,\ 000\ 624\ 223,$$
$$\gamma_{10} = \mathcal{J} - S_{10} = -\ 0,\ 000\ 003\ 050.$$

Как и следовало ожидать, ошибка формулы Симпсона является наименьшей.

Вычислим производные подынтегральной функции $f(x) = 1/x$:

$$f'(x) = -\frac{1}{x^2}; \qquad f''(x) = \frac{2}{x^3};$$

$$f'''(x) = -\frac{6}{x^4}; \qquad f^{(4)}(x) = \frac{24}{x^5}.$$

На отрезке $[1, 2]$ $f''(x) > 0$, $f^{(4)}(x) > 0$, причем их наибольшие значения равны, соответственно, 2 и 24. Найденные ошибки имеют нужные знаки согласно формулам (33), (40), (43), а их величины удовлетворяют неравенствам (34), (41), (44):

$$|\alpha_{10}| \leqslant \frac{2}{24 \cdot 100} = 0,000\,833\,333,$$

$$|\beta_{10}| \leqslant \frac{2}{12 \cdot 100} = 0,001\,666\,666,$$

$$|\gamma_{10}| \leqslant \frac{24}{180 \cdot 10\,000} = 0,000\,013\,333.$$

Рассмотрим теперь второй интеграл: $\mathcal{J} = \int_1^2 (e^{-x}/x)\,dx$. Его нельзя вычислить по формуле Ньютона — Лейбница, но легко подсчитать с помощью формул численного интегрирования.

Воспользуемся для расчета наиболее точной их трех формул — формулой Симпсона, выбирая, как и в предыдущем случае, $n = 10$. Нужные данные приведены в табл. 3. В ее третьем столбце выписаны значения подынтегральной функции $f(x) = e^{-x}/x$ в точках $x_k = 1 + kh$, $h = 0,1$; $k = 0, 1, \ldots, 10$, а в четвертом они умножены на коэффициент 1, 2 или 4 в зависимости от номера k в соответствии с формулой Симпсона (42). Складывая числа из четвертого столбца и умножая сумму на $h/3 = 1/30$, получим

$$\mathcal{J} \approx S_{10} = 0,170\,486\,440.$$

Вычислим четвертую производную функции $f(x) = e^{-x}/x$:

$$f^{(4)}(x) = e^{-x}\left(\frac{1}{x} + \frac{4}{x^2} + \frac{12}{x^3} + \frac{24}{x^4} + \frac{24}{x^5}\right).$$

На отрезке $[1, 2]$ $f^{(4)}(x) > 0$, $\max |f^{(4)}(x)| = 65e^{-1} < 24$. Таким образом, S_{10} дает значение интеграла $\mathcal{J}$ с избытком ($\gamma_{10} < 0$), а для величины ошибки справедлива оценка

$$|\gamma_{10}| < \frac{24}{180 \cdot 10\,000} = 0,000\,013\,333.$$

Попробуйте сами вычислить значения рассматриваемого интеграла по формуле трапеций, используя данные третьего столбца табл. 3. (Не забудьте умножить строки

ТАБЛИЦА 3

k	x_k	$f(x_k)$	$c_k \cdot f(x_k)$
0	1,0	0, 367 879 441	0, 367 879 441
1	1,1	0, 302 010 076	1, 210 440 304
2	1,2	0, 250 995 176	0, 501 990 352
3	1,3	0, 209 639 841	0, 838 559 364
4	1,4	0, 176 140 688	0, 352 281 376
5	1,5	0, 148 753 440	0, 595 013 760
6	1,6	0, 126 185 324	0, 252 370 648
7	1,7	0, 107 460 896	0, 429 843 584
8	1,8	0, 091 832 716	0, 183 665 432
9	1,9	0, 078 720 326	0, 314 881 304
10	2,0	0, 067 667 642	0, 067 667 642

$k = 0$ и $k = 10$ на 1/2.) Сравните полученные результаты. Найдите вторую производную функции $f(x) = e^{-x}/x$, определите по ней знак и оцените величину ошибки формулы трапеций.

В заключение отметим следующее. Оценка погрешности формул численного интегрирования с помощью неравенств типа (34), (41), (44) часто оказывается малоэффективной из-за трудностей, связанных с оценкой производных подынтегральной функции $f(x)$. Поэтому на практике для вычисления ошибки обычно пользуются следующим приемом. Выбирают число n, кратное четырем, находят значения интеграла $\mathcal{J}$ по формуле Симпсона (42) с числом точек n и $n/2$: S_n и $S_{n/2}$ ($n/2$ — целое четное число) и приближенно определяют ошибку численного интегрирования с помощью соотношения

$$\gamma_n = \mathcal{J} - S_n \approx \frac{1}{15}(S_n - S_{n/2}).$$

Глава 8

ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

§ 1. Задача о зеркале прожектора, о колебании маятника и некоторые другие задачи

Знакомство с новым классом математических задач опять начнем с нескольких простых примеров. Мы уже встречали дифференциальные уравнения при обсуждении математических моделей баллистики. Второй пример будет связан с определением формы зеркала прожектора. Соответствующую задачу сформулируем следующим образом: найти такую форму зеркала, чтобы лучи от точечного источника света после отражения в нем образовывали параллельный пучок.

Из соображений симметрии ясно, что поверхность зеркала должна быть поверхностью вращения с осью, проходящей через источник света параллельно отраженным лучам. Проведем через ось какую-нибудь плоскость и рассмотрим сечение нашего зеркала этой плоскостью. Введем в ней систему координат x, y. Начало координат O совместим с источником света, а ось y направим параллельно отраженным лучам (см. рис. 53). Уравнение линии пересечения данной плоскости с зеркалом запишется в виде $y = \varphi(x)$. Наша задача заключается в том, чтобы найти функцию $\varphi(x)$, определяющую форму зеркала.

Рассмотрим рис. 54. Пусть M — произвольная точка искомой линии, ее координаты (x, y), NQ — касательная к линии в точке M, N — точка пересечения касательной с осью y. Луч света, выходящий из точки O, после отражения от зеркала в точке M должен идти параллельно оси y. Согласно закону отражения $\angle OMN = \angle PMQ$. С другой стороны, $\angle ONM = \angle PMQ$, как соответственные при параллельных прямых. Таким образом, в треугольнике OMN углы в вершинах M и N равны, т. е. этот треугольник равнобедренный. В результате будем иметь:

$$ON = OM = \sqrt{x^2 + y^2}.$$

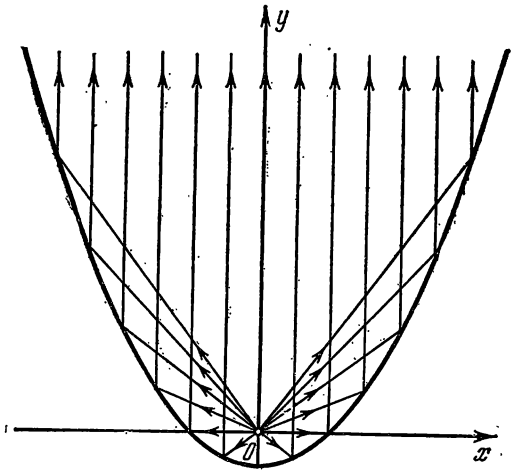

Рис. 53. Ход лучей света в зеркале прожектора.

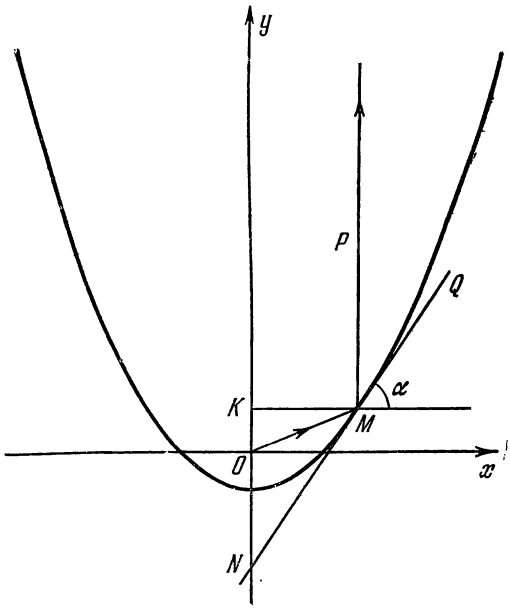

Рис. 54. Геометрическая схема, иллюстрирующая вывод уравнения зеркала.

Теперь рассмотрим ΔKMN и подсчитаем тангенс угла $\angle KMN$, равного углу α:

$$\operatorname{tg}\alpha = \frac{KN}{KM} = \frac{KO + ON}{KM}.$$

Таким образом,

$$\operatorname{tg}\alpha = \frac{y + \sqrt{x^2 + y^2}}{x} = \frac{x}{\sqrt{x^2 + y^2} - y}. \qquad (1)$$

Угол α определяет наклон касательной в точке M к оси x, тангенс этого углу равен, как известно, производной y': $y' = \operatorname{tg}\alpha$. Подставляя это выражение в соотношение (1), получим следующее уравнение:

$$y' = \frac{x}{\sqrt{x^2 + y^2} - y}. \qquad (2)$$

Мы уже говорили в первой главе, что такие уравнения, связывающие независимую переменную x, искомую функцию y и ее производную y', называются дифференциальными уравнениями. Не будем пока обсуждать, как решается уравнение (2), т. е. как из него найти функцию $y = \varphi(x)$, задающую форму зеркала. Об этом будет рассказано ниже. Сейчас же рассмотрим еще одну задачу, которая также приводит к дифференциальному уравнению.

Пусть некоторое тело поместили в среду, в которой поддерживается постоянная температура θ. Как будет изменяться со временем температура тела $T(t)$?

Для того чтобы вывести уравнение для определения функции $T(t)$, нужно знать закон теплообмена. Во многих случаях хорошей математической моделью этого процесса является предположение, что скорость изменения температуры тела пропорциональна разности между температурой среды θ и температурой тела $T(t)$. Из школьного курса математики и физики вы знаете, что скорость изменения какой-нибудь величины определяется ее производной. Таким образом, сформулированное выше предположение можно записать в виде уравнения

$$T' = k(\theta - T), \qquad (3)$$

где k — коэффициент пропорциональности, характеризующий интенсивность теплообмена. Он определяется свойствами тела и окружающей среды.

Соотношение (3), как и (2), является дифференциальным уравнением относительно искомой функции. Оно показывает, что при $T(t) < \theta$, $T'(t) > 0$, т. е. температура тела

возрастает, тело нагревается. Наоборот, при $T(t) > \theta$ $T'(t) < 0$ — температура тела убывает, тело остывает.

Следующий пример, взятый из атомной физики, связан с явлением радиоактивного распада. Экспериментальное исследование этого процесса показывает, что число атомов радиоактивного вещества, спонтанно (самопроизвольно) распадающихся в единицу времени, пропорционально общему числу атомов. Обозначим через $N(t)$ общее число атомов в момент времени t, через λ — коэффициент пропорциональности; тогда сформулированное выше утверждение можно записать в виде дифференциального уравнения:

$$N' = -\lambda N. \tag{4}$$

Уравнение радиоактивного распада (4) с математической точки зрения является частным случаем уравнения (3) и получается из него при $\theta = 0$.

Наш последний пример взят из механики. Это задача о колебании маятника, которая разбирается в учебнике по физике для 10 класса. Однако там авторы обращают основное внимание на выяснение физической картины явления, нас же в первую очередь будет интересовать круг математических вопросов, связанных с его описанием и исследованием.

Маятник — это шарик, подвешенный на нити. Размеры шарика будем считать малыми по сравнению с длиной нити l, так что его можно принять за материальную точку. Нить предположим нерастяжимой. В этом случае шарик будет двигаться по дуге окружности радиуса l (см. рис. 55). Его положение в любой момент времени определяется углом φ, который нить подвеса образует с вертикалью OA (этот угол будем изменять в радианах). Вместо угла φ можно также пользоваться длиной дуги $s = l\varphi$, при этом знак s, совпадающий со знаком φ, показывает, в какую сторону от точки A сместился шарик.

Движение шарика происходит под действием силы тяжести $\boldsymbol{P}$ и натяжения нити $\boldsymbol{F}$ (см. рис. 55). Для анализа этого движения разложим силу $\boldsymbol{P}$ на две составляющие: тангенциальную $\boldsymbol{P}_т$, направленную по касательной к траектории шарика, и нормальную $\boldsymbol{P}_н$, направленную вдоль нити перпендикулярно к траектории. Нормальная составляющая силы тяжести $\boldsymbol{P}_н$ и натяжение нити создают центростремительное ускорение и заставляют шарик двигаться по окружности, меняя все время направление ско-

рости. Однако они не меняют величины скорости. Наоборот, тангенциальная составляющая силы тяжести $P_т$, направленная вдоль скорости, создает тангенциальное ускорение $a_т$ и изменяет величину скорости. Из рис. 55

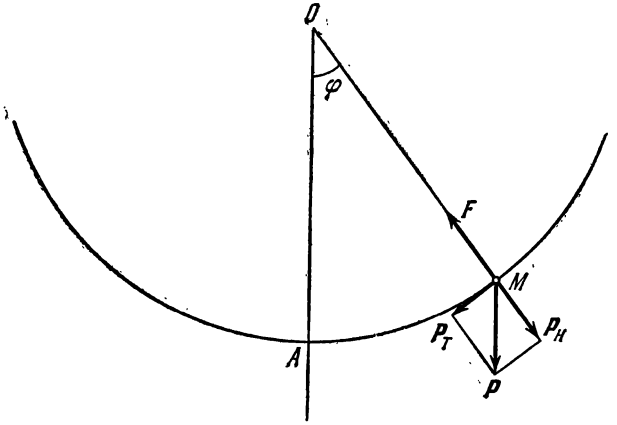

Рис. 55. Геометрическая схема, иллюстрирующая вывод уравнения колебаний маятника.

видно, что величина тангенциальной составляющей вектора P в момент, когда маятник отклонился на угол φ, определяется формулой:

$$P_т = -P \sin \varphi = -mg \sin \varphi. \qquad (5)$$

Знак минус стоит потому, что для положительных углов φ ($\sin \varphi > 0$) тангенциальная составляющая направлена в отрицательную сторону, и наоборот. Это возвращающая сила, которая всегда стремится вернуть шарик в положение равновесия A.

Согласно второму закону Ньютона можно написать:

$$P_т = ma_т. \qquad (6)$$

Для вычисления тангенциального ускорения $a_т$ нужно найти вторую производную от пути $s = l\varphi$ по времени:

$$a_т = s'' = l\varphi''. \qquad (7)$$

При дифференцировании мы учли, что величина l является постоянной и может быть вынесена за знак дифференцирования.

Подставим теперь (5) и (7) в (6), перенесем оба члена в одну сторону и сократим на m. В результате будем иметь:

$$\varphi'' + \frac{g}{l}\sin\varphi = 0. \qquad (8)$$

Это и есть дифференциальное уравнение, описывающее колебания маятника.

Если колебания имеют малую амплитуду, так что маятник отклоняется на небольшие углы, то уравнение (8) можно упростить. При малых углах $\sin\varphi \approx \varphi$, и мы получаем:

$$\varphi'' + \frac{g}{l}\varphi = 0. \qquad (9)$$

Дифференциальное уравнение (9) является уравнением малых колебаний маятника.

Обратите внимание на то, что уравнения (8) и (9) содержат вторую производную искомой функции φ''. Порядок самой старшей производной, входящей в дифференциальное уравнение, называется порядком уравнения. Согласно этому определению уравнения (2), (3) и (4) являются уравнениями первого порядка, уравнения (8) и (9) — уравнениями второго порядка, система (19) главы 1 — системой двух дифференциальных уравнений второго порядка.

Можно предложить множество других задач, приводящих к дифференциальным уравнениям, но в этом нет необходимости. Их важность характеризуется не числом рассмотренных примеров, а типичностью закономерностей, которые удается описать на языке дифференциальных уравнений. Не случайно со времен Ньютона они играют важнейшую роль в теоретической и прикладной математике.

Если в некотором процессе скорость изменения интересующей нас величины определяется значением самой этой величины, то такой процесс описывается дифференциальным уравнением первого порядка. Задача о нагревании тела или о радиоактивном распаде является конкретным примером таких процессов.

В механике второй закон Ньютона устанавливает пропорциональность ускорения тела действующей на него силе. В свою очередь во многих случаях сила определяется положением и скоростью тела. Запись такой связи в математической форме приводит к дифференциальному уравнению или системе дифференциальных уравнений второго порядка. Типичными примерами здесь являются

задачи о колебании маятника или о движении снаряда с учетом сопротивления воздуха. В последнем параграфе главы будет рассмотрен еще один пример, относящийся к космической баллистике: расчет траектории космического корабля в межпланетном пространстве под влиянием сил притяжения Земли, Луны и Солнца. С математической точки зрения такая задача сводится к системе трех дифференциальных уравнений второго порядка.

§ 2. Дифференциальные уравнения первого порядка

Свое знакомство с дифференциальными уравнениями мы начнем с простейшего случая уравнений первого порядка. Такое уравнение можно записать в виде

$$y' = f(x, y), \tag{10}$$

где $f(x, y)$ — некоторая заданная функция независимой переменной x и искомой функции y. Например, для уравнения (2)

$$f(x, y) = \frac{x}{\sqrt{x^2 + y^2} - y}. \tag{11}$$

Всякая функция $y = \varphi(x)$, которая при подстановке в уравнение (10) обращает его в тождество, называется решением этого уравнения. График функции $\varphi(x)$ называется в этом случае интегральной кривой.

Рассмотрим в качестве примера функцию

$$y = \frac{1}{2}(x^2 - 1) \tag{12}$$

и убедимся в том, что она является решением уравнения (2). Для этого вычислим ее производную $y' = x$ и, с другой стороны, подставим эту функцию в правую часть (11) уравнения (2):

$$f(x, y(x)) = \frac{x}{\sqrt{x^2 + \frac{1}{4}(x^4 - 2x^2 + 1)} - \frac{1}{2}(x^2 - 1)} =$$

$$= \frac{x}{\sqrt{\frac{1}{4}(x^4 + 2x^2 + 1)} - \frac{1}{2}(x^2 - 1)} =$$

$$= \frac{x}{\frac{1}{2}\{(x^2 + 1) - (x^2 - 1)\}} = x.$$

В результате мы получаем тождество $x \equiv x$.

Можно также проверить, что функция

$$T(t) = \theta \tag{13}$$

является решением уравнения (3). Действительно, при подстановке ее в левую и правую части уравнения (3) мы получаем тождество $0 = 0$.

Прежде чем продолжать обсуждение уравнения (10), рассмотрим один его частный случай. Предположим, что функция f зависит только от x и не зависит от y, тогда уравнение (10) примет вид

$$y' = f(x), \tag{14}$$

т. е. мы приходим к задаче, обратной дифференцированию: к определению функции $y(x)$ по ее производной $y' = f(x)$. Эта задача вам известна, ее общее решение дается формулой

$$y = F(x) + C, \tag{15}$$

где $F(x)$ — какая-нибудь первообразная функции $f(x)$.

Возвращаясь теперь к обсуждению дифференциального уравнения (10) в общем виде, выясним его геометрический смысл. Примем x и y за декартовы координаты на плоскости. Дифференциальное уравнение (10) ставит в соответствие каждой точке (x, y) определенное значение производной y'. Вспомним, что с геометрической точки зрения производная характеризует направление касательной к графику функции: она равна тангенсу угла наклона касательной к оси x. Таким образом, можно сказать, что дифференциальное уравнение (10) определяет в каждой точке (x, y) некоторое направление, образующее с осью x угол, тангенс которого равен $f(x, y)$. Изобразим эти направления с помощью стрелок. Соответствующие графики для уравнений (2) и (3) приведены на рис. 56 и 57.

С учетом сделанных замечаний задачу решения или, как принято говорить, интегрирования дифференциального уравнения (10) можно интерпретировать следующим образом: требуется найти на плоскости x, y такую кривую $y = \varphi(x)$, чтобы ее касательная в каждой точке имела заданное направление. Грубо говоря, график искомой функции должен всюду касаться расставленных стрелок. Из рис. 56 и 57 интуитивно ясно, что таких кривых можно провести бесчисленное множество.

Обратимся, например, к рис. 57. Будем считать, что мы внесли тело в среду с температурой θ в момент

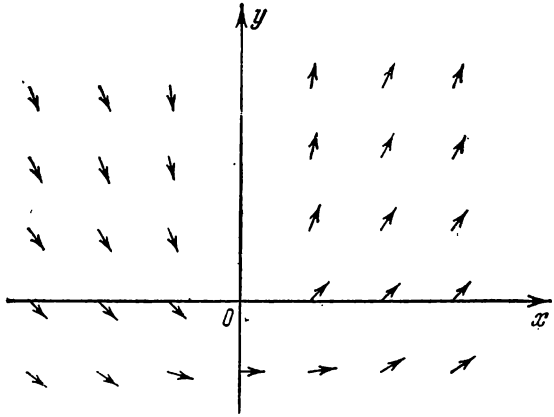

Рис. 56. Направления, определяемые на плоскости x, y дифференциальным уравнением (2).

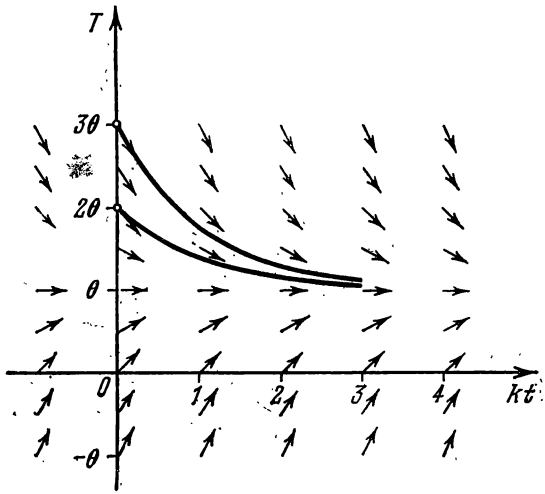

Рис. 57. Направления, определяемые на плоскости t, T дифференциальным уравнением (3). Приведены две интегральные кривые. Одна из них соотвествует случаю, когда температура тела при $t = 0$ равна 2θ, вторая — 3θ.

времени $t = 0$, причем температура внесенного тела была равна 2θ. Точка с координатами $0, 2\theta$ выделена на рис. 57 кружком. Проведя через эту точку кривую вдоль расставленных стрелок, мы получим решение уравнения (3), описывающее изменение температуры тела со временем. Эта кривая показана на рис. 57.

Предположим теперь, что температура внесенного тела была равна не 2θ, а 3θ. Тогда, чтобы узнать дальнейшее изменение температуры тела в этом случае, нужно провести кривую вдоль стрелок, начиная с точки $(0, 3\theta)$. Эта кривая также приведена на рис. 57. Она дает еще одно решение уравнения (3), отличное от предыдущего.

Вообще, задавая совершенно произвольно начальное значение температуры тела T_0: $T(0) = T_0$, мы получим определенное решение уравнения (3), описывающее последующее изменение температуры. Такое условие содержит дополнительные сведения об искомом решении, которые не вытекают из дифференциального уравнения (3). Его называют «начальным» условием. Решение уравнения (3), удовлетворяющее заданному начальному условию, является единственным, оно однозначно определяется этим условием. Таким образом, уравнение (3) описывает все возможные тепловые режимы тела, помещенного в среду с температурой θ, теплообмен с которой подчиняется сформулированному выше предположению. Конкретный же ход процесса нагрева или остывания тела в каждом случае определяется его начальной температурой.

Наши рассуждения относились к определенному уравнению (3) и носили интуитивный характер. Они должны были помочь понять специфику дифференциальных уравнений.

Сформулируем теперь следующую математическую задачу: найти решение дифференциального уравнения (10) $y = \varphi(x)$, принимающее в точке $x = x_0$ заданное значение y_0:

$$\varphi(x_0) = y_0. \tag{16}$$

Задачу подобного типа для дифференциального уравнения (10) принято называть задачей с начальными условиями или задачей Коши.

В теории дифференциальных уравнений доказывается, что при некоторых ограничениях на функцию $f(x, y)$ решение задачи Коши для уравнения (10) существует и является единственным. (Условия, которым должна удов-

летворять функция $f(x, y)$, и идея одного из возможных доказательств утверждения приводятся в следующем параграфе.) Это — очень важный результат. Он обосновывает постановку задачи Коши. Он показывает, что множество всех решений дифференциального уравнения (10), которое принято называть общим решением, зависит от одного параметра, т. е. имеет вид

$$y = \varphi(x, C).$$

За параметр C можно, например, принять начальное значение y_0, однако это необязательно. Придавая C разные значения, мы будем получать из общего решения различные частные решения.

Для некоторых типов дифференциальных уравнений, соответствующих специальному выбору функции $f(x, y)$, общее решение удается получить в виде явной формулы. К числу таких простейших уравнений относятся, в частности, уравнения (2), (3), (4). Познакомимся с этими решениями, чтобы закончить рассмотрение примеров предыдущего параграфа.

Общее решение уравнения (2) имеет вид

$$y = \frac{C}{2}x^2 - \frac{1}{2C}, \tag{17}$$

где C — произвольная постоянная. (При $C = 1$ из формулы (17) получаем частное решение (12), которое мы обсуждали раньше.) Графиком функции (17) при любом C является парабола. Поверхность, которая получается при вращении параболы вокруг ее оси, называется параболоидом. Свойство параболического зеркала собирать лучи в параллельный пучок используется в прожекторах.

Общее решение уравнения (3) дается формулой

$$T = \theta + Ce^{-kt}. \tag{18}$$

Отметим следующую особенность этого решения: при $t \to \infty$ второй член стремится к нулю, так что температура тела $T(t)$ выравнивается с температурой среды θ (см. кривые на рис. 57).

Полагая в формуле (18) $\theta = 0$ и заменяя $T(t)$ и k на $N(t)$ и λ, получим общее решение уравнения (4):

$$N = Ce^{-\lambda t}. \tag{19}$$

Формула показывает, что число атомов радиоактивного вещества в результате их распада экспоненциально убывает со временем.

Подсчитаем промежуток времени t_0, за который число атомов уменьшается в два раза:

$$\frac{N(t+t_0)}{N(t)} = \frac{Ce^{-\lambda(t+t_0)}}{Ce^{-\lambda t}} = e^{-\lambda t_0} = \frac{1}{2}.$$

Отсюда получаем: $t_0 = (1/\lambda) \ln 2$. Этот промежуток времени называется периодом полураспада (см. учебник по физике для 10 класса). Напомним, что коэффициент λ и, соответственно, период полураспада t_0 зависят от того, атомы какого радиоактивного элемента мы рассматриваем. Например, для радия период полураспада равен 1600 лет.

§ 3. Метод ломаных Эйлера

Перейдем к обсуждению основного вопроса: как решаются дифференциальные уравнения? В конце предыдущего раздела мы говорили, что в некоторых случаях их удается проинтегрировать в явном виде. В качестве примера приводились уравнения (2), (3), (4). В течение трех веков, прошедших с момента появления дифференциальных уравнений в математике и физике, разработке аналитических методов уделялось самое серьезное внимание. Они сыграли важную роль в изучении дифференциальных уравнений, позволили рассмотреть множество прикладных задач. В учебниках и справочниках можно найти подробное перечисление различных типов уравнений, которые интегрируются в явном виде. Для каждого из них указывается свой прием решения. Все эти методы носят крайне ограниченный, частный характер и накладывают жесткие ограничения на конкретный вид функции $f(x, y)$. Например, для одного метода необходимо, чтобы она распадалась на произведение двух множителей:

$$f(x, y) = g(x) \cdot h(y),$$

другой предполагает линейную зависимость от y:

$$f(x, y) = p(x)\, y + q(x)$$

и т. д.

Однако в большинстве случаев функция $f(x, y)$ не втискивается в прокрустово ложе таких специальных вариантов, и получить явную формулу для решения не удается. В этих условиях особенно важное значение приобретают численные методы. В отличие от аналитических методов, они являются универсальными и не накла-

дывают ограничений на конкретный вид функции $f(x, y)$. К тому же они сразу дают ответ в виде таблицы чисел, т. е. в форме, удобной для практического применения.

Численные методы решения дифференциальных уравнений начали разрабатываться давно. Их появление стимулировали в первую очередь задачи астрономии, требовавшие большой точности при расчете траекторий движения планет и комет. Вспомним еще раз историю открытия Лаверье планеты Нептун. Однако применение численных методов требует больших вычислений, поэтому их широкое использование стало возможным только теперь — благодаря появлению ЭВМ. В данном разделе мы расскажем о простейшем численном методе решения дифференциальных уравнений, идея которого принадлежит Эйлеру.

Пусть нам нужно на некотором отрезке $x_0 \leqslant x \leqslant a$ найти решение дифференциального уравнения (10), удовлетворяющее при $x = x_0$ начальному условию (16). Возьмем некоторое целое положительное число n и разделим отрезок $[x_0, a]$ на n частей. Точки деления x_k, $k = 0, 1, 2, \ldots, n$, будут при этом иметь координаты:

$$x_k = x_0 + kh, \quad \text{где} \quad h = \frac{a - x_0}{n}. \tag{20}$$

При $k = 0$ формула (20) дает x_0, а при $k = n$ дает $x_n = a$.

Подставим начальные значения x_0, y_0 в выражение для функции f и рассмотрим на первом частичном отрезке $[x_0, x_1]$ вспомогательное дифференциальное уравнение:

$$y' = f(x_0, y_0). \tag{21}$$

При предположении о непрерывности функции $f(x, y)$ уравнение (21) в окрестности точки (x_0, y_0) близко к уравнению (10).

Будем искать решение уравнения (21), удовлетворяющее тем же начальным условиям (16). Согласно (21) производная искомого решения $y(x)$ является постоянной (она равна значению функции f в некоторой фиксированной точке (x_0, y_0)). Это означает, что функция $y(x)$ представляет собой линейную функцию $y = kx + b$ с угловым коэффициентом $k = f(x_0, y_0)$. Свободный член b определяется из начальных условий, так что окончательно получаем:

$$y = f(x_0, y_0)(x - x_0) + y_0, \quad x \in [x_0, x_1]. \tag{22}$$

Действительно, полагая в формуле (22) $x = x_0$, имеем $y(x_0) = y_0$.

Вычислим по формуле (22) значение функции $y(x)$ в граничной точке x_1 рассматриваемого частичного отрезка $[x_0, x_1]$:

$$y(x_1) = f(x_0, y_0)\, h + y_0 = y_1 \tag{23}$$

и, принимая x_1, y_1 за новые начальные значения, повторим ту же процедуру. Подставим эти значения в выражение для функции f и рассмотрим на втором отрезке $[x_1, x_2]$ дифференциальное уравнение

$$y' = f(x_1, y_1). \tag{24}$$

Решение уравнения (24) $y(x)$, принимающее при $x = x_1$ значение $y = y_1$, записывается по формуле, аналогичной (22):

$$y = f(x_1 y_1)\,(x - x_1) + y_1, \quad x \in [x_1, x_2]. \tag{25}$$

При этом в точке $x = x_2$ функция (25) принимает значение

$$y(x_2) = f(x_1, y_1)\, h + y_1 = y_2. \tag{26}$$

Будем продолжать этот процесс. В результате на k-м шаге, решая на отрезке $[x_{k-1}, x_k]$ дифференциальное уравнение

$$y' = f(x_{k-1}, y_{k-1}) \tag{27}$$

с начальным условием

$$y(x_{k-1}) = y_{k-1}, \tag{28}$$

получим следующее выражение для функции $y(x)$:

$$y = f(x_{k-1}, y_{k-1})(x - x_{k-1}) + y_{k-1}, \quad x \in [x_{k-1}, x_k]. \tag{29}$$

В частности, значение функции $y(x)$ в граничной точке x_k рассматриваемого частичного отрезка $[x_{k-1}, x_k]$ имеет вид

$$y(x_k) = f(x_{k-1}, y_{k-1})\, h + y_{k-1} = y_k. \tag{30}$$

Эта рекуррентная формула составляет основу рассматриваемого вычислительного процесса. Она позволяет по y_{k-1}—значению функции $y(x)$ в точке $x = x_{k-1}$—найти y_k — значение функции $y(x)$ в следующей точке $y = y_k$. Применяя ее последовательно n раз, мы дойдем до последней точки $x_n = a$. В результате получим функцию $y(x)$,

определенную на отрезке $[x_0, a]$. На каждом из частичных отрезков $[x_{k-1}, x_k]$ она задается формулой (29), т. е. является линейной функцией. При переходе от одного отрезка к другому наклон прямой меняется. График этой функции представляет собой ломаную линию, состоящую из звеньев с вершинами в точках $(x_k, y_k), k = 0, 1, 2, \ldots, n$. Построенную таким образом линию принято называть ломаной Эйлера. Подчеркнем особо, что процедура вычислений, необходимых для построения ломаной Эйлера, универсальна, она не зависит от конкретного вида функции $f(x, y)$.

Предположим, что функция $f(x, y)$ непрерывна и удовлетворяет по переменной y условию Липшица:

$$|f(x, y_1) - f(x, y_2)| \leqslant \alpha |y_1 - y_2|.$$

В теории дифференциальных уравнений доказывается, что в этом случае последовательность ломаных Эйлера при неограниченном увеличении числа звеньев n (т. е. при $h \to 0$) стремится к определенному пределу $y = \varphi(x)$, который является единственным решением задачи Коши (10), (16).

Доказательство этого утверждения требует специального аппарата, поэтому мы ограничимся лишь его обсуждением. Теорема устанавливает однозначную разрешимость задачи Коши. Это — один из центральных результатов теории дифференциальных уравнений. Чрезвычайно важно также то, что доказательство существования решения является конструктивным: указывается конкретный процесс (построение ломаных Эйлера), который сходится к искомому решению. Тем самым метод ломаных Эйлера получает обоснование как численный метод интегрирования дифференциальных уравнений. Действительно, выбирая достаточно малый шаг h и проводя вычисления по описанной выше схеме, мы построим ломаную, которая, в силу сходимости процесса, будет близка к решению задачи. Чем выше требуемая точность, тем меньший шаг следует выбирать. Это, естественно, потребует увеличения объема вычислений. Вы уже знаете, что за точность нужно «платить».

В качестве примера, иллюстрирующего эти соображения, обратимся к рис. 58. На нем построена интегральная кривая дифференциального уравнения (2), соответствующая начальному условию $y(0) = -0{,}5$ (линия III). Уравнение (12) этой кривой известно. На том же рисунке

для отрезка $0 \leqslant x \leqslant 3$ построены ломаные Эйлера. Одна из них (линия I) соответствует сравнительно большому шагу $h_1 = 0{,}5$. Мы видим, что, хотя она качественно правильно передает ход интегральной кривой, точность в данном случае невелика. Линия II соответствует более мелкому шагу $h_2 = 0{,}1$. Уменьшение шага в пять раз приводит к существенному увеличению точности. Теперь максимальное отклонение приближенного решения от точного не превышает 0,55. Отдельные звенья этой ломаной настолько мелки, что их невозможно показать на рисунке, линия выглядит гладкой. Если уменьшить шаг еще в десять раз: $h_3 = 0{,}01$, то мы получим ломаную Эйлера, которая приближает решение задачи с ошибкой, не превышающей 0,062. При выбранном масштабе она практически сливается с точным решением (линия III). Таким образом, выбирая достаточно маленький шаг h, мы всегда можем получить интегральную кривую с любой нужной степенью точности.

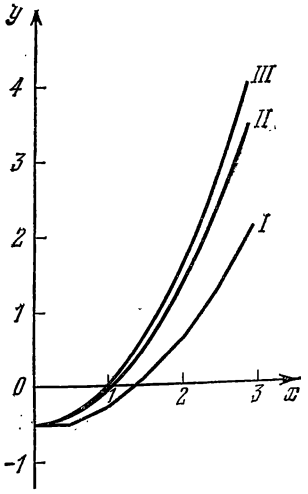

Рис. 58. Ломаные Эйлера, построенные для дифференциального уравнения (2) при начальном условии $y(0) = -0{,}5$. Линия I соответствует шагу $h_1 = 0{,}5$, линия II — шагу $h_2 = 0{,}1$, линия III — шагу $h_3 = 0{,}01$. При выбранном масштабе линия III практически сливается с точным решением задачи $y = \frac{1}{2}(x^2 - 1)$.

Весь процесс вычислений по методу Эйлера основан, в конечном счете, на последовательном применении формулы (30) и легко может быть выполнен на ЭВМ. Он сводится к многократному вычислению значений функции f, которая стоит в правой части дифференциального уравнения (10), в различных точках (x_k, y_k).

В заключение заметим следующее. Существуют различные модификации метода Эйлера, которые имеют более высокую точность. Это позволяет при заданной допустимой погрешности вести расчеты с более крупным шагом h, что уменьшает объем необходимых вычислений. Однако обсуждение методов повышенной точности — специальный вопрос, и мы на нем останавливаться не будем.

§ 4. Дифференциальные уравнения высших порядков и системы дифференциальных уравнений

До сих пор мы обсуждали уравнения первого порядка. Однако в первом параграфе мы встречались с дифференциальными уравнениями второго порядка (8) и (9), содержащими вторые производные искомой функции. Возможны дифференциальные уравнения третьего, четвертого и вообще любого порядка.

Рассмотрим в качестве примера уравнение второго порядка. В общем случае такое уравнение можно записать в виде

$$y'' = f(x, y, y'). \tag{31}$$

Здесь f — некоторая заданная функция своих аргументов; независимой переменной x, искомой функции y и ее первой производной y'.

Всякая функция $y = \varphi(x)$, которая при подстановке в уравнение (31) обращает его в тождество, называется решением этого уравнения. Уравнение (31) имеет бесчисленное множество решений. Если общее решение уравнения первого порядка (16) содержало одну произвольную постоянную, то общее решение уравнения второго порядка зависит от двух произвольных постоянных:

$$y = \varphi(x, C_1, C_2). \tag{32}$$

В соответствии с этим для однозначного определения решения уравнения (31) нужно задать не одно, а два начальных условия.

Задача с начальными условиями формулируется в этом случае следующим образом. Найти решение дифференциального уравнения (31) $y = \varphi(x)$, которое при $x = x_0$ принимает вместе со своей первой производной заданные значения

$$\varphi(x_0) = y_0, \quad \varphi'(x_0) = y_0'. \tag{33}$$

Таким образом, для однозначного определения интегральной кривой уравнения второго порядка нужно задать начальную точку (x_0, y_0) и направление интегральной кривой в этой точке, которое определяется производной $\varphi'(x_0) = y_0'$.

Наряду с отдельными дифференциальными уравнениями рассматриваются также системы дифференциальных уравнений. Возьмем в качестве примера систему двух

дифференциальных уравнений первого порядка. Такую систему можно записать в виде

$$\begin{aligned} y' &= f_1(x, y, z), \\ z' &= f_2(x, y, z). \end{aligned} \tag{34}$$

Здесь f_1 и f_2 — заданные функции независимой переменной x и искомых функций y и z. Пару функций

$$y = \varphi(x), \quad z = \psi(x), \tag{35}$$

которая при подстановке в систему (34) обращает оба уравнения в тождества, называют решением этой системы.

Если рассматривать x, y и z как декартовы координаты в пространстве и построить геометрическое место точек, которое определяет решение (35), то мы получим некоторую пространственную кривую. Ее называют интегральной кривой системы (34).

Задача с начальными условиями, позволяющая выделить единственное решение системы (34) из бесчисленного множества возможных, формулируется следующим образом: найти решение (35) системы дифференциальных уравнений (34), принимающее при $x = x_0$ заданные значения

$$\varphi(x_0) = y_0, \quad \psi(x_0) = z_0. \tag{36}$$

Иными словами, задача с начальными условиями, или задача Коши, заключается в том, чтобы выделить интегральную кривую, проходящую через заданную точку пространства (x_0, y_0, z_0).

Мы говорили для простоты о системах двух дифференциальных уравнений с двумя неизвестными функциями. Однако совершенно аналогично можно рассматривать системы n уравнений с n неизвестными функциями. Такая форма записи задачи является наиболее общей в теории дифференциальных уравнений. Дифференциальные уравнения высших порядков всегда могут быть сведены к системе уравнений первого порядка. Посмотрим, как это делается на примере, уравнения второго порядка (31).

Введем в уравнении (31) новую неизвестную функцию с помощью соотношения

$$z = y'. \tag{37}$$

Продифференцировав это равенство по x, получим:

$$z' = y'' = f(x, y, z). \tag{38}$$

В результате уравнение (31) мы можем заменить системой двух уравнений первого порядка

$$y' = z, \qquad (39)$$
$$z' = f(x, y, z),$$

которая является частным случаем системы (34).

Если для уравнения (31) поставлена задача Коши с начальными условиями (33), то в результате сделанного преобразования она сводится к задаче Коши (39), (36), где, согласно (37),

$$z_0 = y_0'. \qquad (40)$$

Обсуждая уравнения первого порядка, мы отмечали, что аналитическое решение для них возможно только в некоторых специальных случаях. Это замечание тем более справедливо для уравнений высших порядков и систем. В то же время такие численные методы, как метод Эйлера, легко могут быть использованы и в этом случае. Его применение к системам первого порядка является естественным обобщением процедуры, описанной в § 3. На описании деталей мы останавливаться не будем.

§ 5. Задача о колебании маятника

Теперь, когда мы коротко познакомились с дифференциальными уравнениями высших порядков, вернемся к задаче о колебаниях маятника.

Мы уже видели в § 1, что этот процесс описывается дифференциальным уравнением (8), которое в случае малых колебаний упрощается и принимает вид (9). Такое же уравнение получается при описании ряда других колебательных процессов. Кратко оно обсуждается в учебнике математики 10 класса. С анализа этого уравнения мы и начнем.

Важной особенностью уравнения (9) является то, что неизвестная функция φ и ее производная входят в него линейно с коэффициентами, не зависящими от времени. Такие уравнения называются линейными уравнениями с постоянными коэффициентами, они представляют собой самый простой тип дифференциальных уравнений.

Общее решение уравнения (9) имеет вид

$$\varphi = C_1 \cos \omega_0 t + C_2 \sin \omega_0 t, \qquad (41)$$

где

$$\omega_0 = \sqrt{g/l}. \qquad (42)$$

Оно зависит, как это и должно быть для уравнений второго порядка, от двух произвольных постоянных. Придавая им различные численные значения, мы будем получать различные частные решения.

Проверим, что функция (41) действительно удовлетворяет уравнению (9). Ее дифференцирование дает

$$\varphi' = -C_1\omega_0 \sin \omega_0 t + C_2\omega_0 \cos \omega_0 t. \qquad (43)$$

Здесь мы учли, что производная синуса равна косинусу и производная косинуса — минус синусу. Повторное дифференцирование функции (43) позволяет вычислить вторую производную

$$\varphi'' = -C_1\omega_0^2 \cos \omega_0 t - C_2\omega_0^2 \sin \omega_0 t. \qquad (44)$$

Подставляя выражения (41) и (44) в левую часть уравнения (9) и принимая во внимание формулу (42) для частоты ω_0, получаем нуль.

Формула (41) показывает, что движение маятника, подчиняющегося уравнению (9), представляет собой гармонические колебания с круговой частотой ω_0 (42) и периодом

$$T_0 = 2\pi/\omega_0 = 2\pi\sqrt{l/g}. \qquad (45)$$

Постоянные C_1 и C_2 в общем решении (41) определяются из начальных условий. Предположим, например, что в некоторый момент времени, который мы условно примем за момент $t = 0$, маятник отклонили на некоторый угол φ_0 и спокойно, без толчка отпустили. Такой способ возбуждения колебаний можно описать с помощью начальных условий

$$\varphi(0) = \varphi_0, \quad \varphi'(0) = 0. \qquad (46)$$

Первое из этих соотношений очевидно, второе означает, что начальная скорость маятника принимается равной нулю.

Найдем частное решение, описывающее колебания маятника при таком способе возбуждения. Для этого положим в формуле (41) $t = 0$ и воспользуемся первым из условий (46). В результате будем иметь

$$\varphi(0) = C_1 = \varphi_0. \qquad (47)$$

Совершенно аналогично положим в формуле (43) $t = 0$ и воспользуемся вторым из условий (46). Это определяет

постоянную C_2:

$$\varphi'(0) = C_2\omega_0 = 0. \qquad (48)$$

Подставляя найденные значения C_1 и C_2 в формулу (41), получим нужное частное решение

$$\varphi = \varphi_0 \cos \omega_0 t. \qquad (49)$$

Совершенно аналогично можно рассмотреть другую задачу. Пусть маятник находится в положении равновесия (в точке A на рис. 55). В некоторый момент времени,

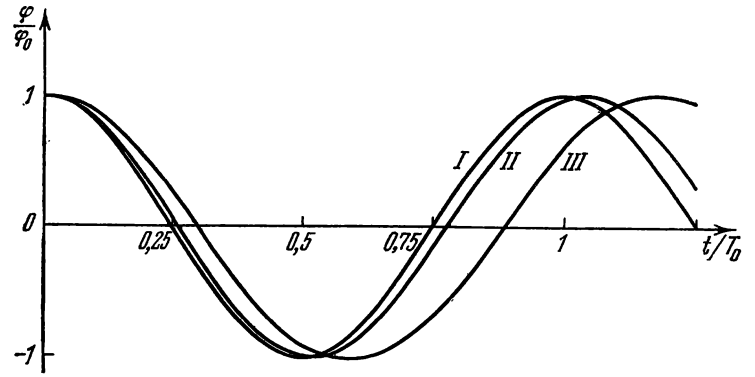

Рис. 59. Интегральные кривые уравнения (8) с начальными условиями (46) при различных начальных отклонениях маятника.

который мы опять примем условно за момент $t = 0$, ему резким ударом сообщили начальную скорость v_0. Попробуйте сами сформулировать для этого случая начальные условия и выделить из общего решения (41) соответствующее частное решение.

Перейдем теперь к обсуждению уравнения (8), которое описывает колебания с произвольной амплитудой. Данное уравнение не является линейным (неизвестная функция φ входит под знак синуса), и это существенно усложняет задачу. Уравнение (8) можно решить аналитически, однако ответ записывается через особый класс специальных функций (так называемые эллиптические интегралы), с которым вы незнакомы. В то же время все интересующие нас вопросы легко могут быть изучены, если воспользоваться численными методами.

Рассмотрим рис. 59. На нем приведены три кривые, полученные с помощью численного решения уравнения (8)

с начальными условиями (46) при разных начальных отклонениях маятника φ_0.

Кривая I соответствует $\varphi_0 = \pi/36$ (т. е. 5°). При такой маленькой амплитуде с высокой степенью точности справедливо линейное уравнение (9), поэтому рассчитанная кривая практически представляет график функции (49). Разница между ними настолько мала, что не может быть показана в масштабе рис. 59.

Кривые II и III соответствуют начальным амплитудам $\varphi_0 = \pi/4$ и $\varphi_0 = \pi/2$ (т. е. 45° и 90°). В этом случае линейное приближение малых колебаний уже не справедливо, и полученные решения отличаются от (49). Колебания перестают быть чисто гармоническими, а их период T оказывается зависящим от амплитуды: $T = T(\varphi_0)$.

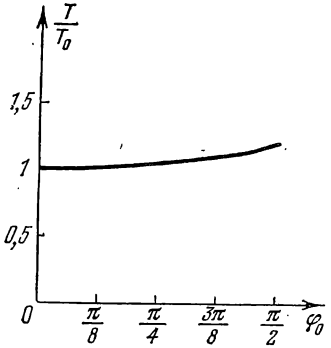

Рис. 60. Зависимость периода колебаний маятника от амплитуды.

На рис. 60 приведен график, характеризующий эту зависимость. При малых амплитудах, когда справедливо уравнение (9) и вытекающие из него следствия, $T = T_0 = 2\pi\sqrt{l/g}$. Однако с увеличением φ_0 условия применимости линейного приближения начинают нарушаться, при этом период колебаний T возрастает и при $\varphi_0 = \pi/2$ он превышает T_0 на 18%.

Заметим, что зависимость периода колебаний от амплитуды — специфическая особенность нелинейных колебательных систем. При их исследовании всегда интересуются кривой типа, изображенной на рис. 60, которая является одной из наиболее важных характеристик.

§ 6. Расчет траектории ядра с учетом сопротивления воздуха

В первой главе была построена математическая модель баллистики пушечного ядра с учетом сопротивления воздуха, которая свелась к системе двух дифференциальных уравнений второго порядка (19) на стр. 22. Для завершения постановки задачи необходимо сформулировать соответствующие начальные условия. Примем момент

вылета ядра из орудия за $t = 0$, тогда в системе координат, описанной в § 1 главы 1 (начало координат совмещено с орудием, ось x направлена горизонтально, ось y — вертикально вверх), начальные условия имеют вид

$$x(0) = 0, \quad y(0) = 0, \tag{50}$$

$$x'(0) = v_0 \cos \alpha, \quad y'(0) = v_0 \sin \alpha, \tag{51}$$

где v_0 — величина начальной скорости, α — угол возвышения.

Предположим, что плотность воздуха ρ можно считать постоянной. В этом случае система (19) главы 1 не содержит явно искомых функций $x(t)$, $y(t)$, в нее входят только производные. Это позволяет снизить порядок уравнений со второго до первого.

Введем в качестве искомых функций компоненты скорости:

$$u(t) = v_x(t) = x'(t), \quad w(t) = v_y(t) = y'(t), \tag{52}$$

тогда

$$u'(t) = x''(t), \quad w'(t) = y''(t). \tag{53}$$

Подставляя (52) и (53) в (19) главы 1, получим систему двух уравнений первого порядка:

$$\begin{aligned} u' &= -\frac{C\pi}{2}\frac{R^2\rho}{m}\sqrt{u^2+w^2}\,u, \\ w' &= -g-\frac{C\pi}{2}\frac{R^2\rho}{m}\sqrt{u^2+w^2}\,w, \end{aligned} \tag{54}$$

которую нужно решать при начальных условиях, вытекающих из (51):

$$u(0) = v_0 \cos \alpha, \quad w(0) = v_0 \sin \alpha. \tag{55}$$

После того как в результате решения задачи Коши (54), (55) найдены компоненты скорости: $u(t) = v_x(t)$, $w(t) = v_y(t)$, функции $x(t)$, $y(t)$ определяются по ним простым интегрированием:

$$x(t) = \int_0^t u(\tau)\,d\tau, \quad y(t) = \int_0^t w(\tau)\,d\tau, \tag{56}$$

при этом будут автоматически выполняться начальные условия (50).

При проведении расчетов, результаты которых представлены в первой главе на рис. 3—7, задача (54), (55)

решалась с помощью метода ломаных Эйлера. В результате получалась таблица значений функций $u(t)$, $w(t)$: $u_k = u(t_k)$, $w_k = w(t_k)$, $t_k = kh$, h — шаг по времени. Значения функций $x(t)$, $y(t)$ находились с помощью интегрирования этих таблиц по формуле трапеций, приводящей к рекуррентным соотношениям:

$$\begin{aligned} x_k &= x(t_k) = x_{k-1} + \frac{h}{2}(u_{k-1} + u_k), \\ y_k &= y(t_k) = y_{k-1} + \frac{h}{2}(w_{k-1} + w_k). \end{aligned} \tag{57}$$

Вычисления прекращались на шаге $k = n$, для которого $y_{n-1} > 0 \geqslant y_n$. Это условие означает, что в некоторый момент времени T, удовлетворяющий условию $t_{n-1} < T \leqslant t_n$, ядро падает на Землю, заканчивая свое движение.

§ 7. Как послать космический корабль к Луне

В заключение этой главы мы расскажем о задаче, относящейся к космической баллистике.

Предположим, что нужно рассчитать траекторию полета космического корабля к Луне. После того как ракета вывела его в космическое пространство и прекратила работу, он движется под действием гравитационного притяжения Земли, Луны и Солнца.

Введем в пространстве декартову систему координат и обозначим через (x, y, z) координаты корабля, через (x_n, y_n, z_n), $n = 1, 2, 3$,— координаты Земли, Луны и Солнца соответственно. При этом нужно иметь в виду, что все четыре тела движутся, т. е. их координаты являются функциями времени.

Будем считать, что движение Земли, Луны и Солнца нам известно, т. е., что $x_n(t)$, $y_n(t)$, $z_n(t)$ являются заданными функциями. Наша задача заключается в том, чтобы рассчитать движение космического корабля. Иными словами, мы должны найти функции $x(t)$, $y(t)$, $z(t)$, определяющие его положение в пространстве в любой момент времени.

Движение корабля подчиняется второму закону Ньютона:

$$\boldsymbol{a} = \frac{1}{m}\boldsymbol{F}, \tag{58}$$

где m — масса корабля, $\boldsymbol{a}$ — ускорение, $\boldsymbol{F}$ — равнодействующая сил притяжения корабля Землей, Луной

и Солнцем:

$$\boldsymbol{F} = \boldsymbol{F}_1 + \boldsymbol{F}_2 + \boldsymbol{F}_3.$$

Согласно закону всемирного тяготения эти силы определяются формулой

$$\boldsymbol{F}_n = \gamma \frac{m m_n}{R_n^3} \{(x_n - x)\,\boldsymbol{i} + (y_n - y)\,\boldsymbol{j} + (z_n - z)\,\boldsymbol{k}\}, \qquad (59)$$

$$n = 1, 2, 3.$$

Здесь γ — гравитационная постоянная, m_n — масса соответствующего небесного тела, R_n — расстояние между его центром и кораблем:

$$R_n = \sqrt{(x_n - x)^2 + (y_n - y)^2 + (z_n - z)^2},$$

$\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ — единичные векторы, направленные вдоль координатных осей и образующие базис в пространстве. Выражение (59) составлено таким образом, что вектор $\boldsymbol{F}_n$ направлен по прямой, которая соединяет корабль с центром небесного тела, и его величина обратно пропорциональна квадрату расстояния между ними:

$$|\boldsymbol{F}_n| = \gamma \frac{m m_n}{R_n^2}.$$

Обратите внимание на то, что масса корабля m в правой части уравнения (58) сокращается.

Векторное уравнение (58) можно записать в виде трех скалярных уравнений, спроектировав его на координатные оси:

$$x'' = X,\ y'' = Y,\ z'' = Z. \qquad (60)$$

Здесь вторые производные функций $x(t)$, $y(t)$, $z(t)$ по времени представляют собой проекции ускорения $\boldsymbol{a}$; X, Y, Z — проекции вектора $\frac{1}{m}\boldsymbol{F}$. Согласно (59)

$$\begin{aligned} X &= \gamma\left(m_1 \frac{x_1 - x}{R_1^3} + m_2 \frac{x_2 - x}{R_2^3} + m_3 \frac{x_3 - x}{R_3^3}\right), \\ Y &= \gamma\left(m_1 \frac{y_1 - y}{R_1^3} + m_2 \frac{y_2 - y}{R_2^3} + m_3 \frac{y_3 - y}{R_3^3}\right), \qquad (61) \\ Z &= \gamma\left(m_1 \frac{z_1 - z}{R_1^3} + m_2 \frac{z_2 - z}{R_2^3} + m_3 \frac{z_3 - z}{R_3^3}\right). \end{aligned}$$

Координаты центра Земли, Луны, Солнца являются известными функциями времени. Подставляя их в (61), мы запишем правые части уравнений (60) X, Y, Z как функции переменных x, y, z, t и получим систему трех дифференциальных уравнений второго порядка:

$$\begin{aligned} x'' &= X(x, y, z, t), \\ y'' &= Y(x, y, z, t), \\ z'' &= Z(x, y, z, t). \end{aligned} \tag{62}$$

В свою очередь ее можно свести к шести уравнениям первого порядка:

$$\begin{aligned} x' &= u, \\ y' &= v, \\ z' &= w, \\ u' &= X(x, y, z, t), \\ v' &= Y(x, y, z, t), \\ w' &= Z(x, y, z, t). \end{aligned} \tag{63}$$

Итак, система дифференциальных уравнений, которая описывает движение корабля в космическом пространстве после выключения двигателей, составлена. При современном уровне развития вычислительных средств найти ее решение при заданных начальных условиях

$$\begin{aligned} x(t_0) &= x_0,\ y(t_0) = y_0,\ z(t_0) = z_0, \\ u(t_0) &= u_0,\ v(t_0) = v_0,\ w(t_0) = w_0 \end{aligned} \tag{64}$$

не составляет труда. Таким образом, если известно, в какую точку околоземного пространства и с какой скоростью ракета выводит корабль, то с помощью интегрирования системы (63) можно определить его дальнейшее движение.

Однако особенность рассматриваемой задачи состоит в том, что никаких начальных условий в ней не задано. В ней поставлена конечная цель космического полета: корабль, выведенный на орбиту, должен после прекращения работы двигателей пролететь от Земли к Луне и попасть в определенную область окололунного пространства. Для достижения этой цели нам самим следует подобрать такие начальные условия (сформулировать «задание» для ракеты), которым соответствует нужная траектория.

В такой постановке задача является намного более сложной. По своему характеру это типичная задача опти-

мизации, в которой ограничения на допустимые решения задаются техническими возможностями ракеты. Поиск начальных условий (64), определяющих наилучшую траекторию, ведется с помощью методов, описанных в главе 5. Он требует многократного интегрирования системы (63) с различными начальными условиями (64), в результате чего мы постепенно приближаемся к искомому ответу.

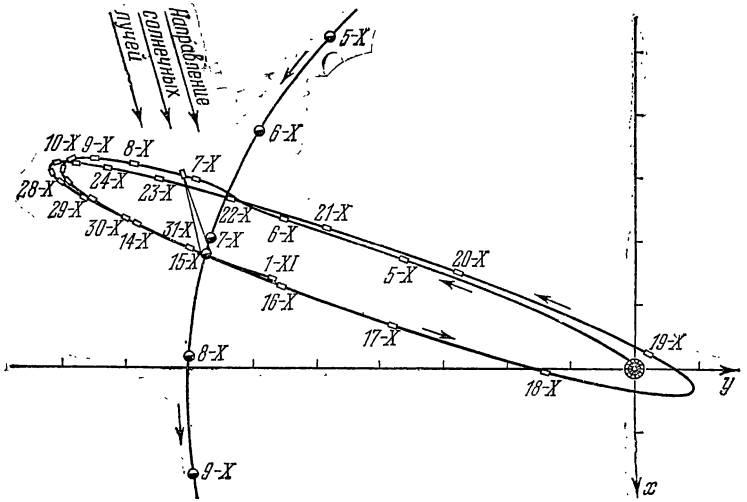

Рис. 61. Траектория советской автоматической межпланетной станции Луна-2, сделавшей первые снимки обратной стороны Луны (проекция на плоскость орбиты Луны).

Предположим теперь, что эта задача решена: найдены такие начальные условия, которые определяют нужную траекторию корабля. После этого необходимо еще рассмотреть вопрос о допусках. Совершенно ясно, что ракета не сможет вывести корабль на орбиту абсолютно точно в соответствии с нашими расчетами. Неизбежны какие-то отклонения в положении и скорости корабля в момент выключения двигателей от рассчитанных начальных условий. В результате вся дальнейшая траектория также будет отличаться от рассчитанной и может далеко уйти от намеченной цели. Поэтому, прежде чем осуществлять запуск, необходимо провести дополнительное исследование с целью определить допустимые погрешности в положении и скорости корабля в начале его траектории.

Мы рассказали, как рассчитывается траектория космического корабля в межпланетном пространстве после того, как он выведен на орбиту и двигатели прекратили работу. Не менее сложной является задача расчета активного участка траектории. При ее решении, кроме сил тяготения, нужно учитывать работу двигателей, сопротивление воздуха, вращение Земли и т. д.

На рис. 61 в качестве примера приведена траектория советской автоматической межпланетной станции Луна-2, которая впервые позволила заглянуть в неизвестное: сделать самые первые снимки обратной стороны Луны и передать их на Землю. Станция была запущена 4 октября 1959 года — через два года после запуска первого искусственного спутника Земли. На ней не было двигателей коррекции, поэтому требовалась очень высокая точность выведения на орбиту для того, чтобы осуществить этот уникальный космический эксперимент.

Дальнейший прогресс в космических исследованиях был основан на применении автоматических и пилотируемых аппаратов, снабженных собственными бортовыми двигателями. Благодаря этим двигателям космические аппараты, выведенные на орбиту ракетой-носителем, перестали быть пассивными «пленниками» гравитационных сил: появилась возможность целенаправленного изменения их траекторий.

Всякое управляемое изменение траектории принято называть в космической баллистике маневром. Примерами маневров могут служить переход с одной орбиты искусственного спутника Земли на другую, стыковка, возвращение аппарата с орбиты на Землю, коррекция траектории при полете к другим планетам, изменение траектории при подлете к какой-нибудь планете с целью перехода на орбиту ее искусственного спутника, мягкая посадка. Маневрирование — это сложная и ответственная операция, от успеха которой обычно зависит выполнение всей намеченной программы полета.

В настоящее время разработаны эффективные численные методы решения математических задач космической баллистики. Они позволяют определить оптимальный режим работы бортового двигателя, обеспечивающий выполнение запланированного маневра. Расчеты проводятся с учетом конкретных ограничений, обусловленных техническими характеристиками двигателя и системы управления. Условие оптимальности режима обычно означает

наименьший расход горючего, запас которого на борту космического корабля жестко ограничен.

Однако ЭВМ используются не только для предварительного расчета маневра, но и для его фактического осуществления. При выполнении маневра они управляют работой двигателя с учетом информации, которая автоматически поступает к ним от приборов, определяющих положение и скорость корабля. Такое активное управление с переработкой поступающей информации в режиме реального времени особенно важно при выполнении одного из самых сложных маневров — мягкой посадки на планету, лишенную атмосферы, например, на Луну. При отсутствии атмосферы не возникает аэродинамического торможения, нельзя использовать парашют. Уменьшение скорости аппарата к моменту его соприкосновения с поверхностью планеты до безопасного уровня осуществляется только за счет бортового двигателя, и требуется предельно точное управление его работой, чтобы запланированная мягкая посадка не превратилась в «жесткую».

Космические исследования базируются на самых новейших достижениях во многих областях человеческих знаний. В частности, они были бы абсолютно невозможны без современной вычислительной математики.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

Аддитивная система записи чисел 49
Алгоритм 31
— Евклида 32
Алгоритмические языки 67
Арифметическое устройство ЭВМ 53

Градиент 119

Дисплей 74

Задача внешней баллистики 19
— космической баллистики 199
— линейного программирования 141
— об использовании ресурсов 134
— о колебании маятника 179, 194
Запоминающее устройство ЭВМ 53

Интегральная сумма 158
Интегрируемая функция 158
Итерация 39

Критерий практики 16

Лемма о переходе к пределу в неравенствах 89
Леммы о суммах Дарбу 161, 162
Линии уровня 120
Лобовое сопротивление 17

Математическая модель 12
Машинное слово 53
Метод вилки 89
— градиентного спуска 125
— итераций 93
— касательных 99
— ломаных Эйлера 188
— наибыстрейшего спуска 126
— Ньютона 99
— покоординатного спуска 123
— последовательных приближений 93

Определенный интеграл 158

Позиционная система записи чисел 49

Рекуррентная формула 38

Симплекс-метод 145
Суммы Дарбу 160

Телетайп 73
Теорема Вейерштрасса 110
— об интегрируемости монотонных функций 163

Теорема о существовании корня у непрерывной функции 85
— о сходимости метода итераций 96
— — — — касательных 100
Транслятор 68
Транспортная задача 131

Управляющее устройство ЭВМ 53
Уравнение колебаний маятника 181
— радиоактивного распада 179
Условие Липшица 94
Устройства ввода-вывода ЭВМ 54, 69

Формула Ньютона — Лейбница 156
Формула прямоугольников 167

— Симпсона 171
— трапеций 171
— удвоения 42

Целевая функция 106
Центральный процессор ЭВМ 55

Частная производная 118

Андрей Николаевич Тихонов
Дмитрий Павлович Костомаров

РАССКАЗЫ О ПРИКЛАДНОЙ МАТЕМАТИКЕ

М., 1979 г., 208 стр. с илл.

Редакторы *И. В. Викторенкова*, *Е. Ю. Ходан*
Технический редактор *В. Н. Кондакова*
Корректоры *О. А. Бутусова*, *А. Л. Ипатова*

ИБ № 11461

Сдано в набор 26.02.79.
Подписано к печати 22.05.79. Т-08999.
Бумага 84×108¹/₃₂, тип. № 1.
Обыкновенная гарнитура. Высокая печать.
Усл. печ. л. 10,92. Уч.-изд. л. 10,5.
Тираж 150 000 экз. Заказ № 1606
Цена книги 35 коп.

Издательство «Наука»
Главная редакция
физико-математической литературы
117071, Москва, В-71, Ленинский проспект, 15

2-я тип. изд-ва «Наука»
121099, Москва, Г-99, Шубинский пер., 10