

ПРЕДИСЛОВИЕ

Концепция математического образования в высшей технической школе была сформулирована еще в начале XX века А. Н. Крыловым¹: "Инженер в своей практической деятельности бывает постоянно вынужден делать свои заключения, руководствуясь "здравым смыслом" или "глазомером и при этом в тех трудных случаях, когда расчет бессилен или когда надо устанавливать сами данные или допущения для расчета. Он изучает математику с целью практической, прикладной и рассматривает ее не как самостоятельный предмет изучения, а как подсобное орудие, как инструмент для решения ряда вопросов, встречаемых в некоторой ограниченной области практической деятельности. Здесь полная строгость рассуждений не может быть проводима целиком [...] Обоснование может быть дано не только чисто умозрительное, сводящее все к основным аксиомам, но и при помощи наглядности, делающее утверждение очевидным. Из этого однако не следует, чтобы прикладное изучение математики сводилось к рецептуре или к умению пользоваться справочниками, ибо тогда оно сводило бы математику к орудию счета по готовым образцам, и ее значение как орудия исследования утратилось бы. Но, понятно, прикладной характер должен оказывать существенное влияние на содержание и изложение курса."

Эта концепция, блестяще подтвержденная всей деятельностью ее автора, сохранила свое значение и сегодня. Однако за прошедшие годы изменилась как сама математика, так и ее роль в прикладных исследованиях. По-видимому, наиболее существенные для приложений изменения связаны с быстрым развитием средств вычислительной техники.

До эпохи "компьютерной революции" математика позволяла в каждой области знания эффективно анализировать сравнительно небольшой круг задач, решаемых в основном точными аналитическими методами. Поэтому исследователь был вынужден при построении математической модели упрощать свою реальную задачу до тех пор, пока она не станет анализируемой (даже за счет существенного уменьшения адекватности). С другой стороны, исследователь должен был владеть "математической технологией": уметь выполнять всякого рода математические преобразования, находить интегралы, решать частные виды дифференциальных

¹Алексей Николаевич КРЫЛОВ (1863-1945) – русский математик, механик и кораблестроитель, член Петербургской АН и АН СССР, основатель теории приближенных вычислений.

уравнений и т. п. На выработку этих знаний и умений и направлен существующий до сих пор традиционный курс математики для технических и естественнонаучных специальностей.

Но непрерывный рост быстродействия и объема памяти компьютера, разработка высокоэффективных прикладных программ и сред конечного пользователя расширяет круг доступных для анализа математических моделей, позволяет использовать иные методы их исследования. Поэтому математическая подготовка современного исследователя должна быть существенно усиlena. В то же время прикладное математическое обеспечение компьютера берет на себя все возрастающую часть "математической технологии и мы считаем целью практических занятий по математике углубленное знакомство с основными математическими понятиями, структурами и моделями, а не подготовку виртуозов аналитических преобразований, которых немало было в прошлые времена.

Предлагаемый курс математики преследует две основные цели:

- 1) подготовить студента к изучению общетехнических и специальных дисциплин, познакомив его с используемыми в этих дисциплинах математическими моделями;
- 2) научить студента решать и анализировать типовые (для избранного им направления подготовки) задачи, эффективно используя современное математическое обеспечение компьютера.

Соответственно, при изучении каждой темы

- на лекции студенту следует объяснить, "как устроена" изучаемая математическая модель, и, по возможности, установить связи с введенными ранее понятиями и моделями. При этом многие доказательства можно заменять *правдоподобными рассуждениями*, не выдавая их, конечно, за доказательства;
- на практических занятиях в аудитории студент должен научиться исследовать эту модель "вручную" в простейших (не содержащих технических трудностей) случаях;
- на лабораторных занятиях в дисплейном классе студент должен быть ознакомлен с современными программными средствами, позволяющими решать реальные задачи по изучаемой теме. При этом особое внимание должно быть уделено работе в средах конечного пользователя (MAPLE, MATLAB) и использованию библиотек стандартных программ на Фортране (таких как NAG, IMSL).

Серьезное предупреждение. Цивилизованный пользователь, на под-

готовку которого рассчитан этот курс, должен научиться пользоваться при решении математических задач библиотеками стандартных программ и не должен даже пытаться программировать вычислительные алгоритмы сам.

Создание "самопальных" программ по вдохновению пользователя или на основе многочисленных "руководств для чайников" может привести к последствиям, не менее печальным, чем изготовление самодельных взрывных устройств.

Это не означает, что "самопал" не сработает никогда. Однако можно уверенно утверждать, что правильный ответ будет получен лишь в такой задаче, которая может быть решена как угодно, даже "при помощи веревочной петли и палки". В случае задачи сколь-нибудь более сложной полученный "ответ" не будет иметь ничего общего с действительностью.

Курс состоит из трех разделов: "Математический анализ" "Линейная алгебра и ее приложения" "Дополнительные главы". Компоновка рассчитана на *одновременное* чтение лекций по первым двум разделам, составивших первый том двухтомника.

Первоначальный вариант нашего курса читался в течение ряда лет в Санкт-Петербургском государственном электротехническом университете и на химическом факультете Санкт-Петербургского государственного университета. Он был напечатан в виде отдельных брошюр в 1992-1994 годах. Переработанный курс вышел в трех томах в 1996-2000 годах. При подготовке настоящего издания некоторые параграфы курса подверглись существенной переработке с учетом замечаний, сделанных нашими коллегами. Были исправлены также замеченные опечатки.

В разные годы отдельные главы рукописи по нашей просьбе читали Я.И. Белопольская, Н.А. Бодунов, Ю.А. Ильин, М.В. Левит, А.С. Меркурьев, М.А. Нарбут, В.В. Некруткин, А.Н. Подкорытов, В.И. Полищук, С.И. Репин, В.М. Рябов, Г.С. Светлова, В.В. Скитович. Мы признательны нашим коллегам, которые способствовали уменьшению количества ошибок. Особенно благодарны мы профессору Санкт-Петербургского Политехнического университета **Владимиру Матвеевичу Чистякову**, трагически погившему летом 2006 года. Он внимательнейшим образом прочел весь курс и высказал более ста замечаний.

Мы заранее признательны всем читателям, которые пожелают прислать нам свои замечания².

²Проще всего это сделать по e-mail nazarov@lek.ru

Раздел 1

МАТЕМАТИЧЕСКИЙ АНАЛИЗ

Глава 1. НЕКОТОРЫЕ ОСНОВНЫЕ ПОНЯТИЯ

1.1. Высказывания. Логические операции

Предложение, содержащее истинное или ложное утверждение, мы будем называть *высказыванием*.

Примеры. 1. Тигр – млекопитающее.

2. "А" – первая буква русского алфавита.

3. В колоде для преферанса 52 карты.

4. $2+3=6$.

Первые два высказывания истинны, вторые два – ложны.

Следующие предложения высказываниями не являются:

Простите пехоте, что так неразумна бывает она.

Социалистическая перспектива.

Розовое платье красивее, чем голубое.

Первые два предложения не содержат утверждений; истинность утверждения, содержащегося в третьем предложении, зависит от вкуса.

Определение. Пусть **A** – высказывание. *Отрицанием* высказывания **A** называется высказывание "неверно, что **A**" (или "не **A**"), которое можно, если **A** истинно, и истинно, если **A** ложно. При записи вместо "не **A**" обычно употребляют знак $\neg A$.

Это определение коротко записывается в виде так называемой *таблицы истинности* (здесь **И** обозначает истину, **Л** – ложь):

A	$\neg A$
И	Л
Л	И

Определение. Пусть **A**, **B** – высказывания. *Конъюнкцией* этих высказываний называется высказывание "**A** и **B**" (пишут **A&B** или **A \wedge B**), которое задается следующей таблицей истинности:

A	B	A \wedge B
И	И	И
И	Л	Л
Л	И	Л
Л	Л	Л

Примеры. 1. Высказывание (В октябре 31 день) $\wedge (\pi > 3)$ истинно, ибо истинны оба операнда конъюнкции.

2. Высказывание (Стрелка компаса всегда указывает на запад) $\wedge (3 \cdot 3 = 9)$ ложно, так как ложен первый операнд конъюнкции.

Определение. Пусть A , B – высказывания. *Дизъюнкцией* этих высказываний называется высказывание "A или B" (пишут $A \vee B$), которое задается следующей таблицей истинности:

A	B	$A \vee B$
И	И	И
И	Л	И
Л	И	И
Л	Л	Л

Примеры. 1. Высказывание (В октябре 31 день) \vee (Стрелка компаса всегда указывает на запад) истинно, ибо первый операнд дизъюнкции истинный.

2. Высказывание (1900 год – високосный) $\vee (2 \cdot 2 = 5)$ ложно, так как ложны оба операнда дизъюнкции.

Замечания. 1. Обратите внимание на то, что союз "или" не имеет в математике разделительного смысла "или-или обычного для употребления этого союза в русском языке".

2. Для запоминания удобны следующие описания: конъюнкция истинна только при истинности обоих ее operandов; дизъюнкция ложна только при ложности обоих ее operandов.

Определение. Пусть A , B – высказывания. *Импликацией* называется высказывание "если A, то B" (пишут $A \Rightarrow B$), которое задается следующей таблицей истинности:

A	B	$A \Rightarrow B$
И	И	И
И	Л	Л
Л	И	И
Л	Л	И

Примеры. 1. Высказывание $(\pi > 3) \Rightarrow (2 \cdot 2 = 3)$ ложно.

2. Высказывание $(2 \cdot 2 = 5) \Rightarrow (2 \cdot 2 = 3)$ истинно.

Замечания. 1. Как видно из второго примера, значение союза "если...то" в математике не всегда совпадает с его значением в русском языке. Поэтому мы советуем не пытаться интерпретировать импликацию на содержательном уровне, а запомнить таблицу истинности и пользоваться ею.

2. В отличие от конъюнкции и дизъюнкции импликация не симметрична: $B \Rightarrow A$ не то же самое, что $A \Rightarrow B$.

Определение. Пусть A , B – высказывания. Эквиваленцией называется высказывание " A равносильно B " (пишут $A \Leftrightarrow B$), которое задается следующей таблицей истинности:

A	B	$A \Leftrightarrow B$
И	И	И
И	Л	Л
Л	И	Л
Л	Л	И

Примеры. 1. $(\neg(A \wedge B)) \Leftrightarrow ((\neg A) \vee (\neg B))$.

2. $(\neg(A \vee B)) \Leftrightarrow ((\neg A) \wedge (\neg B))$.

3. $((A \Rightarrow B) \wedge (B \Rightarrow A)) \Leftrightarrow (A \Leftrightarrow B)$.

Проверьте истинность этих высказываний, построив таблицы истинности.

1.2. Множества

Множество – понятие первичное, неопределяемое. Синонимами слова *множество* являются: *семейство*, *класс*, *стадо*, а также ряд других слов.

Множество состоит из элементов. Описание множества должно быть настолько подробным, чтобы практически о любом предмете (вещи) можно было сказать, является этот предмет (эта вещь) элементом данного множества или нет.

Мы будем обозначать множества большими буквами латинского алфавита, а их элементы – малыми. Если a – элемент множества A (говорят также " a принадлежит множеству A "), будем писать $a \in A$; если a не является элементом множества A (a не принадлежит множеству A), будем писать $a \notin A$.

Часто встречающиеся числовые множества имеют стандартные обозначения:

\mathbb{N} – множество натуральных чисел;

\mathbb{Z} – множество целых чисел;

\mathbb{R} – множество вещественных чисел.

Очевидна истинность следующих высказываний: $2 \in \mathbb{N}$, $3.62 \notin \mathbb{N}$, $3.62 \in \mathbb{R}$.

Если множество состоит из конечного (и не слишком большого) количества элементов, проще всего задать это множество, просто перечислив все его элементы. Например, запись

$$X = \{-1, 0, 1\}$$

означает, что множество состоит из трех элементов: (-1) , (0) , (1) . Отметим, что порядок перечисления элементов множества не играет роли. Так, например,

$$\{-1, 0, 1\} = \{1, -1, 0\}.$$

Если пишут $A = B$, то имеют в виду, что A и B – два имени одного и того же множества (иначе говоря, слева и справа от знака равенства стоят два экземпляра одного и того же множества).

Условимся считать все элементы множества различными (во множестве не может быть одинаковых элементов). Например, если положить в кошелек пять только что отпечатанных сторублевок, то элементами множества следует считать не сторублевки, а *пронумерованные* сторублевки, так как одна купюра отличается от другой только своим номером. Несколько рублевых монет можно рассматривать как множество только при условии, что они различны (например, по году выпуска).

Если множество конечно, но содержит очень много элементов, то целесообразно вместо перечисления этих элементов сформулировать характеристическое свойство, которым обладают все элементы этого множества и только элементы этого множества. Запись

$$A = \{a | \mathcal{P}(a)\}$$

читается так: " A состоит из всех тех и только тех элементов, которые обладают свойством \mathcal{P} (для каждого из которых высказывание $\mathcal{P}(a)$ истинно)". Например, " A – множество всех натуральных чисел,

квадрат которых меньше миллиарда" (попробуйте задать это множество перечислением его элементов!).

Очевидно, что задать множество с бесконечным количеством элементов можно только указанием характеристического свойства этих элементов. Так, множеством четных чисел называют множество целых чисел, делящихся на 2 без остатка, множеством правильных дробей называют множество обыкновенных дробей, у которых модуль числителя строго меньше модуля знаменателя, и т.д.

Замечания. 1. Множество удобно представлять себе как мешок, в который свалены составляющие это множество элементы. Мы намеренно использовали слово "свалены", чтобы еще раз подчеркнуть неупорядоченность элементов множества.

2. Не следует отождествлять близкие по смыслу слова русского языка "много" и "множество". Во множестве (математическом) может быть мало элементов, может быть один элемент (есть даже термин "одноэлементное множество") и, наконец, может вообще не быть ни одного элемента (*пустое множество*). Поэтому часто употребляемое "школьное" выражение "уравнение имеет множество решений" бессодержательно: у уравнения всегда есть множество решений (даже если у него нет ни одного решения).

3. Пустое множество принято обозначать символом \emptyset .

4. Если множество задается характеристическим свойством его элементов, то это свойство должно быть проверяемым. Вряд ли можно говорить о множестве студентов-спортсменов, так как существуют различные точки зрения на то, следует ли считать спортсменом студента, играющего на лекции в бридж³.

³Следует вообще избегать обсуждения несформулированных проблем. Так, в середине прошлого века у философов была модная тема дискуссий: "Может ли машина мыслить?". При этом не определялись ни понятие "машина" ни понятие "мыслить". Математик Тьюринг "проблему" закрыл, опубликовав статью под названием "Может ли машина мыслить?" в которой определил понятия "машина" и "мыслить" и доказал, что машина (в смысле Тьюринга) может мыслить (в смысле Тьюринга). Рекомендуем прочесть эту весьма любопытную работу (русский перевод: А. Тьюринг. Может ли машина мыслить, Физматгиз, М.: 1960).

Алан Матисон ТЬЮРИНГ (A.M. Turing, 1912-1954) – английский инженер и математик, член Лондонского королевского общества. Возглавлял работу по созданию вычислительных машин в Национальной физической лаборатории в Теддингтоне. Математические работы Тьюринга в основном посвящены математической логике и вычислительным машинам.

1.3. Части множества. Операции над множествами

По-видимому, нет необходимости объяснять, что такая часть множества, если само множество уже задано. Отметим только, что иногда вместо понятных слов "часть множества" употребляют менее понятный синоним "подмножество".

Например, множество четных чисел есть часть множества целых чисел (подмножество множества целых чисел). Если множество A является частью множества B , то пишут $A \subset B$. При этом не исключается, что $A = B$, т.е. всякое множество считается *по определению* своей частью. Очевидно, что

$$(A \subset B) \wedge (B \subset A) \iff A = B.$$

Мы будем считать пустое множество *по определению* частью любого множества. В частности, $\emptyset \subset \emptyset$. Это соглашение неизбежно, так как утверждение, что \emptyset не является частью X , означало бы, что во множестве \emptyset (пустом) есть элемент, не содержащийся во множестве X !

Рассмотрим некоторые операции над множествами.

Определение. *Пересечением* множеств A и B (обозначается $A \cap B$) называется множество, состоящее из всех элементов, входящих в A , и в B (т.е. их максимальная общая часть). Утверждение $A \cap B = \emptyset$ читается "множества A и B не пересекаются". Из определения очевидно, что для любых множеств A, B, C

$$A \cap B = B \cap A; \quad A \cap A = A; \quad A \cap \emptyset = \emptyset;$$

$$(A \cap B) \cap C = A \cap (B \cap C).$$

Определение. *Объединением* множеств A и B называется множество, состоящее из всех элементов, входящих хотя бы в одно из объединяемых множеств. Если объединяемые множества A и B представлять себе в виде заполненных их элементами мешков, то объединение получается так: берут пустой мешок C ,сыпают в него все содержимое мешков A и B , а затем удаляют дубликаты элементов (если такие найдутся). Пишут $C = A \cup B$. Из определения очевидно, что для любых множеств A, B, C

$$A \cup B = B \cup A; \quad A \cup A = A; \quad A \cup \emptyset = A;$$

$$(A \cup B) \cup C = A \cup (B \cup C).$$

Графическое изображение сказанного выше называется диаграммой Венна⁴ (рис.1.1). На этой диаграмме одно из исходных множеств заштриховано горизонтально, другое – вертикально. Пересечение (общая часть) несет на себе обе штриховки, объединение – хотя бы одну.

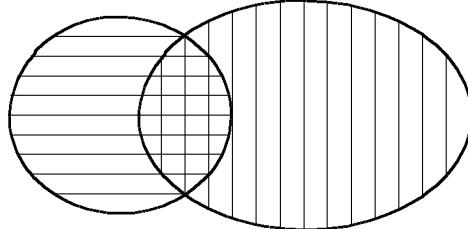


Рис.1.1

Несколько менее очевидны *дистрибутивные* свойства пересечения и объединения множеств: для любых множеств A, B, C

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C); \quad A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$$

Проверьте эти свойства на диаграмме Венна (рис.1.2):

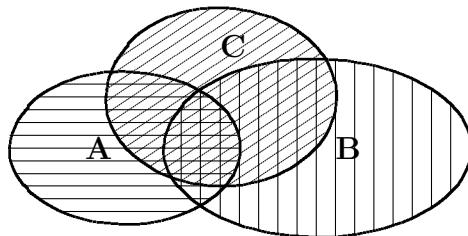


Рис.1.2

Определение. *Разностью* множеств A и B (обозначение $A \setminus B$) называется множество, состоящее из всех тех элементов множества A , которые не входят в множество B .

Для любых множеств A и B

$$(A \setminus B) \cup (A \cap B) = A.$$

В частности, если $A \cap B = \emptyset$, то $A \setminus B = A$.

⁴Джон ВЕНН (J. Venn, 1834-1923) – английский логик и математик.

Определение. *Прямым произведением* непустых множеств A и B (обозначение $A \times B$) называется множество, состоящее из всех *упорядоченных пар* (a, b) , где $a \in A$, $b \in B$.

Замечания. 1. Обратите внимание на то, что *первый* элемент упорядоченной пары (a, b) должен принадлежать *первому* сомножителю в декартовом произведении, а второй – второму. Отсюда очевидно, что $A \times B \neq B \times A$, если $A \neq B$.

2. Вместо $A \times A$ обычно пишут A^2 , и вообще, $A^n = \underbrace{A \times A \times \dots \times A}_{n \text{ сомножителей}}$

3. Прямое произведение множеств называют также их *декартовым⁵ произведением*.

1.4. Функция, отображение, оператор

Считая понятия *функция*, *отображение*, *оператор* первичными, опишем правила их использования.

Пусть X и Y – произвольные непустые множества. Если каждому элементу из X поставлен в соответствие *ровно один* элемент из Y , мы будем говорить:

- на множестве X задана функция со значениями в Y , или
- задано отображение X в Y , или
- задан оператор, действующий из X в Y .

Таким образом, слова *функция*, *отображение*, *оператор* мы будем считать синонимами⁶.

Если имя функции (отображения, оператора) есть f , то пишут

$$f : X \rightarrow Y.$$

Элемент $y \in Y$ (*единственный*), который функция f ставит в соответствие элементу $x \in X$, называют *значением функции* f в точке $x \in X$ или *образом* точки $x \in X$ при отображении f . Элемент $x \in X$ называют *прообразом* точки $y \in Y$ при отображении f . Пишут $y = f(x)$. Договоримся считать, что на наших рисунках прообраз соединяется с образом стрелкой, направленной от прообраза к образу (рис.1.3).

⁵Рене ДЕКАРТ (R. Descartes, 1596-1650) – французский философ, математик, физик и физиолог. Впервые ввел понятия переменной и функции, сформулировал основную теорему алгебры, создал метод координат.

⁶Иногда в математической литературе между этими словами проводят различие. Например, называют оператором только отображение множества в себя.

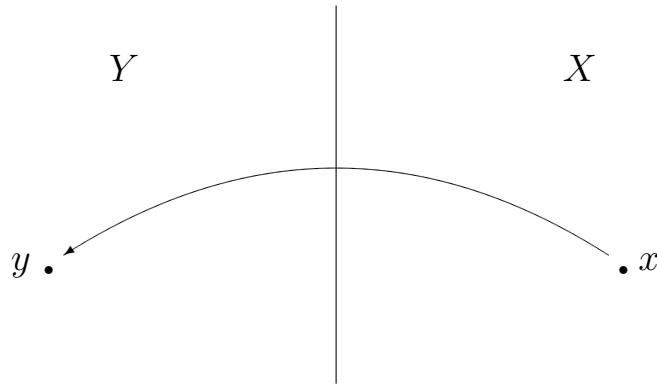


Рис.1.3

Серьезное предупреждение. Обращаем внимание читателя на то, что функция (отображение, оператор) – это тройка (X, Y, f) , т.е. всякий раз, когда употребляется одно из слов "функция" "отображение" "оператор" должны быть названы множество X (область определения функции), множество Y (множество, в котором лежат значения функции⁷) и правило f , позволяющее для каждого $x \in X$ найти соответствующий ему *единственный* $y \in Y$.

Например, $f : \mathbb{R} \rightarrow \mathbb{R}; \quad y = f(x) = x^2$.

Отметим часто допускаемую вольность: говорят о функции $\sin(x)$, хотя в точном понимании $\sin(x)$ не функция, а значение функции \sin в точке x . Мы настоятельно рекомендуем всегда проводить различие между функцией и ее значением в точке.

Терминологическое замечание. В физике употребляется понятие "величина" (то, что можно измерить или вычислить, например, масса, температура и т.д.). В математике аналогом этого понятия служит *переменная*⁸. Объявляя какую-нибудь букву переменной, мы одновременно должны указать некоторое непустое множество, элементы которого могут в дальнейшем замещать эту букву. Так, например, если n – целая числовая переменная ($n \in \mathbb{Z}$), то в выражении $n^2 + 4$ можно вместо n подставлять любое целое число. Если p и q – логические переменные, то в выражении $p \vee q$ можно вместо них подставлять либо **И** (истина), либо **Л** (ложь).

Если x – переменная с множеством значений X , y – переменная с множеством значений Y , $f : X \rightarrow Y$ – функция, то в выражении

⁷В этом контексте Y не есть множество значений функции f !

⁸Переменная в математике – имя существительное.

$y = f(x)$ переменную x часто называют *независимой*, а переменную y – *зависимой*. В нашем курсе эти термины не используются.

1.5. Композиция функций

Рассмотрим такую ситуацию (рис.1.4):

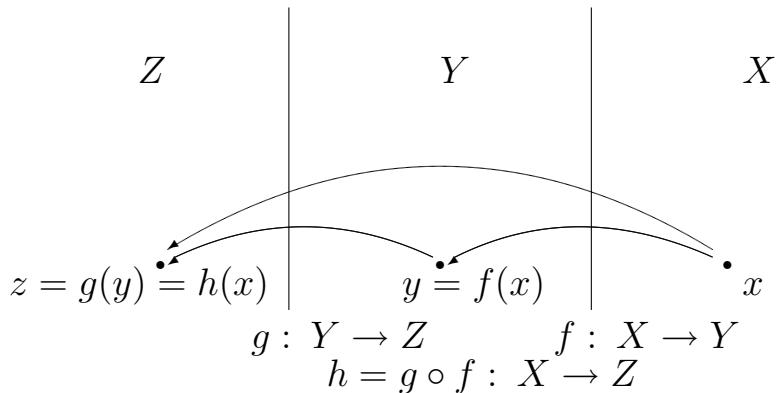


Рис.1.4

Заданы функции $f : X \rightarrow Y$ и $g : Y \rightarrow Z$. Поставим в соответствие каждому элементу $x \in X$ ровно один элемент $z \in Z$ по следующему правилу:

- 1) по заданному $x \in X$ найти соответствующий ему $y = f(x) \in Y$;
- 2) по полученному $y \in Y$ найти соответствующий ему $z = g(y) \in Z$.

Таким образом, оказывается построенной новая функция $h : X \rightarrow Z$. Ее называют *композицией* (иногда – *суперпозицией*) функций f и g . Пишут $h = g \circ f$ (обратите внимание на порядок компонент!). Говорят также "сложная функция но мы не советуем использовать этот термин.

Итак, $z = g(y) = g(f(x)) = (g \circ f)(x) = h(x)$.

Если изображать функцию в виде "черного ящика" со входом x , выходом y и именем f , то композиция функций представится последовательным соединением компонент (рис.1.5).

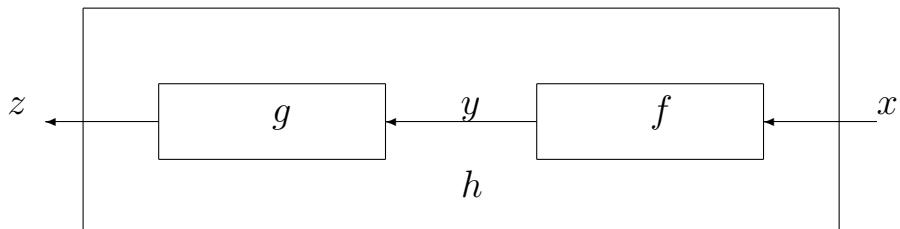


Рис.1.5

Глава 2. ЧИСЛОВЫЕ МНОЖЕСТВА

2.1. Множество всех вещественных чисел (\mathbb{R})

Мы не будем пытаться определить понятие "вещественное число считая, что некоторое представление читатель о нем уже имеет (в частности, знает, что каждое вещественное число изображается точкой на числовой оси). Как правило, мы будем иметь дело со следующими частями \mathbb{R} (здесь $a, b \in \mathbb{R}$, $a < b$):

- 1) $[a, b] = \{x \mid a \leq x \leq b\}$ – замкнутый промежуток, или сегмент;
- 2) $]a, b[= \{x \mid a < x < b\}$ – открытый промежуток, или интервал;
- 3) полуинтервал $]a, b] = \{x \mid a < x \leq b\}$;
- 4) полуинтервал $[a, b[= \{x \mid a \leq x < b\}$;

Серьезное предупреждение. Читателю, несомненно, приходилось и придется в дальнейшем пользоваться компьютером для вычислений. Мы считаем необходимым подчеркнуть, что любой компьютер работает не с множеством всех вещественных чисел (\mathbb{R}), а лишь с *конечной* (и не очень большой) его частью – так называемыми *машинными* числами. Учет этого факта поможет избежать многих досадных ошибок при решении вычислительных задач. Поэтому мы настоятельно рекомендуем ознакомиться с машинными числами хотя бы в объеме Приложения.

Введем теперь *расширенное множество вещественных чисел* ($\overline{\mathbb{R}}$), которое получается добавлением к \mathbb{R} двух элементов: $-\infty$ (*минус бесконечность*) и $+\infty$ (*плюс бесконечность*), т.е. $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$. Эти элементы не являются, конечно, вещественными числами.

Следующие правила обращения с элементами множества $\overline{\mathbb{R}}$ вводятся по определению (здесь x – произвольное вещественное число):

$$-\infty < x < +\infty; \quad x + (-\infty) = -\infty; \quad x + (+\infty) = +\infty;$$

$$(-\infty) + (-\infty) = -\infty; \quad (+\infty) + (+\infty) = +\infty;$$

$(+\infty) + (-\infty)$ **не определено!**

$$x > 0 \implies x \cdot (\pm\infty) = \pm\infty; \quad x < 0 \implies x \cdot (\pm\infty) = \mp\infty;$$

$0 \cdot (\pm\infty)$ **не определено!**

$$(-\infty) \cdot (-\infty) = (+\infty) \cdot (+\infty) = +\infty; \quad (-\infty) \cdot (+\infty) = -\infty;$$

$\frac{x}{\pm\infty} = 0; \quad \frac{x}{0}$ **не определено!**

Делить на нуль даже в $\overline{\mathbb{R}}$ нельзя!

Определение. Пусть X – непустая часть \mathbb{R} . Если существует такое вещественное число M , что для всякого $x \in X$ выполняется неравенство $x \leq M$ (в X нет числа, большего, чем M), то говорят, что множество X *ограничено сверху*, а число M называют *верхней границей* множества X . Обратите внимание на то, что неравенство *нестрогое*!

Если числа с описанным выше свойством не существует, т.е. какое бы мы ни взяли вещественное число, во множестве X найдется большее число, то говорят, что множество X *не ограничено сверху*.

Геометрически X представляется множеством точек вещественной оси. Если M – верхняя граница множества X , то *правее* точки M нет точек из X .

Ясно, что любое число, большее верхней границы множества, также будет верхней границей этого множества. Поскольку $x \in \mathbb{R} \implies x < +\infty$, будем считать $+\infty$ верхней границей любой непустой части \mathbb{R} . Если же у X нет верхних границ из \mathbb{R} , то $+\infty$ будет его *единственной* верхней границей в $\overline{\mathbb{R}}$.

Аналогично определяется ограниченность снизу непустой части \mathbb{R} и ее нижняя граница.

Отметим, что не во всяком непустом множестве вещественных чисел есть наибольшее число. Покажем, например, что в интервале нет наибольшего числа.

Действительно, возьмем любое число x из интервала $]a, b[$. Тогда

$$x \in]a, b[\implies \left(\frac{x+b}{2} \in]a, b[\right) \wedge \left(\frac{x+b}{2} > x \right).$$

В то же время можно показать, что⁹ среди верхних границ любого непустого множества $X \subset \overline{\mathbb{R}}$ есть наименьшая. Ее называют *точной верхней границей* или *верхней гранью* множества X и обозначают символом $\sup(X)$ (от латинского supremum). При этом если X ограничено сверху, то $\sup(X) \in \mathbb{R}$, иначе $\sup(X) = +\infty$.

Аналогично вводится понятие *точной нижней границы (нижней грани)* множества – наибольшей из его нижних границ. Нижнюю грань множества X обозначают символом $\inf(X)$ (от латинского infimum).

Очевидно, что если в множестве есть наибольшее (наименьшее) число, оно, и является верхней (нижней) гранью этого множества.

⁹Выражение "можно показать, что" здесь и далее означает: "Мы не можем или не хотим приводить доказательство. Интересующимся предоставляется возможность ознакомиться с ним по более полным курсам".

2.2. Множество всех комплексных чисел (\mathbb{C})

Известно, что каждая точка M , лежащая в плоскости, взаимно однозначно определяется своими декартовыми координатами: упорядоченной парой вещественных чисел (x, y) (рис.2.1)¹⁰:

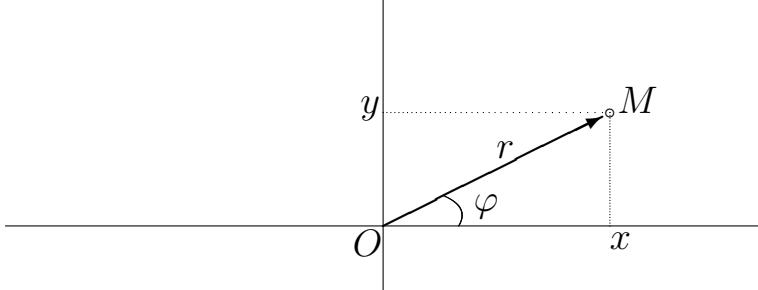


Рис.2.1

Условимся, что в этом пункте буква z (как с индексом, так и без) будет обозначать упорядоченную пару вещественных чисел. Известно также, что каждой точке M взаимно однозначно соответствует направленный отрезок \overrightarrow{OM} .

Назовем суммой двух упорядоченных пар вещественных чисел

$$z_1 = (x_1, y_1), \quad z_2 = (x_2, y_2)$$

упорядоченную пару вещественных чисел $(x_1 + x_2, y_1 + y_2)$:

$$z_1 + z_2 = (x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2).$$

При этом соответствующие направленные отрезки $\overrightarrow{OM_1}$ и $\overrightarrow{OM_2}$ складываются по известному правилу параллелограмма (рис.2.2).

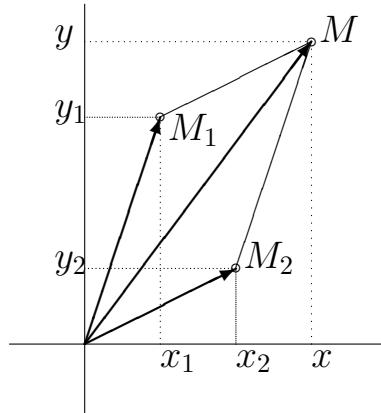


Рис.2.2

¹⁰Таким образом, плоскость можно рассматривать как геометрическую интерпретацию декартова произведения $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$.

Очевидно, что при любых z_1, z_2, z_3

$$z_1 + z_2 = z_2 + z_1; \quad z_1 + (z_2 + z_3) = (z_1 + z_2) + z_3$$

(сложение упорядоченных пар вещественных чисел коммутативно и ассоциативно).

Назовем произведением двух упорядоченных пар вещественных чисел

$$z_1 = (x_1, y_1), \quad z_2 = (x_2, y_2)$$

упорядоченную пару вещественных чисел $((x_1x_2 - y_1y_2), (x_1y_2 + x_2y_1))$:

$$z_1 \cdot z_2 = (x_1, y_1) \cdot (x_2, y_2) = ((x_1x_2 - y_1y_2), (x_1y_2 + x_2y_1)).$$

Убедитесь, что умножение коммутативно, ассоциативно и дистрибутивно относительно сложения.

Избавимся теперь от необходимости повторять длинное словосочетание "упорядоченная пара вещественных чисел" и заменим его более коротким "комплексное число".

Определение. Декартово произведение $\mathbb{R} \times \mathbb{R}$ с определенными в нем выше двумя операциями – сложением и умножением – будем называть *множеством всех комплексных чисел*, а его элементы – *упорядоченные пары вещественных чисел* – *комплексными числами*. Множество всех комплексных чисел обозначается символом \mathbb{C} .

Все среды конечного пользователя, предназначенные для решения вычислительных задач, "знают" комплексные числа и "умеют" работать с ними. Поэтому не следует стараться всегда "приводить выражения к вещественной форме".

Имеет место очевидное взаимно однозначное соответствие между точками плоскости и комплексными числами. Рассмотрим часть \mathbb{C} , соответствующую точкам, лежащим на оси абсцисс. Все эти комплексные числа имеют нулевую вторую компоненту пары. Покажем, что это множество *замкнуто относительно операций сложения и умножения*, т.е. что сумма и произведение чисел из этого множества находятся в нем же:

$$z_1 = (x_1, 0) \wedge z_2 = (x_2, 0) \implies z_1 + z_2 = (x_1 + x_2, 0) \wedge z_1 \cdot z_2 = (x_1x_2, 0).$$

Видно, что фактически мы оперируем лишь с первыми компонентами пар, соответственно складывая или перемножая их как вещественные числа. Это дает основание отождествить комплексное число $(x, 0) \in \mathbb{C}$ с вещественным числом $x \in \mathbb{R}$, и в дальнейшем не различать их.

При таком соглашении каждое комплексное число $z = (x, y)$ может быть представлено в виде

$$z = (x, y) = (x, 0) + (y, 0) \cdot (0, 1). \quad (2.2.1)$$

Если ввести обозначение $i = (0, 1)$, то (2.2.1) перепишется более компактно:

$$z = x + y \cdot i \quad \text{или} \quad z = x + i \cdot y. \quad (2.2.2)$$

Выражение (2.2.2) называют *алгебраической формой* записи комплексного числа. Точку, обозначающую умножение, часто опускают и пишут $z = x + iy = x + yi$. Говорят, что x – *вещественная часть* комплексного числа $z = x + iy$, а y – его *мнимая часть* (не стоит доискиваться исторических причин появления таких названий – проще запомнить их и пользоваться ими).

Пишут

$$x = Re(z) = Re(x + iy), \quad y = Im(z) = Im(x + iy).$$

(*Re* и *Im* – сокращения от *Real* – вещественный и *Imaginary* – мнимый).

Нетрудно заметить, что из "таблицы умножения" комплексных чисел существенно только одно правило

$$i^2 = i \cdot i = -1$$

(мы надеемся, что читатель убежден в отсутствии *вещественного* числа, квадрат которого отрицателен, и надеемся, что эта уверенность у него сохранится. Следовало бы писать $(0, 1)^2 = (-1, 0)$. Однако удобнее запись $i^2 = -1$, и так пишут все).

Сформулируем правила выполнения арифметических операций с комплексными числами, заданными в алгебраической форме.

Пусть $z_1 = x_1 + iy_1$, $z_2 = x_2 + iy_2$. Тогда

$$z_1 + z_2 = (x_1 + iy_1) + (x_2 + iy_2) = (x_1 + x_2) + i(y_1 + y_2);$$

$$z_1 \cdot z_2 = (x_1 + iy_1) \cdot (x_2 + iy_2) = (x_1x_2 - y_1y_2) + i(x_1y_2 + x_2y_1).$$

Видно, что комплексные числа складываются и перемножаются как двучлены. Нужно только помнить, что при появлении произведения $i \cdot i$ его следует сразу же заменить числом (-1) . Заметим еще, что умножение комплексного числа на *вещественное* выполняется покомпонентно:

$$\alpha \cdot (x + iy) = (\alpha x) + i(\alpha y),$$

что соответствует известному правилу умножения направленного отрезка \overrightarrow{OM} на число α .

Вычитание комплексных чисел определяется естественным образом:

$$z_1 - z_2 = (x_1 + iy_1) - (x_2 + iy_2) = (x_1 - x_2) + i(y_1 - y_2).$$

Определение. Числа $x + iy$ и $x - iy$ называются *сопряженными* комплексными числами. Пишут

$$x - iy = \overline{x + iy} \quad \text{или} \quad x + iy = \overline{x - iy}.$$

Точки плоскости, соответствующие паре сопряженных комплексных чисел, симметричны относительно оси абсцисс.

Комплексные числа с нулевой мнимой частью (по нашему соглашению – вещественные числа), и только они, совпадают с сопряженными им, т.е.

$$z = \bar{z} \iff z \in \mathbb{R} \subset \mathbb{C}.$$

Имеют место очевидные равенства:

$$Re(z) = \frac{1}{2} \cdot (z + \bar{z}); \quad Im(z) = \frac{1}{2i} \cdot (z - \bar{z}).$$

Заметим, что

$$z \cdot \bar{z} = (x + iy) \cdot (x - iy) = x^2 + y^2,$$

т.е. произведение пары комплексно сопряженных чисел равно квадрату расстояния от точек плоскости, соответствующих этим числам, до начала координат.

Определение. *Модуль* комплексного числа – это расстояние от начала координат до точки плоскости, соответствующей этому числу (длина направленного отрезка, соответствующего этому числу). Пишут

$$|z| = \sqrt{x^2 + y^2}.$$

Имеет место очевидное неравенство треугольника

$$|z_1 + z_2| \leq |z_1| + |z_2|$$

(длина стороны треугольника не больше суммы длин двух других его сторон).

2.3. Полярная система координат на плоскости и экспоненциальная форма записи комплексных чисел

Положение всякой точки M на плоскости (кроме начала координат) можно задать с помощью упорядоченной пары чисел: r – расстояние от начала координат до этой точки и φ – угол (в радианах), отсчитываемый против часовой стрелки от оси абсцисс до направленного отрезка \overrightarrow{OM} (рис.2.1).

Эту упорядоченную пару чисел называют *полярными координатами* точки M ; r называют *полярным радиусом*, а φ – *полярным углом* точки. Переход от полярных координат точки к ее декартовым координатам осуществляется по формулам

$$x = r \cdot \cos(\varphi); \quad y = r \cdot \sin(\varphi).$$

Полярный радиус точки определяется по ее декартовым координатам также однозначно:

$$r = \sqrt{x^2 + y^2} = |z|.$$

А вот полярный угол определен лишь с точностью до целого числа периодов синуса (косинуса). Действительно, на рис.2.1 вместо φ с одинаковым основанием можно написать

$$\varphi \pm 2\pi, \varphi \pm 4\pi, \dots, \varphi + 2k\pi, k \in \mathbb{Z}.$$

Иногда, чтобы устранить эту неоднозначность, усавливаются считать, что

$$0 \leq \varphi < 2\pi \quad \text{или} \quad -\pi < \varphi \leq \pi.$$

Тогда у каждой точки плоскости (кроме начала координат) полярный угол определяется однозначно¹¹. Для точки $(0, 0)$ (начало координат) полярный угол не определен.

¹¹Иногда встречающаяся формула $\varphi = \operatorname{arctg}(y/x)$ неверна, так как определяемый по ней угол всегда будет лежать в промежутке $[-\frac{\pi}{2}, \frac{\pi}{2}]$. В средах конечного пользователя имеются функции, возвращающие правильное значение полярного угла.

Мы в этом курсе не вводим ограничений на величину полярного угла.

Пусть комплексное число задано в алгебраической форме. Выражая декартовы координаты точки через полярные, получим

$$z = x + i \cdot y = r \cdot \cos(\varphi) + i \cdot r \cdot \sin(\varphi) = |z| \cdot (\cos(\varphi) + i \cdot \sin(\varphi)).$$

Выражение $|z| \cdot (\cos(\varphi) + i \cdot \sin(\varphi))$ называют *тригонометрической формой* записи комплексного числа. Полярный угол точки плоскости, соответствующей комплексному числу, называют также *аргументом* этого числа. Пишут $\arg(z) = \varphi$.

Отметим, что

$$\bar{z} = x - iy = |z| \cdot (\cos(\varphi) - i \cdot \sin(\varphi)),$$

откуда

$$|\bar{z}| = |z|, \quad \arg(\bar{z}) = -\arg(z).$$

Учитывая, что конструкция $\cos(\varphi) + i \cdot \sin(\varphi)$ будет постоянно встречаться, введем для нее специальное обозначение

$$\exp(i\varphi) = \cos(\varphi) + i \cdot \sin(\varphi)$$

и установим некоторые свойства функции $\exp : \mathbb{R} \rightarrow \mathbb{C}$ (иногда вместо $\exp(i\varphi)$ пишут $e^{i\varphi}$, но мы предпочитаем этого не делать):

1. $\exp(0) = 1$.
2. $\exp(i(\varphi + 2k\pi)) \equiv \exp(i\varphi)$ при $k \in \mathbb{Z}$.
3. $|\exp(i\varphi)| \equiv 1$.
4. $\exp(i\varphi) \cdot \exp(i\psi) \equiv \exp(i(\varphi + \psi))$.
5. $(\exp(i\varphi))^n \equiv \exp(in\varphi)$ при $n \in \mathbb{N}$.

Доказательство. 1-2. Очевидно из определения.

3. $|\exp(i\varphi)| = (\cos^2(\varphi) + \sin^2(\varphi))^{1/2} = 1$.
4. $\exp(i\varphi) \cdot \exp(i\psi) = (\cos(\varphi) + i \cdot \sin(\varphi)) \cdot (\cos(\psi) + i \cdot \sin(\psi))$
 $= (\cos(\varphi) \cdot \cos(\psi) - \sin(\varphi) \cdot \sin(\psi)) + i \cdot (\cos(\varphi) \cdot \sin(\psi) + \sin(\varphi) \cdot \cos(\psi))$
 $= \cos(\varphi + \psi) + i \cdot \sin(\varphi + \psi) = \exp(i(\varphi + \psi))$.
5. Полагая в 4 $\psi = \varphi$, получаем $(\exp(i\varphi))^2 = \exp(2i\varphi)$.
Далее, $(\exp(i\varphi))^3 = (\exp(i\varphi))^2 \cdot \exp(i\varphi) = \exp(2i\varphi) \cdot \exp(i\varphi) = \exp(3i\varphi)$, и т.д. ■

Отметим геометрическую интерпретацию свойства 2: точки плоскости, соответствующие значениям функции $\exp(i\varphi)$, $\varphi \in \mathbb{R}$, лежат на единичной окружности с центром в начале координат.

Используя введенное обозначение, будем записывать комплексное число также в виде

$$z = |z| \cdot \exp(i \cdot \arg(z))$$

и называть это выражение *экспоненциальной формой* записи комплексного числа.

При этом, если $z_1 = |z_1| \cdot \exp(i\varphi_1)$ и $z_2 = |z_2| \cdot \exp(i\varphi_2)$, то

$$z_1 \cdot z_2 = |z_1| \cdot |z_2| \cdot \exp(i(\varphi_1 + \varphi_2)).$$

При умножении ненулевых комплексных чисел их модули перемножаются, а аргументы складываются.

Назовем частным от деления z_1 на z_2 такое комплексное число z , что $z \cdot z_2 = z_1$.

Очевидно, что при $z_2 = 0$, $z \cdot z_2 = 0$ для любого $z \in \mathbb{C}$.

Как и во множестве вещественных чисел, деление на нуль в \mathbb{C} не определено.

Очевидно также, что если $z_2 \neq 0$ и $z_1 = 0$, то $z = \frac{z_1}{z_2} = 0$.

Пусть теперь $z_1 \neq 0$, $z_2 \neq 0$. Запишем делимое, делитель и частное в экспоненциальной форме:

$$z_1 = |z_1| \cdot \exp(i\varphi_1), \quad z_2 = |z_2| \cdot \exp(i\varphi_2), \quad z = |z| \cdot \exp(i\varphi).$$

Тогда по определению

$$|z| \cdot |z_2| \cdot \exp(i(\varphi + \varphi_2)) = |z_1| \cdot \exp(i\varphi_1).$$

Отсюда

$$|z| \cdot |z_2| = |z_1|, \quad \varphi + \varphi_2 = \varphi_1 + 2k\pi, \quad k \in \mathbb{Z};$$

$$|z| = \frac{|z_1|}{|z_2|}, \quad \varphi = \varphi_1 - \varphi_2 + 2k\pi, \quad k \in \mathbb{Z},$$

и частное определяется единственным образом:

$$z = \frac{z_1}{z_2} = \frac{|z_1|}{|z_2|} \cdot \exp(i(\varphi_1 - \varphi_2 + 2k\pi)) = \frac{|z_1|}{|z_2|} \cdot \exp(i(\varphi_1 - \varphi_2)). \quad (2.3.1)$$

Если делимое и делитель заданы в алгебраической форме, то частное находят так: умножая числитель и знаменатель дроби $\frac{z_1}{z_2}$ на \bar{z}_2 , имеем

$$\frac{z_1}{z_2} = \frac{x_1 + iy_1}{x_2 + iy_2} = \frac{(x_1 + iy_1) \cdot (x_2 - iy_2)}{(x_1 + iy_1) \cdot (x_2 - iy_2)} = \frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2} + i \cdot \frac{y_1x_2 - x_1y_2}{x_2^2 + y_2^2}.$$

Пример. $\frac{1+i}{2-3i} = \frac{(1+i) \cdot (2+3i)}{2^2 + 3^2} = -\frac{1}{13} + \frac{5}{13}i.$

Число $\frac{1}{z}$ обозначают также z^{-1} . Из формулы (2.3.1) получаем

$$z^{-1} = \frac{1}{|z|} \cdot \exp(-i\varphi) = \frac{\bar{z}}{|z|^2}.$$

В частности, $(\exp(i\varphi))^{-1} = \overline{\exp(i\varphi)} = \exp(-i\varphi)$.

2.4. Уравнение $z^n = c$, $n \in \mathbb{N}$. Квадратное уравнение

Пусть $c = |c| \cdot \exp(i\varphi) \neq 0$ – заданное комплексное число, n – натуральное число. Поскольку, очевидно, $z = 0$ – не корень, будем искать корни в экспоненциальной форме.

Пусть $z = |z| \cdot \exp(i\psi)$. Тогда

$$|z|^n \cdot \exp(in\psi) = |c| \cdot \exp(i\varphi).$$

Отсюда

$$|z|^n = |c|; \quad n\psi = \varphi + 2\pi k, \quad k \in \mathbb{Z}.$$

Поэтому все корни уравнения имеют один и тот же модуль $|z| = |c|^{1/n}$, а их аргументы вычисляются по формуле $\psi_k = \frac{\varphi + 2\pi k}{n}$, $k \in \mathbb{Z}$.

Легко видеть, что при $k = 0, 1, 2, \dots, n-1$ получаются n различных корней уравнения. Точки, изображающие эти корни, лежат на окружности с центром в начале координат и радиусом $|c|^{1/n}$ и делят эту окружность на n равных частей.

В то же время при $m \in \mathbb{Z}$ $\psi_{k+mn} = \psi_k + 2m\pi$. Поэтому всем значениям k , дающим одинаковый остаток при делении на n , соответствует один и тот же корень. Следовательно, других корней нет.

Замечание. При $c = 0$ корни уравнения сливаются: $z_k = 0$, $k = 0, 1, \dots, n - 1$. Мы будем считать, что уравнение $z^n = 0$ также имеет ровно n корней (каждый из них равен нулю).

Решим теперь квадратное уравнение

$$az^2 + bz + c = 0; \quad a, b, c \in \mathbb{C}, \quad a \neq 0.$$

Проделаем тождественное преобразование

$$\begin{aligned} az^2 + bz + c &= a \left(z^2 + 2z \frac{b}{2a} + \left(\frac{b}{2a} \right)^2 + \frac{c}{a} - \left(\frac{b}{2a} \right)^2 \right) = \\ &= a \left(\left(z + \frac{b}{2a} \right)^2 + \left(\frac{c}{a} - \left(\frac{b}{2a} \right)^2 \right) \right). \end{aligned}$$

Обозначив

$$W = z + \frac{b}{2a}, \quad G = \left(\frac{b}{2a} \right)^2 - \frac{c}{a},$$

получим уравнение $W^2 = G$, имеющее, как было показано выше, два корня (совпадающие при $G = 0$).

Мы показали, что всякое квадратное уравнение имеет ровно два комплексных корня. Рассмотрим особо случай, когда числа a, b, c – вещественные. Возможны три случая:

$$1. G = \left(\frac{b}{2a} \right)^2 - \frac{c}{a} > 0. \text{ Тогда } |G| = G, \quad \arg(G) = 0,$$

$$W_1 = |G|^{1/2} \exp(0) = |G|^{1/2}, \quad W_2 = |G|^{1/2} \exp(i \frac{2\pi}{2}) = -|G|^{1/2},$$

или

$$z_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \quad z_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

(уравнение имеет два *различных вещественных* корня).

$$2. G = \left(\frac{b}{2a} \right)^2 - \frac{c}{a} = 0. \text{ Тогда}$$

$$W_1 = W_2 = 0, \quad z_1 = z_2 = -\frac{b}{2a}$$

(уравнение имеет два *совпадающих вещественных* корня).

Случай 1 и 2 были изучены в школе.

3. $G = \left(\frac{b}{2a}\right)^2 - \frac{c}{a} < 0$. Тогда $|G| = -G$, $\arg(G) = \pi$,

$$W_1 = |G|^{1/2} \exp\left(i\frac{\pi}{2}\right) = i|G|^{1/2}, \quad W_2 = |G|^{1/2} \exp\left(i\frac{3\pi}{2}\right) = -i|G|^{1/2},$$

или

$$z_1 = \frac{-b + i\sqrt{4ac - b^2}}{2a}, \quad z_2 = \frac{-b - i\sqrt{4ac - b^2}}{2a}$$

(уравнение имеет два различных комплексно сопряженных корня).

Замечание. Имеют место так называемые формулы Виета¹² (проверяются вычислением):

$$z_1 + z_2 = -\frac{b}{a}, \quad z_1 \cdot z_2 = \frac{c}{a}.$$

¹²Франсуа ВИЕТ (F. Viète, 1540-1603) – французский математик.

Глава 3. ПОЛИНОМЫ (МНОГОЧЛЕНЫ)

3.1. Определение и стандартное представление

Определение. Полиномом *степени* n ($n = 0, 1, \dots$) называется функция $f : \mathbb{C} \rightarrow \mathbb{C}$, действующая по правилу

$$f(z) = a_0 + a_1 z + \dots + a_n z^n, \quad (3.1.1)$$

где a_0, a_1, \dots, a_n – заданные числа (коэффициенты полинома), и $a_n \neq 0$. Коэффициент a_n называется *старшим* коэффициентом полинома, а a_0 – *младшим* коэффициентом, или *свободным членом*.

К полиномам относят также функцию, равную нулю во всех точках \mathbb{C} . Степень этого полинома не определена.

Примеры.

- | | |
|--------------------|--|
| $f(z) \equiv 0$ | (нуль-полином, степень не определена); |
| $f(z) \equiv 5$ | (полином нулевой степени); |
| $f(z) = 1 - 5z$ | (полином первой степени); |
| $f(z) = 2z - 5z^7$ | (полином седьмой степени). |

Отметим, что полиномы и их отношения (так называемые *рациональные дроби*, рассматриваемые в гл. 4) – это единственные функции, вычисляемые компьютером непосредственно, ибо компьютер выполняет только сложение, вычитание, умножение и деление чисел. Поэтому полиномы и рациональные дроби являются сегодня истинными "элементарными функциями и следует освоить технику работы с ними.

В зависимости от задачи один и тот же полином целесообразно записывать в разных формах. Форму (3.1.1) мы будем называть *стандартным представлением полинома*.

Рассмотрим еще одно представление: покажем, что можно разложить полином не по степеням переменной z (как в стандартном представлении), а по степеням двучлена $(z - p)$:

$$f(z) = b_0 + b_1(z - p) + \dots + a_n(z - p)^n, \quad (3.1.2)$$

где p – произвольное число.

Здесь b_k , $k = 0, \dots, n - 1$ – некоторые числа (зависящие, конечно, от p). В частности, $b_0 = f(p)$.

Для доказательства (3.1.2) заменим в стандартном представлении полинома z на $w + p$:

$$f(w + p) = a_0 + a_1(w + p) + \dots + a_n(w + p)^n.$$

Раскрыв скобки и приведя подобные члены, получим

$$f(w + p) = b_0 + b_1w + \dots + b_{n-1}w^{n-1} + a_nw^n,$$

Положив $w = 0$, получим $b_0 = f(p)$. Заменив w на $z - p$, придем к (3.1.2).

3.2. Схема Горнера

При стандартном представлении полинома его значения следует вычислять по так называемой *схеме Горнера*¹³

$$\begin{aligned} f(z) &= a_0 + a_1z + a_2z^2 + \dots + a_nz^n = \\ &= \underbrace{(\dots \dots)}_{n-1}(a_nz + a_{n-1}) \cdot z + a_{n-2} + \dots + a_1 \cdot z + a_0. \end{aligned}$$

Такой способ вычисления имеет два преимущества.

Первое – очевидное: минимизируется количество арифметических операций. Действительно, для вычисления значения полинома степени n по схеме Горнера требуется n сложений и n умножений. При вычислении же по формуле (3.1.1) потребуется n сложений и $\frac{n(n+1)}{2}$ умножений (проверьте!).

Второе – совсем не очевидное: при вычислении по схеме Горнера существенно уменьшается вычислительная погрешность. Чтобы убедиться в этом, вычислите, используя микрокалькулятор, $f(199)$, если

$$f(z) = 2z^6 - 396z^5 - 396z^4 - 396z^3 - 396z^2 - 396z - 197,$$

двумя способами: по формуле (3.1.1) и по схеме Горнера.

3.3. Корни полинома. Разложение полинома на множители первой степени

Определение. Число $c \in \mathbb{C}$ называется *корнем* полинома f , если $f(c) = 0$.

Примеры. 1. Полином нулевой степени не имеет корней:

$$f(z) = a_0 \neq 0.$$

2. Полином первой степени имеет один корень:

¹³Вильям Джордж ГОРНЕР (W.J. Horner, 1786-1837) – английский математик.

$$f(z) = a_0 + a_1 z = 0 \implies z_1 = -\frac{a_0}{a_1}.$$

Отметим, что полином первой степени представим в виде

$$f(z) = a_0 + a_1 z = a_1 \cdot (z - z_1).$$

3. Полином второй степени имеет два корня (они могут и совпадать):

$$f(z) = a_0 + a_1 z + a_2 z^2 = 0 \implies z_{1,2} = -\frac{a_1}{2a_2} \pm \sqrt{\left(\frac{a_1}{2a_2}\right)^2 - \frac{a_0}{a_2}}$$

(здесь символ \sqrt{A} обозначает любое из решений уравнения $z^2 = A$).

Используя формулы Виета, легко проверить, что полином второй степени представим в виде

$$f(z) = a_0 + a_1 z + a_2 z^2 = a_2 \cdot (z - z_1) \cdot (z - z_2).$$

Эта формула верна и при $z_1 = z_2$.

Можно показать, что всякий полином степени n имеет ровно n корней и может быть представлен в виде

$$f(z) = a_0 + a_1 z + a_2 z^2 + \dots + a_n z^n = a_n \cdot (z - z_1) \cdot \dots \cdot (z - z_n), \quad (3.3.1)$$

где z_1, \dots, z_n – корни полинома (не обязательно попарно различные). Это утверждение называют *основной теоремой алгебры*.

Естественно объединить одинаковые сомножители в (3.3.1):

$$f(z) = a_n \cdot (z - z_1)^{k_1} \cdots (z - z_m)^{k_m}. \quad (3.3.2)$$

Здесь z_1, \dots, z_m – *попарно различные* корни полинома, а натуральные числа k_1, \dots, k_m называются *кратностями* соответствующих корней. Очевидно, что $k_1 + \dots + k_m = n$. Корень кратности 1 обычно называют *простым*.

Формулу (3.3.2) называют *разложением полинома на множители первой степени* или *мультипликативным представлением полинома*.

Если выполнить умножение в правой части (3.3.2), получатся формулы, связывающие коэффициенты полинома с его корнями (так же, как в случае квадратного уравнения, они называются формулами Виета). Приведем две из них:

$$\sum_{j=1}^n z_j = -\frac{a_{n-1}}{a_n}; \quad \prod_{j=1}^n z_j = (-1)^n \frac{a_0}{a_n}$$

(здесь каждый корень считается столько раз, какова его кратность).

Пример. Если $f(z) = 3 \cdot (z - 2) \cdot (z - 5)^3 \cdot (z - i)^2$, то $z_1 = 2$ – простой корень, $z_2 = -5$ – корень кратности 3, $z_3 = i$ – корень кратности 2.

Серьезное предупреждение. Задача отыскания корней полинома численно неустойчива, т.е. малые изменения коэффициентов полинома могут вызывать большие изменения его корней. Приведем простой пример.

Пусть требуется найти корни полинома $f(z) = (z - 1)^n$, и в свободном члене имеется абсолютная погрешность ε , т.е. фактически вместо уравнения $(z - 1)^n = 0$ решается уравнение $(z - 1)^n = \varepsilon$. Все его корни расположены на окружности с центром в точке $(1, 0)$ и радиусом $\varepsilon^{\frac{1}{n}}$.

Абсолютная погрешность найденных корней (она же и относительная) равна $\varepsilon^{\frac{1}{n}}$. Следовательно, погрешность определения корня превышает относительную погрешность свободного члена в $\varepsilon^{\frac{1}{n}-1}$ раз. Например, допустив погрешность в свободном члене полинома шестой степени всего в седьмой значащей цифре ($\varepsilon = 10^{-6}$), получим "коэффициент усиления" погрешности 10^5 !

3.4. Деление полинома на полином

Из курса линейной алгебры известно, что пространством полиномов порядка n называется множество, состоящее из всех полиномов, степень которых строго меньше, чем n , и нуль-полинома.

Пусть P – полином степени n , Q – полином степени m ($m \leq n$). Можно показать, что существуют такие полиномы S (степени $n - m$) и R (порядка m), что

$$P = Q \cdot S + R, \quad (3.4.1)$$

причем полиномы S и R определяются единственным образом. По аналогии с делением чисел полином P называют делимым, полином Q – делителем, полином S – частным и полином R – остатком.

Сформулированное утверждение мы проиллюстрируем примером, из которого будет ясен способ доказательства. Итак, пусть

$$\begin{aligned} P(z) &= p_5 z^5 + p_4 z^4 + p_3 z^3 + p_2 z^2 + p_1 z + p_0 \quad (p_5 \neq 0) \\ Q(z) &= q_2 z^2 + q_1 z + q_0 \quad (q_2 \neq 0) \end{aligned}$$

Полином S должен иметь степень 3, а полином R – порядок 2. Запишем эти полиномы в виде (здесь $r_0, r_1, s_0, s_1, s_2, s_3$ подлежат определению)

$$S(z) = s_3 z^3 + s_2 z^2 + s_1 z + s_0 \quad (s_3 \neq 0), \quad R(z) = r_1 z + r_0.$$

Подставив P, Q, S, R в (3.4.1), получим

$$\begin{aligned} p_5z^5 + p_4z^4 + p_3z^3 + p_2z^2 + p_1z + p_0 = \\ (q_2z^2 + q_1z + q_0) \cdot (s_3z^3 + s_2z^2 + s_1z + s_0) + r_1z + r_0. \end{aligned}$$

Раскрыв скобки и приравняв коэффициенты при одинаковых степенях переменной в обеих частях уравнения, получим линейную систему относительно искомых коэффициентов.

$$\begin{array}{l|lcl} z^5 & q_2s_3 & & = p_5 \\ z^4 & q_1s_3 + q_2s_2 & & = p_4 \\ z^3 & q_0s_3 + q_1s_2 + q_2s_1 & & = p_3 \\ z^2 & q_0s_2 + q_1s_1 + q_2s_0 & & = p_2 \\ z^1 & q_0s_1 + q_1s_0 + 1 \cdot r_1 & & = p_1 \\ z^0 & q_0s_0 + 0 \cdot r_1 + 1 \cdot r_0 & = p_0 \end{array}.$$

Определитель матрицы коэффициентов этой системы (нижней треугольной) отличен от нуля (он равен q_2^4), следовательно, система имеет единственное решение. Кроме того, из первого уравнения получаем $s_3 = \frac{p_5}{q_2} \neq 0$, т.е., действительно, частное – полином *третьей степени*. В то же время остаток – полином *второго порядка* (в том числе может оказаться и нуль-полиномом – тогда говорят, что P делится на Q без остатка).

3.5. Непрерывность полинома

Пусть $f(z) = a_0 + a_1z + \dots + a_nz^n$ – произвольный полином степени n . Зафиксируем на комплексной плоскости точку p , а точку z сделаем переменной. Положим в формуле (3.1.2) $z = p + h$:

$$f(p+h) - f(p) = f(p) + b_1 \cdot h + \dots + b_n \cdot h^n - f(p) = h \cdot (b_1 + \dots + b_n \cdot h^{n-1}).$$

Оценим модуль этой разности, считая, что точка z лежит внутри круга с центром в точке p и радиусом ρ , т.е. $|h| = |z - p| < \rho$ (рис.3.1).

$$|b_1 + \dots + b_n \cdot h^{n-1}| \leq |b_1| + \dots + |b_n| \cdot \rho^{n-1}.$$

Отсюда

$$|f(z) - f(p)| \leq M \cdot |h| = M \cdot |z - p|,$$

где $M = |b_1| + \dots + |b_n| \cdot \rho^{n-1}$ – положительное число.

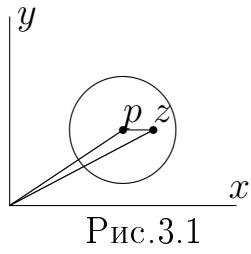


Рис.3.1

Из последнего неравенства следует, что значение полинома в точке z можно сделать как угодно близким к его значению в точке p за счет приближения z к p . Точнее,

Для любого положительного числа ε можно указать такое положительное число δ , что

$$|z - p| < \delta \implies |f(z) - f(p)| < \varepsilon. \quad (3.5.1)$$

Очевидно, что достаточно взять $\delta = \min\{\rho, \frac{\varepsilon}{M}\}$.

Определение. Функция, обладающая свойством (3.5.1), называется непрерывной в точке p .

Поскольку в нашем случае комплексное число p произвольно, мы доказали, что полином непрерывен в любой точке своей области определения (комплексной плоскости).

Глава 4. РАЦИОНАЛЬНЫЕ ДРОБИ (ДРОБНО-РАЦИОНАЛЬНЫЕ ФУНКЦИИ)

4.1. Определение и важное соглашение

Определение. Дробно-рациональной функцией (рациональной дробью) называется отношение полиномов

$$f(z) = \frac{a_m z^m + a_{m-1} z^{m-1} + \dots + a_1 z + a_0}{b_n z^n + b_{n-1} z^{n-1} + \dots + b_1 z + b_0}, \quad (a_m \neq 0, \quad b_n \neq 0, \quad n > 0).$$

Важное соглашение

Может оказаться, что полином-числитель и полином-знаменатель имеют общий корень – число c . Тогда оба полинома делятся без остатка на двучлен $(z - c)$, и рациональную дробь можно сократить, т.е. разделить и числитель и знаменатель на этот двучлен.

Договоримся не рассматривать рациональные дроби, которые можно сократить, т.е. будем всегда считать, что числитель и знаменатель не имеют общих корней.

Если $m < n$ (степень числителя строго меньше степени знаменателя), рациональная дробь называется правильной, если $m \geq n$ – неправильной. Неправильная рациональная дробь может быть (см. п.3.4) единственным образом представлена в виде суммы полинома и правильной рациональной дроби. Действительно, из (3.4.1) следует, что

$$\frac{P}{Q} = S + \frac{R}{Q},$$

и дробь $\frac{R}{Q}$ – правильная (случай, когда числитель делится на знаменатель без остатка, интереса, очевидно, не представляет). Поэтому в дальнейшем мы рассматриваем, в основном, правильные рациональные дроби.

Итак, мы рассматриваем *правильные и несократимые* рациональные дроби. Корни знаменателя такой дроби называются ее *полюсами*. Кратность корня знаменателя называется кратностью полюса.

Рациональная дробь определена на всей комплексной плоскости за исключением полюсов.

4.2. Разложение правильной рациональной дроби на простейшие

Определение. Рациональная дробь $\frac{A}{(z - c)^k}$ (A, c – комплексные числа, k – натуральное число) называется *простейшей дробью*.

Теорема. Всякая правильная рациональная дробь может быть представлена в виде суммы простейших дробей (*разложена на простейшие дроби*).

Доказательство. В силу **Важного соглашения** (п.4.1) числитель и знаменатель дроби не имеют общих корней. Предположим, что наша дробь имеет полюс с кратности k , т.е. она имеет вид

$$f(z) = \frac{P(z)}{(z - c)^k Q(z)}, \quad P(c) \neq 0, \quad Q(c) \neq 0.$$

Покажем, что существует *единственный* полином R порядка k и *единственный* полином S , такие что

$$f(z) = \frac{P(z)}{(z - c)^k Q(z)} = \frac{R(z)}{(z - c)^k} + \frac{S(z)}{Q(z)}.$$

Рассмотрим разность

$$\frac{P(z)}{(z - c)^k Q(z)} - \frac{R(z)}{(z - c)^k} = \frac{P(z) - R(z) \cdot Q(z)}{(z - c)^k Q(z)}, \quad (4.2.1)$$

где R – произвольный полином порядка k . Попробуем построить этот полином так, чтобы разность $P - R \cdot Q$ делилась без остатка на $(z - c)^k$. Для этого расположим полиномы P, Q, R по возрастающим степеням двучлена $(z - c)$ (см. (3.1.2)):

$$\begin{aligned} P(z) &= p_0 + p_1(z - c) + \dots + p_{k-1}(z - c)^{k-1} + \dots, \\ Q(z) &= q_0 + q_1(z - c) + \dots + q_{k-1}(z - c)^{k-1} + \dots, \\ R(z) &= r_0 + r_1(z - c) + \dots + r_{k-1}(z - c)^{k-1} \end{aligned}$$

(многоточия в полиномах P и Q означают, что эти полиномы могут содержать и более высокие степени двучлена $(z - c)$). Заметим, что $q_0 = Q(c) \neq 0$.

Вычислим теперь коэффициенты при степенях двучлена в числителе дроби (4.2.1) вплоть до $(k - 1)$ -го и приравняем их нулю:

$$\begin{array}{l|lll} (z-c)^0 & p_0 - r_0 q_0 & = & 0 \\ (z-c)^1 & p_1 - r_0 q_1 - r_1 q_0 & = & 0 \\ (z-c)^2 & p_2 - r_0 q_2 - r_1 q_1 - r_2 q_0 & = & 0 \\ \dots & \dots & & \dots \\ (z-c)^{k-1} & p_{k-1} - r_0 q_{k-1} - r_1 q_{k-2} - r_2 q_{k-3} - \dots - r_{k-1} q_0 & = & 0 \end{array}.$$

Полученная для определения коэффициентов полинома R система уравнений имеет вид:

$$\begin{bmatrix} q_0 & 0 & 0 & \dots & 0 \\ q_1 & q_0 & 0 & \dots & 0 \\ q_2 & q_1 & q_0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ q_{k-1} & q_{k-2} & q_{k-3} & \dots & q_0 \end{bmatrix} \times \begin{bmatrix} r_0 \\ r_1 \\ r_2 \\ \dots \\ r_{k-1} \end{bmatrix} = \begin{bmatrix} p_0 \\ p_1 \\ p_2 \\ \dots \\ p_{k-1} \end{bmatrix}.$$

Определитель матрицы коэффициентов этой системы равен $q_0^k \neq 0$. Поэтому числа r_0, \dots, r_{k-1} (а потому и полином R) определяются однозначно. Теперь полином $P(z) - R(z) \cdot Q(z)$ делится на $(z - c)^k$ без остатка (по построению). Обозначая полином-частное S , получим

$$P(z) - R(z) \cdot Q(z) = (z - c)^k \cdot S(z),$$

откуда

$$f(z) = \frac{P(z)}{(z - c)^k Q(z)} = \frac{R(z)}{(z - c)^k} + \frac{S(z)}{Q(z)}.$$

Выделяя таким образом последовательно все полюсы, получим представление правильной несократимой рациональной дроби в виде

$$f(z) = \frac{P(z)}{(z - c_1)^{k_1} \dots (z - c_m)^{k_m}} = \frac{R_1(z)}{(z - c_1)^{k_1}} + \dots + \frac{R_m(z)}{(z - c_m)^{k_m}}. \quad (4.2.2)$$

Поскольку полиномы R_j ($j = 1, \dots, m$) уже расположены по возрастающим степеням $(z - c_j)$, получим далее

$$\begin{aligned} \frac{R_j(z)}{(z - c_j)^{k_j}} &= \frac{r_{0,j} + r_{1,j}(z - c_j) + \dots + r_{k_j-1,j}(z - c_j)^{k_j-1}}{(z - c_j)^{k_j}} = \\ &= \frac{r_{0,j}}{(z - c_j)^{k_j}} + \frac{r_{1,j}}{(z - c_j)^{k_j-1}} + \dots + \frac{r_{k_j-1,j}}{(z - c_j)}, \end{aligned}$$

что завершает доказательство теоремы. ■

Можно показать, что разложение правильной рациональной дроби на простейшие единственно. Числа, стоящие в чисителях простейших дробей, могут быть найдены так называемым методом неопределенных коэффициентов, который мы продемонстрируем на примере.

Пример. Разложить на простейшие дроби

$$\frac{z^2 + 1}{(z - 1)^3(z^2 + 3)}.$$

Выделим полюсы и представим эту дробь в виде (4.2.2) с неизвестными пока коэффициентами:

$$\begin{aligned} \frac{z^2 + 1}{(z - 1)^3(z + i\sqrt{3})(z - i\sqrt{3})} &= \\ &= \frac{r_0 + r_1(z - 1) + r_2(z - 1)^2}{(z - 1)^3} + \frac{s}{z + i\sqrt{3}} + \frac{v}{z - i\sqrt{3}}. \end{aligned}$$

Приведем правую часть к общему знаменателю и приравняем в чисителях дробей, стоящих слева и справа от знака равенства, коэффициенты при одинаковых степенях переменной z :

$$\begin{array}{l|lll} z^4 & r_2 + s + v & = & 0 \\ z^3 & r_1 - 2r_2 - (3 + \sqrt{3}i)s - (3 - \sqrt{3}i)v & = & 0 \\ z^2 & r_0 - r_1 + 4r_2 + (3 + \sqrt{3}i)s + (3 - \sqrt{3}i)v & = & 1 \\ z^1 & 3r_1 - 6r_2 - (1 + 3\sqrt{3}i)s - (1 - 3\sqrt{3}i)v & = & 0 \\ z^0 & 3r_0 - 3r_1 + 3r_2 + \sqrt{3}is - \sqrt{3}iv & = & 1 \end{array}.$$

Применяя алгоритм полного исключения, получим

$$r_0 = 1/2, \quad r_1 = 1/4, \quad r_2 = 0, \quad s = -i\sqrt{3}/24, \quad v = i\sqrt{3}/24.$$

Серьезное предупреждение. С "докомпьютерных" времен у человека сохранилось естественное отвращение к работе с комплексными числами. Поэтому при разложении вещественных рациональных дробей на простейшие иногда используют в случае сопряженных комплексных полюсов так называемые "простейшие дроби второго типа":

$$\frac{Ax + B}{x^2 + px + q},$$

где A, B, p, q – вещественные числа, а корни квадратного трехчлена в знаменателе – комплексные.

Это существенно усложняет задачу и не дает никакого выигрыша, поскольку (в отличие от человека) компьютер легко справляется с "комплексной арифметикой". Мы настоятельно рекомендуем не пользоваться простейшими дробями второго типа.

4.3. Непрерывность рациональной дроби

Рассмотрим простейшую дробь $f(z) = \frac{1}{z^n}$ с полюсом в начале координат. Зафиксируем точку $p \neq 0$ и оценим разность $f(z) - f(p)$, считая, что переменная точка z берется внутри круга с центром в точке p и радиусом $|p|/2$ (рис.4.1). Тогда

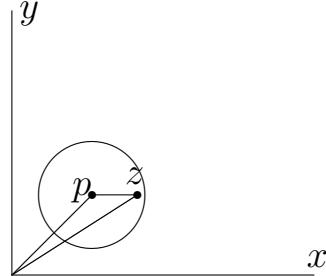


Рис.4.1

$$f(z) - f(p) = \frac{p^n - z^n}{p^n z^n} = (p - z) \cdot \frac{p^{n-1} + p^{n-2}z + \dots + pz^{n-2} + z^{n-1}}{p^n z^n}.$$

$$|f(z) - f(p)| \leq \frac{|p - z|}{|p|^n |z|^n} \cdot (|p|^{n-1} + |p|^{n-2}|z| + \dots + |p||z|^{n-2} + |z|^{n-1}). \quad (4.3.1)$$

Так как $\frac{|p|}{2} < |z| < \frac{3|p|}{2}$, усилим неравенство (4.3.1), заменив в правой его части: в знаменателе $|z|$ – на меньшее число $|p|/2$, а в скобке $|z|$ – на большее число $3|p|/2$.

$$\begin{aligned} |f(z) - f(p)| &\leq |p - z| \cdot \frac{\left(|p|^{n-1} + \frac{3}{2}|p|^{n-1} + \dots + \left(\frac{3}{2}\right)^{n-1}|p|^{n-1}\right)}{\frac{|p|^n}{2^n}|p|^n} = \\ &= |z - p| \cdot \frac{2(3^n - 2^n)}{|p|^{n+1}} = M \cdot |z - p|. \end{aligned}$$

Здесь $M = \frac{2(3^n - 2^n)}{|p|^{n+1}}$ – положительное число.

Из полученного неравенства следует, что значение нашей простейшей дроби в точке z можно сделать как угодно близким к

ее значению в точке p за счет приближения z к p . Точнее: для любого положительного числа ε можно указать такое положительное число $\delta = \min\{|p|/2, \varepsilon/M\}$, что

$$|z - p| < \delta \implies |f(z) - f(p)| < \varepsilon.$$

Поскольку точка p произвольна, мы доказали непрерывность нашей простейшей дроби в любой точке ее области определения.

Небольшое усложнение проведенного выше рассуждения позволяет установить и непрерывность простейшей дроби $\frac{1}{(z - c)^n}$ в любой точке ее области определения.

Докажем теперь важное вспомогательное утверждение. Пусть функции f_1 и f_2 определены в некоторой окрестности точки p , и существует такое положительное число M , что в этой окрестности выполняются неравенства

$$|f_1(z) - f_1(p)| \leq M|z - p|, \quad |f_2(z) - f_2(p)| \leq M|z - p|.$$

Тогда для линейной комбинации этих функций $f = \alpha_1 f_1 + \alpha_2 f_2$ имеем

$$\begin{aligned} |f(z) - f(p)| &= \left| (\alpha_1 f_1 + \alpha_2 f_2)(z) - (\alpha_1 f_1 + \alpha_2 f_2)(p) \right| \leq \\ &\leq |\alpha_1| |f_1(z) - f_1(p)| + |\alpha_2| |f_2(z) - f_2(p)| \leq \\ &\leq (|\alpha_1| + |\alpha_2|)M \cdot |z - p|. \end{aligned}$$

Это утверждение очевидным образом распространяется на линейную комбинацию любого конечного числа функций. Учитывая, что любая рациональная дробь представима в виде суммы полинома и линейной комбинации конечного числа простейших дробей, получаем

Следствие. Любая рациональная дробь непрерывна в любой точке области ее определения.

4.4. Поведение рациональной дроби в окрестности полюса

Ограничимся рассмотрением случая простейшей дроби

$$f(z) = \frac{1}{(z - c)^n}.$$

Перейдем к полярным координатам с началом в полюсе

$$z = c + r \cdot \exp(i\varphi) \implies f(z) = \frac{\exp(-i\varphi)}{r^n} \implies |f(z)| = \frac{1}{r^n}.$$

Очевидно, что модуль $f(z)$ можно сделать как угодно большим за счет приближения точки z к полюсу. Точнее: для любого положительного числа E можно указать такое положительное число $\delta = E^{-1/n}$, что

$$0 < |z - c| < \delta \implies |f(z)| > E.$$

Полученный результат обычно выражают словами: "простейшая дробь *не ограничена* в окрестности своего полюса".

Можно показать, что всякая рациональная дробь не ограничена в окрестности каждого своего полюса.

4.5. Почему не следует работать с сократимыми рациональными дробями

Начнем с примера. Пусть

$$f(x) = x \quad (x \in \mathbb{R}); \quad g(x) = \frac{x^2}{x} \quad (x \in \mathbb{R}, \quad x \neq 0) \quad (\text{рис.4.2}).$$

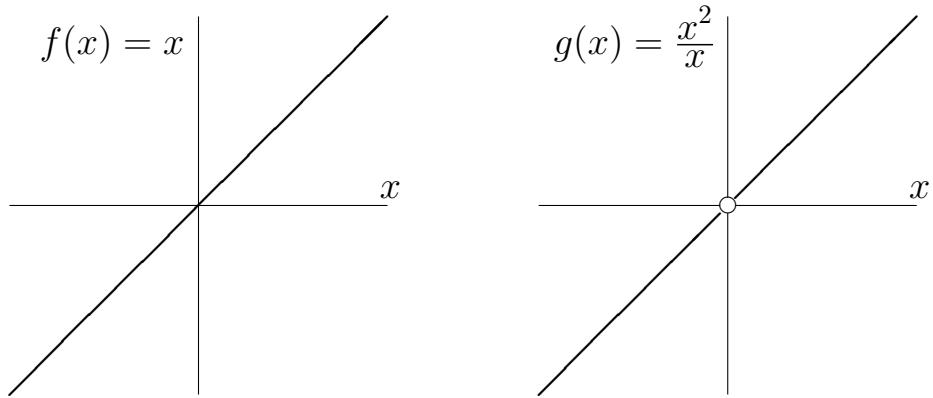


Рис.4.2

Эти функции совпадают всюду, кроме одной точки ($x = 0$), где g не определена (график функции g получен удалением из графика f начала координат). Если доопределить g , положив $g(0) = 0$, то она совпадет с f уже всюду и, в частности, станет *непрерывной* в точке $x = 0$. В то же время очевидно, что формально f получается из g *сокращением* последней на x .

Далее возможны два пути:

1. Отказаться от рассмотрения сократимых рациональных дробей, т.е., в частности, не различать функции

$$f(x) = x \quad (x \in \mathbb{R}) \quad \text{и} \quad g(x) = \frac{x^2}{x} \quad (x \in \mathbb{R}, \quad x \neq 0),$$

что равносильно *доопределению* функции g в нуле по непрерывности.

2. Считать f и g разными функциями, а число ноль называть *пределом функции* g в точке $x = 0$. Пишут:

$$\lim_{x \rightarrow 0} g(x) = 0 \quad (\text{или} \quad \lim_{x \rightarrow 0} g(x) = 0).$$

С нашей точки зрения первый путь проще, и последовательное его проведение существенно облегчает *пользователю* изучение математики. В то же время нам не известны случаи возникновения на этом пути каких-либо затруднений и, тем более, противоречий.

Поэтому мы не будем работать с сократимыми рациональными дробями (а позже – и с другими функциями, имеющими точки *устранимого разрыва*).

Мы будем такие разрывы устранять!

При этом предел функции в точке либо будет совпадать с ее значением (и поэтому не будет представлять интереса), либо не будет существовать. Так, например, в полюсе, т.е. в точке, являющейся корнем знаменателя *несократимой* рациональной дроби, никаким доопределением этой дроби сделать ее непрерывной нельзя – рациональная дробь в полюсе не имеет предела.

Такая точка зрения согласуется со взглядами известных математиков XX века Н. Н. Лузина¹⁴ и Л. Берса¹⁵.

Рассматривая рациональную дробь

$$\frac{x^m}{x^n} \cdot \frac{\phi(x)}{\psi(x)}, \quad \text{где } \psi(0) \neq 0 \quad \text{и} \quad m \geq n,$$

Н.Н.Лузин пишет: "Естественно рассматривать этот случай как случай кажущегося разрыва, обязанного не недостаткам самой функции (в геометрическом смысле), а лишь некоторому недостатку дающей эту функцию формулы, утрачивающей свой числовой смысл при } x = 0. Чем

¹⁴Николай Николаевич ЛУЗИН (1883-1950) – действительный член АН СССР и ряда зарубежных академий, основатель московской школы теории функций. Цитируется учебник В. Грэнвиль и Н. Лузин, Элементы дифференциального и интегрального исчисления, "ГИЗ М.-Л.: 1931.

¹⁵Липман БЕРС (L. Bers, 1914-1993) – профессор Колумбийского университета (США), президент Американского математического общества, глава отделения математики Национальной Академии наук США. Цитируется учебник Л. Берс, Математический анализ, "Высшая школа М.: 1975.

подобного рода утрата числового смысла той или иной формулой может произойти чисто случайно, читатель заметит из того обстоятельства, что достаточно написать любую непрерывную функцию $f(x)$ в виде $\frac{x \cdot f(x)}{x}$ или в виде $1/x + f(x) - 1/x$, как уже новая формула утрачивает числовой смысл в точке $x = 0$ ".

Л. Берс: "Работая с рациональными функциями, мы часто будем предполагать, что каждый линейный множитель, входящий одновременно в числитель и в знаменатель, уже сокращен, и что область определения функции соответствующим образом расширена".

Против этой точки зрения категорически возражают лишь преподаватели-репетиторы, ибо она отнимает у них возможность "обучать" всевозможным хитроумным приемам "вычисления" пределов.

Глава 5. ЧИСЛОВАЯ ПОСЛЕДОВАТЕЛЬНОСТЬ И ЕЕ ПРЕДЕЛ

5.1. Основные понятия. Примеры

Определение. Числовой последовательностью называют числовую функцию, заданную на множестве натуральных чисел ($f : \mathbb{N} \rightarrow \mathbb{C}$). Числовая последовательность каждому натуральному числу – номеру – ставит в соответствие одно комплексное число.

Замечания. 1. Часто областью определения числовой последовательности считают множество натуральных чисел, пополненное нулем (начинают считать не с единицы, а с нуля).

2. По традиции значения функции-последовательности часто обозначают не $f(n)$, а f_n , саму же последовательность – (f_n) или $(f_n)_1^{+\infty}$.

3. Удобно представлять себе последовательность в виде бесконечной вправо таблицы, в первой строке которой – номера, а во второй – соответствующие им значения последовательности:

n	1	2	3	...
f_n	f_1	f_2	f_3	...

Примеры. 1. $f(n) \equiv 1$. Множество значений последовательности состоит из одного числа. Такую последовательность принято называть *последовательность-константа*.

Серьезное предупреждение. Не следует путать функцию-константу (в частности, последовательность-константу) и число – единственное значение этой функции.

$$2. f(n) = \exp\left(i\frac{\pi}{2}n\right); \quad n = 0, 1, 2, \dots$$

Множество значений этой последовательности состоит из четырех чисел: $\{1, i, -1, -i\}$. Последовательность *периодическая*, и ее период равен четырем:

$$f(n + 4 \cdot m) = f(n) \quad \text{для всех } n, m = 0, 1, 2, \dots$$

$$3. f_n = n + i \cdot n^2; \quad |f_n| = \sqrt{n^2 + n^4} = n^2 \cdot \sqrt{1 + \frac{1}{n^2}}.$$

Видно, что при увеличении номера модуль значения последовательности неограниченно растет.

$$4. f_n = \left(1 + \frac{4}{n}\right) + i\left(2 + \frac{3}{n^2}\right).$$

При больших номерах значения этой последовательности мало отличаются от числа $1 + 2i$:

$$|f_n - (1 + 2i)| = \left| \frac{4}{n} + i \frac{3}{n^2} \right| = \sqrt{\frac{16}{n^2} + \frac{9}{n^4}} = \frac{1}{n} \sqrt{16 + \frac{9}{n^2}} \leq \frac{5}{n}.$$

Точнее: для любого положительного числа ε можно указать номер¹⁶ $n_0 = \text{entier}\left(\frac{5}{\varepsilon}\right) + 1$, начиная с которого выполняется неравенство $|f_n - (1 + 2i)| < \varepsilon$.

Определение. Будем называть r -окрестностью точки $a \in \mathbb{C}$ внутренность круга с центром в этой точке и радиусом r , т.е. множество $\{x \in \mathbb{C} \mid |x - a| < r\}$.

Используя это понятие, можно описать поведение последовательности из примера 4 так:

Каково бы ни было положительное число ε , найдется номер n_0 , начиная с которого все значения последовательности лежат в ε -окрестности точки $1 + 2i$.

Иначе, вне любой окрестности точки $1 + 2i$ лежат значения последовательности лишь для конечного числа номеров¹⁷.

И, наконец, говорят, что число $1 + 2i$ есть предел последовательности $(1 + \frac{4}{n}) + i(2 + \frac{3}{n^2})$, и пишут

$$\lim \left(\left(1 + \frac{4}{n} \right) + i \left(2 + \frac{3}{n^2} \right) \right) = 1 + 2i.$$

Определение. Число $A \in \mathbb{C}$ называется пределом последовательности (a_n) , если

для любого положительного числа ε найдется такой номер n_0 , что

$$n \geq n_0 \implies |a_n - A| < \varepsilon;$$

или

вне любой окрестности точки A лежат значения последовательности (a_n) лишь для конечного числа номеров;

или

внутри любой окрестности точки A лежат значения последовательности (a_n) почти для всех номеров.

¹⁶entier (фр.) – целый; $\text{entier}(x)$ – целая часть вещественного числа x , т.е. наибольшее целое число, не превосходящее x .

¹⁷Если некоторое свойство не выполняется лишь для конечного числа номеров, мы будем говорить, что оно выполняется почти для всех номеров.

Если последовательность (a_n) имеет предел A , то пишут

$$\lim a_n = A \quad \text{или} \quad \lim_{n \rightarrow +\infty} a_n = A.$$

Говорят также, что последовательность (a_n) *сходится* к числу A , и пишут

$$a_n \rightarrow A.$$

Нетрудно заметить, что пределом последовательности-константы является значение этой константы (пример 1).

Определение. Если все значения последовательности лежат внутри некоторого круга, то последовательность называется *ограниченной*.

Сравнивая это определение с определением предела, устанавливаем, что всякая сходящаяся (имеющая предел) последовательность ограничена.

Обратное, очевидно, не верно. Так, периодическая последовательность (пример 2) ограничена, но предела не имеет.

Определение. Если вне *любого* круга найдутся значения последовательности, то эту последовательность называют *неограниченной*. Неограниченная последовательность предела не имеет. Такова, в частности, последовательность из примера 3.

5.2. Свойства пределов

Следующие утверждения носят название *теорем о пределах*:

$$1. \quad f_n \equiv a \implies \lim f_n = a.$$

Это свойство было установлено в предыдущем пункте.

$$2. \quad (\lim f_n = a) \wedge (\lim g_n = b) \implies \lim(f + g)_n = a + b.$$

Доказательство. Из очевидного неравенства

$$|(f + g)_n - (a + b)| = |(f_n - a) + (g_n - b)| \leq |f_n - a| + |g_n - b|$$

следует, что при попадании f_n в $\varepsilon/2$ -окрестность точки a , а g_n – в $\varepsilon/2$ -окрестность точки b , $(f + g)_n$ попадает в ε -окрестность точки $a + b$. Но по условию вне $\varepsilon/2$ -окрестности точки a лежат значения последовательности (f_n) лишь для конечного числа номеров, и вне $\varepsilon/2$ -окрестности точки b лежат значения последовательности (g_n) также

лишь для конечного числа номеров. Следовательно, вне ε -окрестности точки $a + b$ лежат значения последовательности $((f + g)_n)$ лишь для конечного числа номеров. В силу произвольности положительного числа ε $\lim(f + g)_n = a + b$. ■

$$3. \quad (\lim f_n = a) \bigwedge (\lim g_n = b) \implies \lim(f \cdot g)_n = a \cdot b.$$

$$4. \quad (\lim f_n = a) \bigwedge (\lim g_n = b) \bigwedge (b \neq 0) \implies \lim(f/g)_n = a/b.$$

Доказательства утверждений **3** и **4** мы не приводим. Отметим лишь, что концентрация значений последовательности (f_n) около точки (числа) a , а значений последовательности (g_n) около точки (числа) b приводит, очевидно, к концентрации значений последовательности $((f \cdot g)_n)$ около точки (числа) $a \cdot b$, а значений последовательности $((f/g)_n)$ – около точки (числа) a/b . В последнем случае особо оговорено, что $b \neq 0$.

$$5. \quad (\lim f_n = 0) \bigwedge ((g_n) \text{ ограничена}) \implies \lim(f \cdot g)_n = 0.$$

Как и в утверждении **3**, здесь рассматривается предел произведения двух последовательностей. Но здесь не требуется существование предела второго сомножителя!

Доказательство. Пусть все значения ограниченной последовательности (g_n) лежат в M -окрестности начала координат, т.е. $|g_n| < M$. Тогда

$$|f_n \cdot g_n| < M \cdot |f_n|.$$

Возьмем произвольное положительное число ε . Вне ε/M -окрестности нуля лежат значения (f_n) лишь для конечного числа номеров. Поэтому вне $M \cdot \varepsilon/M = \varepsilon$ -окрестности нуля лежат значения $((f \cdot g)_n)$ для тех же (конечного числа) номеров. ■

$$6. \quad \lim f_n = a + ib \quad (a, b \in \mathbb{R}) \iff (\lim \operatorname{Re}(f_n) = a) \bigwedge (\lim \operatorname{Im}(f_n) = b).$$

Доказательство. Обозначим $x_n = \operatorname{Re}(f_n)$, $y_n = \operatorname{Im}(f_n)$. Если $\lim x_n = a$ и $\lim y_n = b$, то согласно утверждениям **1 – 3**,

$$\lim(x_n + iy_n) = \lim x_n + \lim(i \cdot y_n) = a + \lim(i) \cdot \lim y_n = a + ib.$$

Наоборот, пусть $\lim f_n = a + ib$. Из очевидного равенства

$$|f_n - (a + ib)| = |\overline{f_n - (a + ib)}| = |\overline{f_n} - (a - ib)|$$

следует, что $\lim \overline{f_n} = a - ib$. Теперь утверждения **1 – 3** дают:

$$\begin{aligned}\lim x_n &= \lim \left(\frac{1}{2}(f_n + \overline{f_n}) \right) = \lim \left(\frac{1}{2} \right) \cdot (\lim f_n + \lim \overline{f_n}) = \\ &= \frac{1}{2} \cdot (a + ib + a - ib) = a,\end{aligned}$$

$$\lim y_n = \lim \frac{1}{2i}(f_n - \overline{f_n}) = \frac{1}{2i} \cdot (a + ib - a + ib) = b. \quad \blacksquare$$

Утверждения **1 – 6** верны для любых комплексных последовательностей. При изучении *вещественных* последовательностей важную роль играет следующая

Теорема. Если последовательность *вещественных* чисел (x_n) не убывает и ограничена сверху, то она имеет предел.

Доказательство. Множество X значений последовательности (x_n) ограничено сверху и, следовательно (см. п.2.1), имеет верхнюю грань. Обозначим $x = \sup(X)$ и покажем, что $\lim x_n = x$.

Действительно, $x_n \leq x$ при всех $n \in \mathbb{N}$ (из определения верхней границы множества). Далее, для любого $\varepsilon > 0$ существует такое $n_0 \in \mathbb{N}$, что $x_{n_0} > x - \varepsilon$ (иначе число $x - \varepsilon$ было бы верхней границей множества X , что противоречит определению $\sup(X)$). Но последовательность (x_n) не убывает, и потому для всех $n \geq n_0$

$$x - \varepsilon < x_n \leq x,$$

откуда $|x_n - x| < \varepsilon$. ■

Аналогично, если последовательность вещественных чисел (x_n) не возрастает и ограничена снизу, то $\lim x_n = \inf(X)$.

Глава 6. ЧИСЛОВЫЕ РЯДЫ

6.1. Определение. Примеры

Рассмотрим две последовательности:

- 1) $f_n = 2n + 3i/n;$
- 2) $g_1 = 1; g_2 = 1; \text{ при } n > 2 \quad g_n = g_{n-1} + g_{n-2}.$

В первом случае значение последовательности с любым номером можно вычислить, не вычисляя ее значения с предшествующими номерами. Во втором – для вычисления, например, g_6 , придется найти g_3, g_4, g_5 : $g_3 = g_2 + g_1 = 2, g_4 = g_3 + g_2 = 3, g_5 = g_4 + g_3 = 5$ и, наконец, $g_6 = g_5 + g_4 = 8$.

Будем говорить, что последовательность (f_n) задана *явно*, а последовательность (g_n) – *неявно*. Иногда (к сожалению, редко) удается превратить неявное задание последовательности в явное. Напомним два примера из школьного курса:

Геометрическая прогрессия: $u_0 = a \neq 0$; при $n > 0 \quad u_n = q \cdot u_{n-1}$ ($q \neq 0$). Известно, что $u_n = a \cdot q^n$ для всех $n = 0, 1, 2, \dots$

Арифметическая прогрессия: $v_0 = a$; при $n > 0 \quad v_n = q + v_{n-1}$ ($q \neq 0$). Известно, что $v_n = a + q \cdot n$ для всех $n = 0, 1, 2, \dots$

Пусть теперь *явно* задана последовательность (a_n) . Определим *неявно* новую последовательность (A_n) так:

$$\begin{aligned} A_1 &= a_1, \\ A_2 &= A_1 + a_2 = a_1 + a_2, \\ &\dots \\ A_n &= A_{n-1} + a_n = a_1 + a_2 + \dots + a_n, \\ &\dots \end{aligned}$$

Такую пару последовательностей называют *числовым рядом*. При этом *явно заданная* последовательность (a_n) называется *последовательность членов ряда*, а *неявно заданная* последовательность (A_n) – *последовательность частных сумм ряда*. Мы будем обозначать последовательности членов ряда малыми латинскими буквами, а последовательности частных сумм – соответствующими большими.

Определение. Если последовательность частных сумм ряда сходится (имеет предел), то говорят, что *ряд сходится*, а этот предел (число) называют *суммой ряда*.

Если $\lim A_n = A$, то пишут $a_1 + a_2 + \dots + a_n + \dots = \sum_{n=1}^{+\infty} a_n = A$.

Если $\lim A_n$ не существует, то говорят, что *ряд расходится*.

Замечания. 1. Символы $a_1 + a_2 + \dots + a_n + \dots$ и $\sum_{n=1}^{+\infty} a_n$ по определению обозначают сумму сходящегося ряда, т.е. число. Однако по традиции этими символами обозначают и сам ряд (даже в случае его расходимости). Можно встретить, например, утверждение: "ряд $\sum_{n=1}^{+\infty} \frac{1}{n}$ расходится".

2. Если изменить значения последовательности членов ряда для k номеров, взяв a'_{n_1} вместо a_{n_1} , a'_{n_2} вместо a_{n_2}, \dots, a'_{n_k} вместо a_{n_k} (номера идут в порядке возрастания), то, начиная с номера n_k для частных сумм исходного ряда – A_n и измененного – A'_n будет выполняться условие

$$A'_n - A_n = (a'_{n_1} - a_{n_1}) + \dots + (a'_{n_k} - a_{n_k}), \quad n \geq n_k.$$

Таким образом, эти частные суммы отличаются на фиксированное число, и либо оба ряда – исходный и измененный – сходятся, либо оба расходятся. Более того, если оба ряда сходятся, то их суммы отличаются на то же число:

$$A' - A = (a'_{n_1} - a_{n_1}) + \dots + (a'_{n_k} - a_{n_k}).$$

Следовательно, при решении вопроса о сходимости ряда можно не обращать внимания на значения последовательности его членов для любого *конечного* числа номеров. В частности, если некоторое свойство имеет место *почти для всех номеров* (т.е. за исключением конечного их числа), то при решении вопроса о сходимости можно считать, что оно выполнено *для всех номеров*. Иногда этот прием существенно упрощает рассуждения.

Рассмотрим три примера.

$$1. f_n = a \cdot q^n \quad (a \neq 0, q \neq 0, n = 0, 1, \dots). \quad F_n = a + a \cdot q + \dots + a \cdot q^n.$$

Если $q = 1$, то $F_n = a \cdot n$. Если же $q \neq 1$, то

$$(1 - q) \cdot F_n = F_n - q \cdot F_n = a \cdot (1 - q^{n+1}) \implies F_n = \frac{a}{1 - q} - \frac{a}{1 - q} \cdot q^{n+1}.$$

Если $|q| < 1$, то $\lim q^{n+1} = 0$, и $F = \lim F_n = \frac{a}{1 - q}$. Если $|q| \geq 1$, то $\lim F_n$ не существует (ряд расходится). При $|q| > 1$ это очевидно, так как (F_n) не ограничена. Случай $|q| = 1$ будет рассмотрен ниже.

Итак, $\sum_{n=0}^{+\infty} aq^n = \frac{a}{1-q}$ при $|q| < 1$.

$$2. g_n = a + n \cdot q \quad (q \neq 0, n = 0, 1, \dots). \quad G_n = (n+1) \cdot a + \frac{n(n+1)}{2} \cdot q.$$

(G_n) не ограничена и, следовательно, предела не имеет. Ряд расходится.

$$3. h_n = (-1)^n \quad (n = 0, 1, \dots). \quad H_n = \begin{cases} 1 & \text{при четном } n \\ 0 & \text{при нечетном } n \end{cases}.$$

(H_n) – периодическая последовательность (период равен двум) и, следовательно, предела не имеет. Ряд расходится.

В этих примерах мы смогли ответить на вопрос о сходимости ряда "по определению так как нам удалось получить явное выражение для последовательностей частных сумм.

Сформулируем две основные задачи теории числовых рядов:

1) по известным свойствам (a_n) – последовательности членов ряда – решить вопрос о сходимости (A_n) – последовательности частных сумм этого ряда (*не находя явного выражения для последовательности частных сумм*).

2. если ряд сходится, то вычислить его сумму, т.е. $\lim A_n$.

Теоремы, позволяющие решить первую задачу, называют обычно *признаками сходимости рядов*.

Теорема (признак расходимости ряда). Если неверно, что $\lim a_n = 0$, то ряд расходится.

Доказательство. Предположим, что ряд сходится, т.е. существует $\lim A_n = A$.

Так как $a_n = A_n - A_{n-1}$, то по теореме о пределе суммы

$$\lim a_n = \lim(A_n - A_{n-1}) = \lim A_n - \lim A_{n-1} = A - A = 0,$$

что противоречит условию. ■

Замечание. Из этой теоремы следует, в частности, расходимость ряда в примере 1 при $|q| = 1$.

Приведем еще несколько утверждений, полезных при исследовании рядов. Они являются очевидными переформулировками теорем 2, 3 и 6 о пределах последовательностей (п.5.2).

1. Если ряды $\sum_{n=1}^{+\infty} a_n$ и $\sum_{n=1}^{+\infty} b_n$ сходятся, и их суммы равны A и B соответственно, то сходится и ряд $\sum_{n=1}^{+\infty} (\alpha a_n + \beta b_n)$ ($\alpha, \beta \in \mathbb{C}$), причем его сумма равна $\alpha A + \beta B$.

$$2. \sum_{n=1}^{+\infty} a_n = A \iff \sum_{n=1}^{+\infty} \operatorname{Re}(a_n) = \operatorname{Re}(A) \wedge \sum_{n=1}^{+\infty} \operatorname{Im}(a_n) = \operatorname{Im}(A).$$

6.2. Положительные ряды

Определение. Если значения последовательности членов ряда – вещественные неотрицательные числа, то ряд называется *положительным*.

Последовательность частных сумм положительного ряда не убывает, так как при $a_n \geq 0$ $A_n = A_{n-1} + a_n \geq A_{n-1}$. Следовательно, либо она не ограничена, либо (теорема из п.5.2) имеет предел.

Определение. Если для всех номеров $0 \leq a_n \leq b_n$, то говорят, что последовательность (a_n) *мажорируется* последовательностью b_n . Можно также говорить, что последовательность (b_n) *больше* последовательности (a_n) (последовательность (a_n) *меньше* последовательности (b_n)). Те же термины применяются и по отношению к положительным рядам.

Теорема (признак сравнения положительных рядов). Из сходимости большего ряда следует сходимость меньшего. Из расходимости меньшего ряда следует расходимость большего.

Доказательство. Пусть $0 \leq a_n \leq b_n$, $\sum_{n=1}^{+\infty} b_n = B$. Тогда

$$A_n = a_1 + \dots + a_n \leq b_1 + \dots + b_n \leq B,$$

т.е. неубывающая последовательность (A_n) ограничена и, следовательно, имеет предел.

Второе утверждение доказывается от противного. ■

Теорема (пределная форма признака сравнения). Если существует предел $\lim\left(\frac{a_n}{b_n}\right) = L$, то:

- 1) при $L \neq 0$ либо оба ряда сходятся, либо оба расходятся;
- 2) при $L = 0$ из сходимости (B_n) следует сходимость (A_n) , а из расходимости (A_n) – расходимость (B_n) .

Доказательство. 1) По определению предела последовательности в любой окрестности точки L лежат значения последовательности *почти для всех номеров*. Положим $\varepsilon = L/2$. Тогда, учитывая замечание 2 из п.6.1, можно считать, что *для всех номеров*

$$\left| \frac{a_n}{b_n} - L \right| < \frac{L}{2} \quad \text{или} \quad \frac{L}{2} < \frac{a_n}{b_n} < \frac{3L}{2}, \quad \text{или} \quad \frac{L}{2} b_n < a_n < \frac{3L}{2} b_n.$$

Пусть $\sum_{n=1}^{+\infty} b_n = B$. Тогда

$$A_n = a_1 + \dots + a_n < \frac{3L}{2}(b_1 + \dots + b_n) < \frac{3LB}{2},$$

т.е. неубывающая последовательность (A_n) ограничена и, следовательно, имеет предел.

Если же $\sum_{n=1}^{+\infty} a_n = A$, то

$$B_n = b_1 + \dots + b_n < \frac{2}{L}(a_1 + \dots + a_n) < \frac{2A}{L}$$

т.е. неубывающая последовательность (B_n) ограничена и, следовательно, имеет предел.

2) Рассмотрим ε -окрестность нуля. Для всех номеров (см. замечание 2 из п.6.1) имеем

$$\left| \frac{a_n}{b_n} - 0 \right| < \varepsilon \quad \text{или} \quad -\varepsilon < \frac{a_n}{b_n} < \varepsilon \quad \Rightarrow \quad a_n < \varepsilon b_n.$$

Пусть $\sum_{n=1}^{+\infty} b_n = B$. Тогда $A_n = a_1 + \dots + a_n < \varepsilon(b_1 + \dots + b_n) < \varepsilon B$,

т.е. неубывающая последовательность (A_n) ограничена и, следовательно, имеет предел.

Вторая часть утверждения 2) доказывается от противного. ■

Пользоваться признаками сравнения можно, имея эталонные положительные ряды (как сходящиеся, так и расходящиеся). Примерами таких эталонов являются уже известные:

- 1) семейство, порожденное арифметическими прогрессиями $\sum_{n=1}^{+\infty} (a + nq)$,
 $a > 0, q > 0$ (все ряды этого семейства расходятся);

2) семейство, порожденное геометрическими прогрессиями $\sum_{n=1}^{+\infty} (a \cdot q^n)$, $a > 0$, $q > 0$ (ряды сходятся при $q < 1$ и расходятся при $q \geq 1$).

Рассмотрим еще семейство положительных рядов $\sum_{n=1}^{+\infty} \frac{1}{n^p}$ (*ряды Дирихле*¹⁸).

Пусть $p > 1$. Сравним ряд Дирихле

$$\sum_{n=1}^{+\infty} \frac{1}{n^p} = \frac{1}{1^p} + \left(\frac{1}{2^p} + \frac{1}{3^p} \right) + \left(\frac{1}{4^p} + \frac{1}{5^p} + \frac{1}{6^p} + \frac{1}{7^p} \right) + \left(\frac{1}{8^p} + \dots + \frac{1}{15^p} \right) + \dots$$

(в каждой следующей скобке вдвое больше слагаемых, чем в предыдущей) с рядом

$$\frac{1}{1^p} + \left(\frac{1}{2^p} + \frac{1}{2^p} \right) + \left(\frac{1}{4^p} + \frac{1}{4^p} + \frac{1}{4^p} + \frac{1}{4^p} \right) + \left(\frac{1}{8^p} + \dots + \frac{1}{8^p} \right) + \dots$$

Очевидно, что второй ряд *больше*. Перепишем его в виде

$$1 + \frac{2}{2^p} + \frac{4}{4^p} + \frac{8}{8^p} + \dots = 1 + \frac{1}{2^{p-1}} + \frac{1}{(2^{p-1})^2} + \frac{1}{(2^{p-1})^3} + \dots$$

Мы получили ряд, порожденный геометрической прогрессией со знаменателем $q = 1/2^{p-1} < 1$ ($p > 1$), который сходится. Поэтому сходится и *меньший* ряд Дирихле.

Пусть теперь $p \leq 1$. Сравним ряд Дирихле

$$\sum_{n=1}^{+\infty} \frac{1}{n^p} = \frac{1}{1^p} + \frac{1}{2^p} + \left(\frac{1}{3^p} + \frac{1}{4^p} \right) + \left(\frac{1}{5^p} + \dots + \frac{1}{8^p} \right) + \left(\frac{1}{9^p} + \dots + \frac{1}{16^p} \right) + \dots$$

с рядом

$$\frac{1}{1^p} + \frac{1}{2^p} + \left(\frac{1}{4^p} + \frac{1}{4^p} \right) + \left(\frac{1}{8^p} + \frac{1}{8^p} + \frac{1}{8^p} + \frac{1}{8^p} \right) + \left(\frac{1}{16^p} + \dots + \frac{1}{16^p} \right) + \dots$$

Очевидно, что второй ряд *меньше*. Перепишем его в виде

$$\frac{1}{1^p} + \frac{1}{2^p} + \frac{2}{4^p} + \frac{4}{8^p} + \frac{8}{16^p} + \dots = 1 + \frac{1}{2} \cdot \left(\frac{1}{2^{p-1}} + \frac{1}{(2^{p-1})^2} + \frac{1}{(2^{p-1})^3} + \dots \right).$$

¹⁸Пьер Густав Лежен ДИРИХЛЕ (P.G.L. Dirichlet, 1805-1859) – немецкий математик.

Полученный ряд расходится, так как порожден геометрической прогрессией, знаменатель которой $q = 1/2^{p-1} \geq 1$ ($p \leq 1$). Поэтому расходится и *больший* ряд Дирихле. ■

Итак, ряд Дирихле $\sum_{n=1}^{+\infty} \frac{1}{n^p}$ сходится при $p > 1$ и расходится при $p \leq 1$.

В частности, расходится так называемый *гармонический ряд* $\sum_{n=1}^{+\infty} \frac{1}{n}$.

Замечание. При исследовании ряда Дирихле мы "расставляли скобки т.е. заменяли группы соседних по номерам слагаемых их суммами. *Можно показать*, что такая операция преобразует *сходящийся* ряд в *сходящийся ряд с той же суммой*, а *расходящийся положительный* (!) ряд – в *расходящийся*.

Для произвольного числового ряда эта операция, вообще говоря, недопустима. Возьмем, например, расходящийся ряд (см. пример 3 п.6.1)

$$\sum_{n=1}^{+\infty} (-1)^n = 1 - 1 + 1 - 1 + \dots$$

и расставим в нем скобки двумя способами:

$$(1 - 1) + (1 - 1) + \dots = 0; \quad 1 + (-1 + 1) + (-1 + 1) + \dots = 1.$$

Результат не нуждается в комментариях.

6.3. Вещественный ряд с чередованием знаков

Определение. Ряд

$$\sum_{n=1}^{+\infty} (-1)^{n+1} a_n = a_1 - a_2 + a_3 - a_4 + \dots,$$

где $a_n > 0$ для всех номеров, называется *рядом с чередованием знаков*.

Теорема (признак Лейбница¹⁹). Если

$$1) \lim a_n = 0; \quad 2) a_{n+1} < a_n \text{ для всех } n,$$

то ряд с чередованием знаков сходится, причем $|A - A_n| < a_{n+1}$.

¹⁹Готфрид Вильгельм ЛЕЙБНИЦ (G.W. Leibniz, 1646-1716) – немецкий математик, физик и философ. Один из основоположников математического анализа, организатор и первый президент Берлинской АН. Член Лондонского Королевского общества и Парижской АН.

Доказательство. Рассмотрим последовательность *четных* частных сумм (A_{2n}) :

$$A_{2n} = (a_1 - a_2) + \dots + (a_{2n-1} - a_{2n}) = A_{2n-2} + (a_{2n-1} - a_{2n}).$$

По условию (2) $a_{2n-1} - a_{2n} > 0$, т.е. $A_{2n} > A_{2n-2}$.

Далее,

$$A_{2n} = a_1 - (a_2 - a_3) - \dots - (a_{2n-2} - a_{2n-1}) - a_{2n}.$$

По условию (2) все скобки положительны, т.е. $A_{2n} < a_1$.

Возрастающая ограниченная сверху последовательность (A_{2n}) имеет предел. Обозначим $A = \lim A_{2n}$. По условию (1) $\lim a_{2n+1} = 0$. Поэтому

$$\lim A_{2n+1} = \lim(A_{2n} + a_{2n+1}) = \lim A_{2n} + \lim a_{2n+1} = \lim A_{2n} = A.$$

Итак, $\lim A_{2n} = \lim A_{2n+1} = A$. Это значит, что в любой окрестности числа A лежат почти все частные суммы (как четные, так и нечетные), т.е. $\lim A_n = A$, и первое утверждение теоремы доказано.

Для доказательства второго утверждения заметим, что

$$A - A_n = (-1)^{n+2}a_{n+1} + (-1)^{n+3}a_{n+2} + \dots = (-1)^{n+2}(a_{n+1} - a_{n+2} + \dots).$$

Рассмотрим ряд с чередованием знаков

$$a_{n+1} - a_{n+2} + a_{n+3} - a_{n+4} + \dots$$

Все его частные суммы положительны, ибо

$$(a_{n+1} - a_{n+2}) + (a_{n+3} - a_{n+4}) + \dots + (a_{n+2m-1} - a_{n+2m}) > 0,$$

$$(a_{n+1} - a_{n+2}) + (a_{n+3} - a_{n+4}) + \dots + (a_{n+2m-1} - a_{n+2m}) + a_{n+2m+1} > 0;$$

в то же время все они меньше, чем a_{n+1} :

$$a_{n+1} - (a_{n+2} - a_{n+3}) - \dots - a_{n+2m} < a_{n+1},$$

$$a_{n+1} - (a_{n+2} - a_{n+3}) - \dots - (a_{n+2m} - a_{n+2m+1}) < a_{n+1}$$

Следовательно, $|A - A_n| = a_{n+1} - a_{n+2} + a_{n+3} - a_{n+4} + \dots < a_{n+1}$. ■

Пример. Ряд $\sum_{n=1}^{+\infty} \frac{(-1)^{n+1}}{n}$ сходится, так как $\lim \frac{1}{n} = 0$ и $\frac{1}{n+1} < \frac{1}{n}$.

6.4. Исследование сходимости произвольных числовых рядов

Если ряд $\sum_{n=1}^{+\infty} a_n$ не является ни положительным, ни рядом с чередованием знаков, то рекомендуется следующий путь исследования его на сходимость.

1. Применить признак расходимости, т.е. попытаться вычислить предел $\lim a_n$. Если этот предел не существует, или существует, но не равен нулю, то вопрос решен – ряд расходится.

2. Если $\lim a_n = 0$, то признак расходимости не работает. В этом случае следует рассмотреть *положительный* ряд $\sum_{n=1}^{+\infty} |a_n|$, к которому можно применить теорему сравнения.

Теорема. Если ряд $\sum_{n=1}^{+\infty} |a_n|$ сходится, то сходится и ряд $\sum_{n=1}^{+\infty} a_n$ (в этом случае говорят, что ряд $\sum_{n=1}^{+\infty} a_n$ *абсолютно сходится*).

Доказательство. Рассмотрим сначала случай, когда последовательность (a_n) – вещественная, и ряд $\sum_{n=1}^{+\infty} |a_n|$ сходится. Рассмотрим два ряда:

$\sum_{n=1}^{+\infty} b_n$ и $\sum_{n=1}^{+\infty} c_n$, где

$$b_n = |a_n| + a_n, \quad c_n = |a_n| - a_n.$$

Очевидно, что для всех номеров либо $b_n = 0$, $c_n = 2 \cdot |a_n|$, либо $b_n = 2 \cdot |a_n|$, $c_n = 0$, т.е. эти ряды положительны и каждый из них меньше сходящегося ряда $\sum_{n=1}^{+\infty} 2 \cdot |a_n|$ – оба ряда сходятся. Тогда сходится

и их почленная разность (см. п.6.1) – ряд $\sum_{n=1}^{+\infty} (b_n - c_n) = \sum_{n=1}^{+\infty} a_n$.

Пусть теперь последовательность (a_n) – комплексная. Тогда $|Re(a_n)| \leq |a_n|$, $|Im(a_n)| \leq |a_n|$, и по признаку сравнения сходятся положительные ряды $\sum_{n=1}^{+\infty} |Re(a_n)|$ и $\sum_{n=1}^{+\infty} |Im(a_n)|$. Тогда по уже доказанному сходятся ряды $\sum_{n=1}^{+\infty} Re(a_n)$ и $\sum_{n=1}^{+\infty} Im(a_n)$, что равносильно сходимости ряда $\sum_{n=1}^{+\infty} a_n$ (утверждение 2 в конце п.6.1). ■

Пример. Дан комплексный ряд $\sum_{n=1}^{+\infty} \frac{1}{n^2 + i \cdot n}$. Построим положительный ряд

$$\sum_{n=1}^{+\infty} \left| \frac{1}{n^2 + i \cdot n} \right| = \sum_{n=1}^{+\infty} \frac{1}{\sqrt{n^4 + n^2}}.$$

Этот ряд сходится, так как $\frac{1}{\sqrt{n^4 + n^2}} < \frac{1}{n^2}$, а ряд Дирихле $\sum_{n=1}^{+\infty} \frac{1}{n^2}$, как известно, сходится. Мы показали, таким образом, что исследуемый комплексный ряд $\sum_{n=1}^{+\infty} \frac{1}{n^2 + i \cdot n}$ сходится абсолютно.

Замечания. 1. Проверка ряда на абсолютную сходимость результативна только при положительном ответе: абсолютно сходящийся ряд сходится. Если же абсолютной сходимости нет, то вопрос о сходимости остается открытым.

2. Для абсолютно сходящегося ряда имеет место неравенство

$$\left| \sum_{n=1}^{+\infty} a_n \right| \leq \sum_{n=1}^{+\infty} |a_n|.$$

Рассмотрим два простых признака сходимости, применимых к произвольным числовым рядам.

Теорема. (Признак Коши²⁰). Если существует предел

$$K = \lim |a_n|^{1/n},$$

то

- 1) при $K < 1$ ряд $\sum_{n=1}^{+\infty} a_n$ сходится (абсолютно);
- 2) при $K > 1$ этот ряд расходится.
- 3) при $K = 1$ признак Коши не работает – существуют и сходящиеся, и расходящиеся ряды с $K = 1$.

Доказательство. 1. Пусть $K < 1$. По определению предела почти для всех номеров выполняется неравенство

$$\left| |a_n|^{1/n} - K \right| < \frac{1-K}{2} \quad \left(\frac{1-K}{2} > 0 \right).$$

²⁰Огюстен Луи КОШИ (A.L. Cauchi, 1789-1857) – французский математик, член Лондонского Королевского общества и почти всех академий мира, один из крупнейших математиков XIX века.

Отсюда

$$|a_n|^{1/n} - K < \frac{1-K}{2} \implies |a_n|^{1/n} < \frac{1+K}{2} \implies |a_n| < \left(\frac{1+K}{2}\right)^n.$$

Мы показали, что положительный ряд $\sum_{n=1}^{+\infty} |a_n|$ мажорируется сходящимся рядом, порожденным геометрической прогрессией со знаменателем $\frac{1+K}{2} < 1$. Утверждение доказано.

2. Пусть $K > 1$. По определению предела почти для всех номеров

$$\left| |a_n|^{1/n} - K \right| < \frac{K-1}{2} \quad \left(\frac{K-1}{2} > 0 \right),$$

откуда

$$-\frac{K-1}{2} < |a_n|^{1/n} - K \implies |a_n|^{1/n} > \frac{1+K}{2} > 1,$$

т.е. $|a_n| > 1$, и неверно, что $\lim a_n = 0$. По признаку расходимости ряд расходится.

3. Для очевидно расходящегося ряда с $a_n \equiv 1$ $K = 1$, но и для сходящегося ряда Дирихле $\sum_{n=1}^{+\infty} \frac{1}{n^2}$ тоже $K = \lim \left(\frac{1}{n^2} \right)^{1/n} = 1$. ■

Теорема (признак Д'Аламбера²¹). Если существует предел

$$D = \lim \left| \frac{a_{n+1}}{a_n} \right|,$$

то

- 1) при $D < 1$ ряд $\sum_{n=1}^{+\infty} a_n$ сходится (абсолютно);
- 2) при $D > 1$ этот ряд расходится.
- 3) при $D = 1$ признак Д'Аламбера не работает – существуют и сходящиеся и расходящиеся ряды с $D = 1$.

Доказательство аналогично предыдущей теореме. Попробуйте провести его самостоятельно.

Если рассмотренные выше простейшие приемы исследования ряда на сходимость не срабатывают, мы рекомендуем обратиться за консультацией к математику-профессионалу.

²¹Жан Лерон Д'АЛАМБЕР (J. le Rond d'Alembert, 1717-1783) – французский математик и механик, член многих академий, один из авторов знаменитой "Энциклопедии наук, искусств и ремесел".

6.5. Оценивание суммы сходящегося числового ряда

Если числовой ряд сходится, то его сумма $A = \lim A_n = \sum_{n=1}^{\infty} a_n$ есть некоторое комплексное число.

Определение. Оценкой суммы сходящегося числового ряда будем называть любой круг на комплексной плоскости, накрывающий точку A (сумму ряда).

Круг задается своими центром и радиусом. Очевидно, из двух кругов-оценок лучше тот, радиус которого меньше.

Примеры. 1. Для вещественного ряда с чередованием знаков $\sum_{n=1}^{\infty} a_n$, удовлетворяющего условиям теоремы Лейбница, $|A - A_n| < a_{n+1}$, т.е. сумма A лежит внутри круга с центром в точке A_n и радиусом a_{n+1} . Вследствие вещественности ряда этот круг превращается в интервал $]A_n - a_{n+1}, A_n + a_{n+1}[$.

2. Пусть сходимость ряда $\sum_{n=1}^{\infty} a_n$ установлена с помощью признака Коши, т.е. существует конечный предел $K = \lim |a_n|^{1/n} < 1$. Выберем число q между K и единицей ($K < q < 1$).

По определению предела последовательности, начиная с некоторого номера n_1 будет выполняться неравенство $|a_n|^{1/n} < q$, или $|a_n| < q^n$. Поэтому при $n \geq n_1$

$$|A - A_n| \leq |a_{n+1}| + |a_{n+2}| + \dots < q^{n+1} + q^{n+2} + \dots = \frac{q^{n+1}}{1-q}.$$

Пусть требуется построить круг-оценку с радиусом ε . Если выполнено неравенство $\frac{q^{n_1+1}}{1-q} < \varepsilon$, то в качестве центра круга можно взять число A_{n_1} . Иначе следует найти такой номер n_2 (больший, чем n_1), что $\frac{q^{n_2+1}}{1-q} \leq \varepsilon$. Такой номер, очевидно, найдется.

Рассмотрим ряд $\sum_{n=1}^{\infty} \frac{n^2}{3^n}$. Здесь $K = \lim \left| \frac{n^2}{3^n} \right|^{1/n} = \frac{1}{3}$. Возьмем $q = \frac{1}{2}$ и найдем номер n_1 , начиная с которого $\left| \frac{n^2}{3^n} \right|^{1/n} \leq \frac{1}{2}$. Перебором первых натуральных чисел находим, что $n_1 = 13$. Пусть, например, $\varepsilon = 10^{-6}$. Решая неравенство $\frac{0.5^n}{1-0.5} < 10^{-6}$, найдем, что $n_2 = 20$. Итак, круг с центром в точке $A_{20} = \sum_{n=1}^{20} \frac{n^2}{3^n} \approx 1.4999999$ и радиусом 10^{-6} заведомо накрывает сумму ряда (можно показать, что сумма равна 1.5).

3. Аналогично строится оценка суммы, если сходимость ряда установлена с помощью признака Д'Аламбера. Оценим, например, сумму ряда²² $\sum_{n=1}^{\infty} \frac{1}{n!}$. Здесь $D = \lim \left| \frac{1/(n+1)!}{1/n!} \right| = \lim \frac{1}{n+1} = 0$.

Возьмем $q = \frac{1}{5}$. Тогда из неравенства $\left| \frac{a_{n+1}}{a_n} \right| = \frac{1}{n+1} < \frac{1}{5}$ находим, что $n_1 = 5$. Следовательно, при $n \geq 5$

$$|A - A_n| = a_n + a_{n+1} + \dots < a_n \cdot (0.2 + 0.2^2 + \dots) = \frac{0.25}{n!}.$$

Полагая, например, $\varepsilon = 10^{-6}$ и решая (перебором натуральных чисел) неравенство $\frac{0.25}{n!} < 10^{-6}$, находим, что $n_2 = 9$. Мы установили, что круг с центром в точке $A_9 = 1 + \frac{1}{1!} + \frac{1}{2!} + \dots + \frac{1}{9!} \approx 2.7182815$ и радиусом 10^{-6} заведомо накрывает сумму ряда.

Замечания. 1. Сумма ряда из последнего примера встречается так часто, что для нее введено стандартное обозначение $e = \sum_{n=1}^{\infty} \frac{1}{n!}$.

2. Отметим, что за улучшение оценки, т.е. за уменьшение радиуса круга-оценки, всегда приходится платить увеличением количества слагаемых в частной сумме A_n , принимаемой за центр этого круга.

3. В практических вычислениях часто употребляют термин "скорость сходимости ряда". Рассмотрим два примера.

$$1) \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^4}; \quad 2) \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{\lg(n+1)}.$$

Признак Лейбница показывает, что оба ряда сходятся. Однако для получения круга-оценки с радиусом 10^{-4} необходимо просуммировать:

для первого ряда 9 слагаемых, так как $|A - A_9| < a_{10} = 10^{-4}$;

для второго ряда $10^{10^4} - 1$ слагаемых, так как только начиная с этого номера выполняется неравенство $\frac{1}{\lg(n+1)} < 10^{-4}$.

Ясно, что вычислить сумму девяти слагаемых в первом примере нетрудно (можно и вручную). Что же касается второй суммы, то даже располагая компьютером, который вычисляет одно слагаемое ряда за 10^{-10} секунды, придется затратить на ее вычисление $10^{10^4-10} = 10^{9990}$ секунд, т.е. примерно 10^{9983} лет (!!).

Эти два простых примера показывают, какой смысл вкладывают в слова "скорость сходимости ряда".

²²Для любого натурального числа n $n! = 1 \cdot 2 \cdot 3 \dots n$ (читается " n -факториал"). По определению полагают $0! = 1$.

Глава 7. СТЕПЕННОЙ РЯД. АНАЛИТИЧЕСКИЕ ФУНКЦИИ

7.1. Степенной ряд

Определение. *Степенным рядом* называют семейство числовых рядов вида

$$a_0 + a_1 z + \dots + a_n z^n + \dots = \sum_{n=0}^{+\infty} a_n z^n, \quad (7.1.1)$$

где $(a_n)_0^{+\infty}$ – заданная последовательность комплексных чисел, а z – комплексная переменная. Фиксируя значение этой переменной, мы извлечем из семейства различные числовые ряды.

Числа a_k , $k = 0, 1, \dots$, называют *коэффициентами* степенного ряда.

Замечание. При $z = 0$ и $n = 0$ в (7.1.1) возникает неопределенность: $a_0 \cdot 0^0$. Поэтому лучше было бы писать $a_0 + \sum_{n=1}^{+\infty} a_n z^n$. Однако пишут короче $- \sum_{n=0}^{+\infty} a_n z^n$, понимая это как $a_0 + \sum_{n=1}^{+\infty} a_n z^n$.

Пример. Пусть дан степенной ряд

$$1 + \frac{z}{1} + \frac{z^2}{2} + \dots + \frac{z^n}{n} + \dots = 1 + \sum_{n=1}^{+\infty} \frac{z^n}{n}.$$

Положив $z = -1$, получим числовой ряд

$$1 - 1 + \frac{1}{2} - \frac{1}{3} + \dots = \sum_{n=2}^{+\infty} \frac{(-1)^n}{n} \quad (\text{сходящийся!}).$$

Положив $z = 1$, получим числовой ряд

$$1 + 1 + \frac{1}{2} + \frac{1}{3} + \dots = 2 + \sum_{n=2}^{+\infty} \frac{1}{n} \quad (\text{расходящийся!}).$$

Из этого примера видно, что при одних значениях переменной степенной ряд может превращаться в сходящийся числовой ряд, а при других – в расходящийся. Поэтому представляет интерес множество точек комплексной плоскости, в которых степенной ряд сходится.

Отметим сначала, что при $z = 0$ сходится любой степенной ряд, так как его сумма есть просто a_0 . Если $z \neq 0$, то применим к степенному ряду (7.1.1) признак Д'Аламбера, т.е. попробуем вычислить предел

$$D(z) = \lim \left| \frac{a_{n+1} z^{n+1}}{a_n z^n} \right| = |z| \cdot \lim \left| \frac{a_{n+1}}{a_n} \right|.$$

Мы сознательно обозначили этот предел $D(z)$, чтобы подчеркнуть, что (в отличие от случая числового ряда) он зависит от точки z , в которой вычисляется.

Если $\lim \left| \frac{a_{n+1}}{a_n} \right| = 0$, то $D(z) = 0$ при всех $z \in \mathbb{C}$, т.е. степенной ряд сходится (абсолютно) на всей плоскости.

Если $\lim \left| \frac{a_{n+1}}{a_n} \right| = +\infty$, то $D(z) = +\infty$ при всех $z \neq 0$, т.е. степенной ряд сходится только в нуле.

Если $\lim \left| \frac{a_{n+1}}{a_n} \right| = \alpha > 0$, то степенной ряд сходится (абсолютно) при $D(z) = \alpha \cdot |z| < 1$, т.е. внутри круга с центром в начале координат и радиусом $R = \frac{1}{\alpha}$. Вне этого круга $D(z) = \alpha \cdot |z| > 1$, т.е. степенной ряд расходится. На границе круга $D(z) = 1$, и признак Д'Аламбера не дает ответа на вопрос о сходимости степенного ряда.

Замечания. 1. Аналогичные утверждения можно было бы получить, пользуясь вместо признака Д'Аламбера признаком Коши.

2. Если последовательности $\left| \frac{a_{n+1}}{a_n} \right|$ и $|a_n|^{1/n}$ не имеют пределов, то признаки Д'Аламбера и Коши не работают. Однако можно показать, что и в этом случае у степенного ряда существует круг сходимости с центром в начале координат.

Если договориться считать начало координат "кругом" нулевого радиуса, а комплексную плоскость – "кругом" бесконечного радиуса, то мы видим, что справедлива

Теорема. У всякого степенного ряда есть круг сходимости (с центром в начале координат). Внутри этого круга степенной ряд сходится (абсолютно), а вне его – расходится.

Радиус круга сходимости степенного ряда принято называть радиусом сходимости ряда.

Примеры. 1. Применим к ряду $\sum_{n=0}^{+\infty} \frac{z^n}{n!}$ признак Д'Аламбера:

$$D(z) = \lim \left| \frac{z^{n+1} n!}{(n+1)! z^n} \right| = |z| \cdot \lim \frac{1}{n+1} \equiv 0$$

Ряд сходится абсолютно на всей комплексной плоскости.

2. К ряду $\sum_{n=0}^{+\infty} (2 + (-1)^n) \cdot z^n$ признак Д'Аламбера неприменим, так как предел

$$D(z) = \lim \left| \frac{(2 + (-1)^{n+1}) \cdot z^{n+1}}{(2 + (-1)^n) \cdot z^n} \right| = |z| \cdot \lim \frac{2 + (-1)^{n+1}}{2 + (-1)^n}$$

не существует ни при каком значении переменной z (кроме нуля). Используя признак Коши

$$K(z) = \lim |(2 + (-1)^n) \cdot z^n|^{1/n} = |z| \cdot \lim |2 + (-1)^n|^{1/n} = |z|,$$

получаем, что этот ряд сходится внутри круга единичного радиуса. Отдельно следует изучить поведение ряда на границе круга, где признак Коши не работает. Нетрудно видеть, что при $|z| = 1$ $\lim |(2 + (-1)^n)z^n|$ не существует, и, следовательно, во всех точках границы круга ряд расходится (признак расходимости!).

3. Применив к ряду $\sum_{n=1}^{+\infty} \frac{z^n}{n}$ признак Д'Аламбера, получим

$$D(z) = \lim \left| \frac{z^{n+1} n}{(n+1) z^n} \right| = |z| \cdot \lim \frac{n}{n+1} = |z|.$$

Ряд сходится абсолютно при $|z| < 1$ и расходится при $|z| > 1$. Как показано выше, на окружности $|z| = 1$ есть точки, в которых ряд сходится (например, $z = -1$), и точки, в которых ряд расходится (например, $z = 1$).

4. Применив к ряду $\sum_{n=1}^{+\infty} \frac{z^n}{n^2}$ признак Д'Аламбера, получим

$$D(z) = \lim \left| \frac{z^{n+1} n^2}{(n+1)^2 z^n} \right| = |z| \cdot \lim \frac{n^2}{(n+1)^2} = |z|.$$

Ряд сходится абсолютно при $|z| < 1$ и расходится при $|z| > 1$. Во всех точках границы круга сходимости ряд сходится абсолютно ($\sum_{n=1}^{+\infty} \frac{1}{n^2}$ – ряд Дирихле с $p > 1$).

5. Применив к ряду $\sum_{n=0}^{+\infty} n^n \cdot z^n$ признак Коши, получим

$$K(z) = \lim |n^n z^n|^{1/n} = \lim(|z| \cdot n) = +\infty \quad (z \neq 0).$$

Этот ряд сходится только в нуле.

Замечания. 1. Как видно из примеров **2 – 4**, на границе круга сходимости степенные ряды могут вести себя по-разному.

2. Можно рассматривать степенные ряды, центр круга сходимости которых расположен не в нуле, а в некоторой точке p комплексной плоскости. Эти ряды имеют вид

$$a_0 + a_1(z - p) + \dots + a_n(z - p)^n + \dots = \sum_{n=0}^{+\infty} a_n(z - p)^n.$$

7.2. Аналитические функции

Если радиус круга сходимости степенного ряда отличен от нуля, то внутри этого круга (а может быть, и в некоторых точках его границы) задана функция

$$f(z) = a_0 + a_1z + \dots + a_nz^n + \dots = \sum_{n=0}^{+\infty} a_nz^n.$$

Определение. *Аналитической функцией* называется функция, являющаяся суммой степенного ряда.

Замечание. В дальнейшем, если не оговорено противное, мы рассматриваем только аналитические функции, определяемые степенными рядами, у которых центр круга сходимости расположен в начале координат.

Аналитическая функция есть естественное обобщение полинома и наследует ряд его полезных свойств. Так, например, *можно показать*, что она непрерывна внутри круга сходимости.

В свою очередь, полином оказывается частным случаем аналитической функции (все коэффициенты определяющего его ряда, начиная с некоторого, равны нулю).

Можно показать, что сумма и произведение двух аналитических функций суть аналитические функции (в меньшем из их кругов сходимости), причем сложение и умножение аналитических функций выполняются так же, как соответствующие операции над полиномами.

Деление аналитических функций рассмотрим на примере. Пусть

$$f(z) = 1 - z + z^2 - \dots + (-1)^n z^n + \dots \quad (|z| < 1),$$

$$g(z) = 1 + z + z^2 + \dots + z^n + \dots \quad (|z| < 1).$$

Запишем функцию-частное в виде степенного ряда, коэффициенты которого подлежат определению.

$$(f/g)(z) = \frac{1 - z - \dots + (-1)^n z^n + \dots}{1 + z + \dots + z^n + \dots} = a_0 + a_1 z + a_2 z^2 + \dots + a_n z^n + \dots$$

Приведем обе части равенства к общему знаменателю и приравняем коэффициенты при одинаковых степенях переменной в числителях:

$$\begin{aligned} a_0 &= 1 \\ a_0 + a_1 &= -1 \\ a_0 + a_1 + a_2 &= 1 \\ a_0 + a_1 + a_2 + a_3 &= -1 \\ \dots &\dots \end{aligned}$$

Отсюда

$$a_0 = 1, a_1 = -2, a_2 = 2, a_3 = -2, \dots, a_n = 2 \cdot (-1)^n, \dots$$

Итак,

$$(f/g)(z) = 1 - 2z + 2z^2 - \dots + 2(-1)^n z^n + \dots \quad (|z| < 1).$$

Аналогично производится деление любых аналитических функций (при условии, что $g(0) \neq 0$).

Замечание. Можно показать, что если R – меньший из радиусов сходимости рядов, определяющих функции f и g , а r – наименьший из модулей нулей функции g , то радиус сходимости ряда-частного есть меньшее из чисел R и r .

7.3. Примеры аналитических функций

Знакомство с аналитическими функциями начнем с *экспоненты*

$$\exp(z) = 1 + \frac{z}{1} + \frac{z^2}{2!} + \dots + \frac{z^n}{n!} + \dots = \sum_{n=0}^{+\infty} \frac{z^n}{n!}.$$

Как было установлено в примере 1 п.7.1, эта функция определена на всей комплексной плоскости. Рассмотрим некоторые ее свойства.

1. $\exp(0) = 1$ (получается подстановкой).
2. $\exp(x) \cdot \exp(y) = \exp(x + y)$.

Доказательство. Перемножая ряды

$$\exp(x) = 1 + \frac{x}{1} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots, \quad \exp(y) = 1 + \frac{y}{1} + \frac{y^2}{2!} + \frac{y^3}{3!} + \dots,$$

получим

$$\begin{aligned} \exp(x) \cdot \exp(y) &= 1 + \frac{x+y}{1!} + \frac{x^2 + 2xy + y^2}{2!} + \frac{x^3 + 3x^2y + 3xy^2 + y^3}{3!} + \dots = \\ &= 1 + \frac{x+y}{1!} + \frac{(x+y)^2}{2!} + \frac{(x+y)^3}{3!} + \dots = \exp(x+y). \end{aligned} \quad \blacksquare$$

$$3. \exp(z) \cdot \exp(-z) = \exp(z + (-z)) = \exp(0) = 1.$$

Из этого *тождества* следует, во-первых, что экспонента не обращается в нуль, а, во-вторых, что

$$\exp(-z) = \frac{1}{\exp(z)}.$$

$$4. \exp(1) = \sum_{n=0}^{+\infty} \frac{1}{n!} = e. \text{ Далее, если } m \in \mathbb{N}, \text{ то}$$

$$\exp(m) = \exp(\underbrace{1 + 1 + \dots + 1}_m \text{ слагаемых}) = (\exp(1))^m = e^m.$$

5. На вещественной оси экспонента положительна и возрастает.

Доказательство. Если $x \in \mathbb{R}$, $x > 0$, то

$$\exp(x) = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \dots + \frac{x^n}{n!} + \dots > 0; \quad \exp(-x) = \frac{1}{\exp(x)} > 0.$$

$$y > x \implies y - x > 0 \implies \exp(y - x) > 1,$$

$$\exp(y) = \exp(y - x + x) = \exp(y - x) \cdot \exp(x) \implies \exp(y) > \exp(x).$$

■

6. На мнимой оси ($y \in \mathbb{R}$)

$$\begin{aligned} \exp(iy) &= \sum_{n=0}^{+\infty} \frac{(iy)^n}{n!} = \sum_{k=0}^{+\infty} \frac{(iy)^{2k}}{(2k)!} + i \cdot \sum_{k=0}^{+\infty} \frac{i^{2k} y^{2k+1}}{(2k+1)!} = \\ &= \sum_{k=0}^{+\infty} (-1)^k \frac{y^{2k}}{(2k)!} + i \cdot \sum_{k=0}^{+\infty} (-1)^k \frac{y^{2k+1}}{(2k+1)!}. \end{aligned} \quad (7.3.1)$$

Не имея возможности в рамках нашего курса доказать это, сообщим, что ряды, стоящие в вещественной и мнимой частях формулы (7.3.1) определяют "всем известные" функции косинус и синус

$$\cos(y) = \sum_{m=0}^{+\infty} (-1)^m \frac{y^{2m}}{(2m)!}, \quad \sin(y) = \sum_{m=0}^{+\infty} (-1)^m \frac{y^{2m+1}}{(2m+1)!}. \quad (7.3.2)$$

Повторим: эти формулы – не определения. Они могут быть доказаны.

Теперь мы можем записать (7.3.1) в виде

$$\exp(iy) = \cos(y) + i \cdot \sin(y) \quad (7.3.3)$$

(эта формула уже была "декларирована" в п.2.3).

Заменяя y на $(-y)$ и учитывая четность косинуса и нечетность синуса (которые, кстати, следуют из формул (7.3.2)), получим

$$\exp(-iy) = \cos(y) - i \cdot \sin(y). \quad (7.3.4)$$

Формулы (7.3.3) и (7.3.4) называются формулами Эйлера²³.

Формулы Эйлера можно переписать так (здесь $y \in \mathbb{R}$):

$$\cos(y) = \frac{\exp(iy) + \exp(-iy)}{2}, \quad \sin(y) = \frac{\exp(iy) - \exp(-iy)}{2i}. \quad (7.3.5)$$

Определение. Аналитические функции косинус и синус задаются на комплексной плоскости формулами Эйлера (7.3.5)

$$\cos(z) = \frac{\exp(iz) + \exp(-iz)}{2}, \quad \sin(z) = \frac{\exp(iz) - \exp(-iz)}{2i}; \quad z \in \mathbb{C}.$$

Замечание. Не следует ожидать, что у этих *новых* функций (имена те же, но область определения другая!) сохранятся все, известные из школы, свойства.

Например, "основное тригонометрическое тождество" сохраняется

$$\begin{aligned} \cos^2(z) + \sin^2(z) &= \left(\frac{\exp(iz) + \exp(-iz)}{2}\right)^2 + \left(\frac{\exp(iz) - \exp(-iz)}{2i}\right)^2 = \\ &= \frac{\exp(2iz) + 2 + \exp(-2iz) - \exp(2iz) + 2 - \exp(-2iz)}{4} \equiv 1, \end{aligned}$$

²³Леонард ЭЙЛЕР (L. Euler, 1707-1783) – математик, физик, механик и астроном, член Петербургской АН, Парижской АН, один из создателей вариационного исчисления, аналитической теории чисел, дифференциальной геометрии, автор многочисленных открытий в математическом анализе и других областях математики.

а неравенства $|\cos(z)| \leq 1$, $|\sin(z)| \leq 1$ – нет. Например,

$$\cos(i) = \frac{\exp(i \cdot i) + \exp(-i \cdot i)}{2} = \frac{\exp(-1) + \exp(1)}{2} = \frac{e + 1/e}{2} > 1.$$

7. Если $z = x + iy$ ($x, y \in \mathbb{R}$), то

$$\exp(z) = \exp(x + iy) = \exp(x) \cdot \exp(iy) = \exp(x) \cdot (\cos(y) + i \cdot \sin(y)).$$

Эта формула сводит вычисление комплексной экспоненты к вычислению вещественных экспоненты, косинуса и синуса.

8. Из 7 следует, что комплексная экспонента – периодическая функция с периодом $2\pi i$:

$$\begin{aligned} \exp(z + 2\pi i) &= \exp(x + i(y + 2\pi)) = \\ &= \exp(x) \cdot (\cos(y + 2\pi) + i \cdot \sin(y + 2\pi)) = \\ &= \exp(x) \cdot (\cos(y) + i \cdot \sin(y)) = \exp(x + iy) = \exp(z). \end{aligned}$$

9. Из свойства 7 следует также, что

$$|\exp(z)| = \exp(\operatorname{Re}(z)), \quad \arg(\exp(z)) = \operatorname{Im}(z).$$

7.4. Обратная функция

Пусть $f : X \rightarrow Y$, $x \in X$, $y = f(x) \in Y$ (рис.7.1).

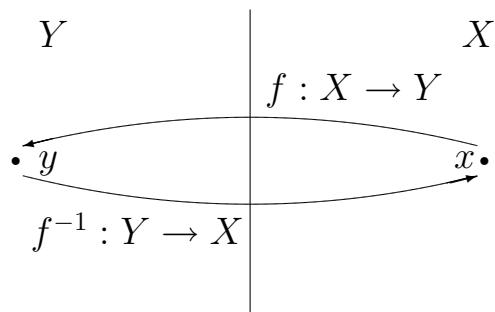


Рис.7.1

В соответствии с принятым пониманием слова функция (оператор, отображение) из каждой точки множества X , на котором определена f , выходит *ровно одна* стрелка. В множестве Y , где лежат значения функции f (они не обязательно исчерпывают это множество) ситуация иная: в некоторых его точках может не оказаться *ни одного* окончания стрелки, в других точках может оканчиваться *как угодно много* стрелок

и, наконец, в каких-то точках может оканчиваться *ровно по одной* стрелке. Договоримся считать в этом пункте, что Y – множество значений функции f , т.е. в каждой его точке оканчивается хотя бы одна стрелка. В таком случае говорят, что f отображает X на (а не в) Y и пишут $Y = f(X)$.

Примеры. 1. $f : \mathbb{C} \rightarrow \mathbb{C}$, $w = f(z) = z^2$. В точке $w = 0$ оканчивается только одна стрелка: единственным прообразом нуля является ноль. Во всех остальных точках плоскости оканчивается по две стрелки (у каждой из них два прообраза $z_1 = |w|^{1/2} \cdot \exp\left(i \cdot \frac{\arg(w)}{2}\right)$ и $z_2 = -z_1$).

2. $f : \mathbb{C} \rightarrow \{0\}$, $w = f(z) \equiv 0$. Множество значений этой функции (функции-константы) состоит всего из одной точки, в которой оканчиваются стрелки, выходящие из всех точек комплексной плоскости – области определения f .

3. $\sin : \mathbb{R} \rightarrow [-1, 1]$, $y = \sin(x)$. В силу периодичности "школьного" синуса в каждой точке сегмента $[-1, 1]$ оканчивается бесконечно много стрелок. Так, например, *полный прообраз* нуля состоит из точек $x_k = k\pi$, $k \in \mathbb{Z}$.

4. $\sin : [-\pi/2, \pi/2] \rightarrow [-1, 1]$, $y = \sin(x)$. Эта функция отлична от рассмотренной в примере 3, поскольку у них разные области определения. Функцию из примера 4 принято называть *сужением* "школьного" синуса на сегмент $[-\pi/2, \pi/2]$. Теперь в каждой точке множества значений оканчивается ровно одна стрелка. Говорят, что *полный прообраз* точки $y \in [-1, 1]$ состоит из одной точки $x \in [-\pi/2, \pi/2]$.

Поставим теперь вопрос об изменении "направления действия" отображения f . При этом каждый образ должен поменяться местами со своим полным прообразом. Очевидно, что это возможно только в ситуации, когда полный прообраз каждой точки $y = f(x) \in Y$ состоит ровно из одной точки $x \in X$, т.е. когда в каждой точке множества Y оканчивается ровно одна стрелка. Если это условие выполнено, то следует лишь изменить направление каждой стрелки. Тогда образы станут прообразами, а прообразы – образами, множество значений – областью определения, а область определения – множеством значений. Тем самым будет построена новая функция, которую называют *обратной* к функции f и обозначают f^{-1} (рис.7.1).

Из четырех рассмотренных выше примеров лишь в четвертом выполнены условия для построения обратной функции. Эту функцию

следует называть $\sin^{-1} : [-1, 1] \rightarrow [-\pi/2, \pi/2]$ (ее часто обозначают \arcsin). Точно так же сужение косинуса на $[0, \pi]$ имеет обратную функцию $\cos^{-1} : [-1, 1] \rightarrow [0, \pi]$, обозначаемую часто \arccos .

Для пары функций

$$f : X \rightarrow Y = f(X), \quad \text{и} \quad f^{-1} : Y \rightarrow X = f^{-1}(Y)$$

имеет место тождество

$$(f^{-1} \circ f)(x) \equiv x, \quad x \in X.$$

"Если из произвольной точки $x \in X$ перейти по единственной, начинающейся в ней стрелке, в точку $y \in Y$, а из нее пойти по единственной, оканчивающейся в ней стрелке в обратном направлении, то придем опять в точку x ".

Аналогично,

$$(f \circ f^{-1})(y) \equiv y, \quad y \in Y.$$

Отметим, наконец, что $(f^{-1})^{-1} = f$, и естественно называть f и f^{-1} *взаимно обратными* функциями.

7.5. Логарифм. Степенная функция

Вследствие своей периодичности экспонента не имеет обратной функции. В этом пункте мы рассмотрим сужение экспоненты на полосу

$$-\pi < \operatorname{Im}(z) \leq \pi,$$

которое будем по-прежнему обозначать \exp . Покажем, что

$$\exp(z_1) = \exp(z_2) \implies z_1 = z_2.$$

1. Из равенства $\exp(x_1 + iy_1) = \exp(x_2 + iy_2)$, следует, что

$$\exp(x_1) = |\exp(x_1 + iy_1)| = |\exp(x_2 + iy_2)| = \exp(x_2).$$

Но на вещественной оси экспонента возрастает (п.7.3, свойство 5). Поэтому из равенства $\exp(x_1) = \exp(x_2)$ следует, что $x_1 = x_2$.

2. Из равенства $\exp(z_1) = \exp(z_2)$ следует, что

$$y_2 = \arg(\exp(z_2)) = \arg(\exp(z_1)) + 2k\pi = y_1 + 2k\pi, \quad k \in \mathbb{Z}.$$

Но $y_1, y_2 \in]-\pi, \pi]$ и, следовательно, $y_1 = y_2$.

Таким образом, $z_1 = z_2$, т.е. у каждого образа есть единственный прообраз!

Итак, сужение экспоненты на полосу $-\pi < \operatorname{Im}(z) \leq \pi$ обратимо. Обратная функция называется *логарифм* и обозначается \ln или \log_e .

Замечание. Рассматриваемый нами логарифм называется *натуральным* или *неперовым*²⁴. В докомпьютерную эпоху широко применялся при вычислениях *десятичный* (бриггов²⁵) логарифм $\log_{10}(x) = \frac{\ln(x)}{\ln(10)}$. В теории информации удобен *двоичный* логарифм, определяемый равенством $\log_2(x) = \frac{\ln(x)}{\ln(2)}$. Репетиторы до сих пор любят противоестественные игры в никому, кроме них самих, не нужные "логарифмы x по основанию a " (например, $\log_{\sqrt{5}}(x)$).

Логарифм определен на множестве значений экспоненты, т.е. на всей комплексной плоскости, кроме нуля. Значения логарифма заполняют полосу $-\pi < \operatorname{Im}(z) \leq \pi$. Имеют место тождества

$$\exp(\ln(z)) \equiv z \ (z \neq 0); \quad \ln(\exp(z)) \equiv z \ (-\pi < \operatorname{Im}(z) \leq \pi).$$

Из свойства 9 экспоненты имеем

$$\operatorname{Re}(\ln(z)) = \ln(|z|); \quad \operatorname{Im}(\ln(z)) = \arg(z) \implies \ln(z) = \ln(|z|) + i \cdot \arg(z).$$

Эта формула сводит вычисление комплексного логарифма к вычислению вещественного логарифма.

Можно показать, что логарифм – аналитическая функция в любом круге, не содержащем начала координат. Степенной ряд, суммой которого является логарифм, будет получен позже.

Рассмотрим теперь *степенную* функцию. Если $n \in \mathbb{N}$, то выражение z^n , $z \in \mathbb{C}$ – это сокращенная запись произведения:

$$z^n = \underbrace{z \cdot z \cdot \dots \cdot z}_{n \text{ сомножителей}} .$$

²⁴Джон НЕПЕР (J. Napier, 1550-1617) – шотландский математик. В работе "Описание таблиц логарифмов" (1614) Непер изложил свойства логарифмов, правила пользования таблицами и примеры их применений.

²⁵Генри БРИГГ (H. Briggs, 1561-1630) – английский математик. Составил и издал таблицы логарифмов с 14 десятичными знаками. Опубликовал несколько астрономических и географических работ, в которых пропагандировал идеи И. Кеплера. Десятичные логарифмы были необходимым пособием в вычислениях до появления калькуляторов. Сейчас они практически забыты.

Выражение z^{-n} , $z \in \mathbb{C}$, $z \neq 0$, $n \in \mathbb{N}$ – это рациональная дробь $\frac{1}{z^n}$.

А как понимать выражения z^π или z^i ? Как вычислить произведение, в котором π сомножителей или i сомножителей?

При $z, x \in \mathbb{C}$, $z \neq 0$ полагают *по определению*

$$z^x = \exp(x \cdot \ln(z)).$$

Эта функция при фиксированном $x \in \mathbb{C}$ определена на всей комплексной плоскости, за исключением начала координат. Для вещественных z и x она определена при $z > 0$ и всех $x \in \mathbb{R}$. Заметим, что при *целых* показателях степени получаем известный ранее результат (проверьте это, используя свойства экспоненты).

7.6. Матричная экспонента

Пусть A – квадратная $(n \times n)$ -матрица. Рассмотрим матричный степенной ряд

$$I + \frac{A}{1!} + \frac{A^2}{2!} + \frac{A^3}{3!} + \dots + \frac{A^n}{n!} + \dots = \sum_{k=0}^{+\infty} \frac{A^k}{k!}. \quad (7.6.1)$$

Покажем, что этот ряд сходится абсолютно при любой матрице A , т.е. абсолютно сходятся все n^2 числовых рядов

$$\sum_{k=0}^{+\infty} \frac{(A^k)_{jm}}{k!}, \quad (j, m = 1, \dots, n). \quad (7.6.2)$$

Обозначим $M = \max_{j,m}(|a_{jm}|)$. Тогда при всех $j, m = 1, \dots, n$

$$|(A^2)_{jm}| = \left| \sum_{r=1}^n a_{jr} a_{rm} \right| \leq \sum_{r=1}^n |a_{jr}| \cdot |a_{rm}| \leq nM^2;$$

$$|(A^3)_{jm}| = \left| \sum_{r=1}^n (A^2)_{jr} \cdot a_{rm} \right| \leq n \cdot (nM^2) \cdot M = n^2M^3,$$

и вообще $|(A^k)_{jm}| \leq n^{k-1}M^k$. Поэтому

$$\begin{aligned} \sum_{k=0}^K \frac{|(A^k)_{jm}|}{k!} &\leq 1 + M + \frac{nM^2}{2!} + \dots + \frac{n^{K-1}M^K}{K!} \leq \\ &\leq 1 + nM + \frac{(nM)^2}{2!} + \dots + \frac{(nM)^K}{K!} \leq \exp(nM). \end{aligned}$$

Из ограниченности частных сумм положительного ряда следует, как известно, его сходимость. Таким образом, все ряды (7.6.2) сходятся абсолютно.

Определение. Сумма ряда (7.6.1), т.е. матрица, элементы которой – суммы числовых рядов (7.6.2), называется *экспонентой матрицы* A и обозначается $\exp(A)$.

Пример. Пусть $A = \begin{bmatrix} 0 & x \\ -x & 0 \end{bmatrix}$. Тогда

$$A^2 = \begin{bmatrix} -x^2 & 0 \\ 0 & -x^2 \end{bmatrix}, \quad A^3 = \begin{bmatrix} 0 & -x^3 \\ x^3 & 0 \end{bmatrix}, \quad A^4 = \begin{bmatrix} x^4 & 0 \\ 0 & x^4 \end{bmatrix}, \quad A^5 = \begin{bmatrix} 0 & x^5 \\ -x^5 & 0 \end{bmatrix},$$

и т.д. Таким образом,

$$\begin{aligned} \exp(A) &= \begin{bmatrix} 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots & x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots \\ -\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots\right) & 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots \end{bmatrix} = \\ &= \begin{bmatrix} \cos(x) & \sin(x) \\ -\sin(x) & \cos(x) \end{bmatrix}. \end{aligned}$$

Отметим некоторые свойства матричной экспоненты.

1. $\exp(Ax) \cdot \exp(Ay) = \exp(A(x + y))$ ($x, y \in \mathbb{C}$, A – квадратная матрица).

Действительно, перемножая ряды

$$\exp(Ax) = I + Ax + \frac{A^2 x^2}{2!} + \frac{A^3 x^3}{3!} + \dots, \quad \exp(y) = I + Ay + \frac{A^2 y^2}{2!} + \frac{A^3 y^3}{3!} + \dots,$$

получим

$$\begin{aligned} \exp(Ax) \cdot \exp(Ay) &= \\ &= I + A(x + y) + \frac{A^2}{2!}(x^2 + 2xy + y^2) + \frac{A^3}{3!}(x^3 + 3x^2y + 3xy^2 + y^3) + \dots = \\ &= I + A(x + y) + \frac{A^2(x + y)^2}{2!} + \frac{A^3(x + y)^3}{3!} + \dots = \exp(A(x + y)). \end{aligned}$$

Следствие. $\exp(A) \cdot \exp(-A) = \exp(A + (-A)) = \exp(\emptyset) = I$. Поэтому матрица $\exp(A)$ обратима, и $(\exp(A))^{-1} = \exp(-A)$.

Серьезное предупреждение. Для двух произвольных квадратных матриц одного порядка равенство $\exp(A) \cdot \exp(B) = \exp(A + B)$, вообще говоря, места не имеет. Только если матрицы A и B коммутируют ($AB = BA$), то

$$\exp(A) \cdot \exp(B) = \exp(B) \cdot \exp(A) = \exp(A + B).$$

2. Если $A = \text{diag}[a_1, \dots, a_n]$, то легко проверить, что

$$\exp(A) = \text{diag}[\exp(a_1), \dots, \exp(a_n)].$$

3. Покажем, что если матрицы A и B подобны ($A = S^{-1}BS$), то матрицы $\exp(A)$ и $\exp(B)$ подобны с той же матрицей, осуществляющей подобие, т.е. $\exp(A) = S^{-1} \cdot \exp(B) \cdot S$.

Действительно,

$$A^2 = (S^{-1}BS)^2 = S^{-1}BSS^{-1}BS = S^{-1}B^2S.$$

Аналогично, при всех $k \in \mathbb{N}$ $A^k = S^{-1}B^kS$. Поэтому

$$\exp(A) = \sum_{k=0}^{+\infty} \frac{A^k}{k!} = \sum_{k=0}^{+\infty} \frac{S^{-1}B^kS}{k!} = S^{-1} \cdot \sum_{k=0}^{+\infty} \frac{B^k}{k!} \cdot S = S^{-1} \cdot \exp(B) \cdot S.$$

Пример. Непосредственным вычислением можно получить, что

$$\begin{bmatrix} 0 & x \\ -x & 0 \end{bmatrix} = \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix}^{-1} \cdot \begin{bmatrix} ix & 0 \\ 0 & -ix \end{bmatrix} \cdot \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix}; \quad \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix}^{-1} = \frac{1}{2} \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix}.$$

Поэтому

$$\begin{aligned} \exp\left(\begin{bmatrix} 0 & x \\ -x & 0 \end{bmatrix}\right) &= \frac{1}{2} \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix} \cdot \begin{bmatrix} \exp(ix) & 0 \\ 0 & \exp(-ix) \end{bmatrix} \cdot \begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix} = \\ &= \begin{bmatrix} \cos(x) & \sin(x) \\ -\sin(x) & \cos(x) \end{bmatrix}. \end{aligned}$$

Замечание. Аналогично можно определить другие *аналитические* функции, заданные на множестве квадратных матриц. Некоторые из них реализованы в средах конечного пользователя.

Глава 8. ПРОИЗВОДНАЯ

8.1. Определение производной

Напомним одну известную из школьного курса задачу, приводящую к понятию производной. На графике функции f взяты две точки — фиксированная $(a, f(a))$ и переменная $(x, f(x))$ (рис.8.1). Через эти точки проведена прямая (секущая)

$$y - f(a) = k_c \cdot (x - a).$$

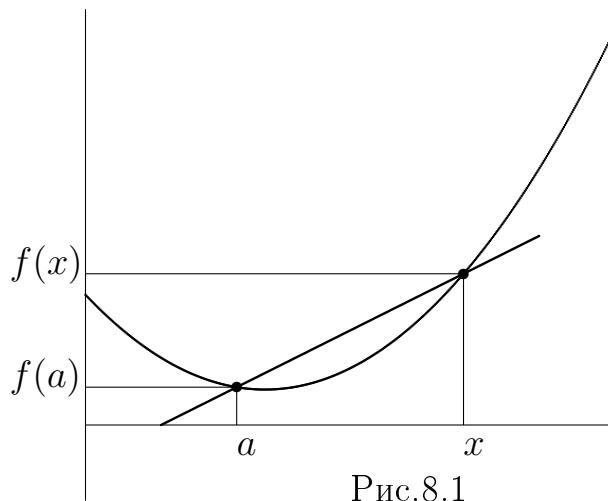


Рис.8.1

Угловой коэффициент этой секущей

$$k_c = \frac{f(x) - f(a)}{x - a}.$$

Что произойдет, если переменная точка совпадет с фиксированной?
Пусть, например, $f(x) = x^2$. Тогда

$$k_c = \frac{x^2 - a^2}{x - a} = x + a$$

(если рациональная дробь сократима, то ее необходимо сократить!).
Поэтому

$$\lim_{x \rightarrow a} \frac{x^2 - a^2}{x - a} = x + a \Big|_{x=a} = 2a.$$

Таким образом, при совпадении переменной точки (x, x^2) с фиксированной (a, a^2) секущая превратится в прямую, проходящую через точку (a, a^2) и имеющую угловой коэффициент $k = 2a$:

$$y - a^2 = 2a \cdot (x - a).$$

Как известно, эта прямая именуется *касательной* к графику функции $y = x^2$ в точке (a, a^2) .

Введем теперь

Определение. Пусть задана функция $f : \mathbb{C} \rightarrow \mathbb{C}$. Зафиксируем точку $a \in \mathbb{C}$, точку $z \in \mathbb{C}$ сделаем переменной и рассмотрим *разностное отношение*

$$\frac{f(z) - f(a)}{z - a}.$$

Значение этого разностного отношения при $z = a$ (если оно определено в этой точке) называется *производной* функции f в точке a и обозначается $f'(a)$.

Итак,

$$f'(a) = \frac{f(z) - f(a)}{z - a} \Big|_{z=a} = \lim_{z \rightarrow a} \frac{f(z) - f(a)}{z - a}.$$

Пример. $f(z) = z^n$ ($n \in \mathbb{N}$).

$$\begin{aligned} f'(a) &= \lim_{z \rightarrow a} \frac{z^n - a^n}{z - a} = \lim_{z \rightarrow a} \frac{(z - a) \cdot (z^{n-1} + z^{n-2}a + \dots + za^{n-2} + a^{n-1})}{z - a} = \\ &= (z^{n-1} + z^{n-2}a + \dots + za^{n-2} + a^{n-1}) \Big|_{z=a} = na^{n-1}. \end{aligned}$$

В рассмотренном примере точка a , в которой вычисляется производная функции $f(x) = x^n$, произвольна. Это дает основание рассматривать новую функцию, значение которой в каждой точке (число) есть производная функции f в этой точке. Эту новую функцию называют *производной функцией* от функции f и обозначают f' (читается "эф штрих"). Итак,

$$f(z) = z^n \implies f'(z) = nz^{n-1}.$$

Далее можно ввести понятие *второй производной функции* от функции f

$$f'' = (f')'; \quad f''(a) = \lim_{z \rightarrow a} \frac{f'(z) - f'(a)}{z - a}.$$

Аналогично вводится (по индукции) понятие *производной функции порядка n* от функции f .

$$f^{(n)} = (f^{(n-1)})'; \quad f^{(n)}(a) = \lim_{z \rightarrow a} \frac{f^{(n-1)}(z) - f^{(n-1)}(a)}{z - a}.$$

Пример. $f(z) = z^4$. Последовательно вычисляем

$$f'(z) = 4 \cdot z^3; \quad f''(z) = 3 \cdot 4 \cdot z^2; \\ f^{(3)}(z) = 2 \cdot 3 \cdot 4 \cdot z; \quad f^{(4)}(z) = 1 \cdot 2 \cdot 3 \cdot 4 = 4! = \text{const.}$$

Прежде, чем вычислять пятую производную функцию, покажем, что в силу наших определений производная функция от функции-константы равна нулю тождественно.

Действительно, пусть $f(z) \equiv p \in \mathbb{C}$. Вычислим производную этой функции в точке $a \in \mathbb{C}$:

$$f'(a) = \lim_{z \rightarrow a} \frac{f(z) - f(a)}{z - a} = \lim_{z \rightarrow a} \frac{p - p}{z - a} = \frac{0}{z - a} \Big|_{z=a}.$$

Дробь $\frac{0}{z - a}$ равна нулю во всех точках, кроме $z = a$, где она не определена. В силу нашего Важного соглашения она доопределяется в этой точке *по непрерывности*, т.е. нулем. Поэтому

$$f'(a) = \lim_{z \rightarrow a} \frac{0}{z - a} = 0.$$

Итак, $f^{(n)}(z) \equiv 0$ при $n > 4$.

8.2. Техника дифференцирования

В нашем распоряжении есть следующие способы "конструирования функций".

1. Образование линейных комбинаций: заданы (на одном и том же множестве $Z \subset \mathbb{C}$) функции f_1, \dots, f_n . Новая функция $f = \sum_{k=1}^n \alpha_k f_k$ ($\alpha_1, \dots, \alpha_k$ – заданные числа) строится по правилу

$$f(z) = \sum_{k=1}^n \alpha_k f_k(z) \quad \text{для всех } z \in Z.$$

2. Степенной ряд.

$$f(z) = \sum_{k=0}^{+\infty} a_k z^k \quad (\text{функция задана внутри круга сходимости ряда}).$$

3. Умножение: заданы (на одном и том же множестве $Z \subset \mathbb{C}$) функции f_1 и f_2 . Новая функция $f = f_1 \cdot f_2$ строится по правилу

$$f(z) = f_1(z) \cdot f_2(z) \quad \text{для всех } z \in Z.$$

4. Деление: заданы (на одном и том же множестве $Z \subset \mathbb{C}$) функции f_1 и f_2 ; f_2 не обращается в нуль. Новая функция $f = f_1/f_2$ строится по правилу

$$f(z) = f_1(z)/f_2(z) \quad \text{для всех } z \in Z.$$

5. Композиция функций (см. п.1.5).

6. Обратная функция (см. п.7.4).

Для вычисления производных функций (*дифференцирования функций*), построенных с помощью описанных выше приемов, необходимо:

1) знать таблицу производных функций для некоторого набора "простейших" функций;

2) уметь находить производные функции для аналитической функции, для линейной комбинации, произведения, частного, композиции, обратной функции.

Сводку таких правил и называют обычно "техникой дифференцирования".

1. Если существуют

$$f'_1(a) = \lim_{z \rightarrow a} \frac{f_1(z) - f_1(a)}{z - a}, \dots, f'_n(a) = \lim_{z \rightarrow a} \frac{f_n(z) - f_n(a)}{z - a}$$

и $f = \sum_{k=1}^n \alpha_k f_k$, то

$$\begin{aligned} f'(a) &= \lim_{z \rightarrow a} \frac{f(z) - f(a)}{z - a} = \lim_{z \rightarrow a} \left(\sum_{k=1}^n \alpha_k \frac{f_k(z) - f_k(a)}{z - a} \right) = \\ &= \sum_{k=1}^n \alpha_k \lim_{z \rightarrow a} \frac{f_k(z) - f_k(a)}{z - a} = \sum_{k=1}^n \alpha_k f'_k(a). \end{aligned}$$

Производная линейной комбинации функций равна линейной комбинации производных этих функций (с теми же коэффициентами).

Пример: производная полинома. Если

$$f(z) = a_0 + a_1 z + a_2 z^2 + \dots + a_n z^n = \sum_{k=0}^n a_k z^k,$$

то

$$f'(z) = a_1 + 2a_2 z + \dots + n a_n z^{n-1} = \sum_{k=1}^n k a_k z^{k-1},$$

$$f''(z) = 2a_2 + \dots + n(n-1)a_n z^{n-2} = \sum_{k=2}^n k(k-1)a_k z^{k-2},$$

...

$$f^{(n)}(z) = n!,$$

$$f^{(m)}(z) \equiv 0 \text{ при } m > n.$$

Обратите внимание на *нижний* предел суммирования!

2. Можно показать, что аналитическая функция внутри круга сходимости определяющего ее степенного ряда имеет производные всех порядков, и они находятся "почленным дифференцированием т.е.

$$\left(\sum_{k=0}^{+\infty} a_k z^k \right)' = \sum_{k=1}^{+\infty} k a_k z^{k-1},$$

$$\left(\sum_{k=0}^{+\infty} a_k z^k \right)'' = \sum_{k=2}^{+\infty} k(k-1)a_k z^{k-2}$$

и т.д. (Сравните с производной полинома!).

Пример.

$$\exp'(z) = \left(\sum_{k=0}^{+\infty} \frac{z^k}{k!} \right)' = \sum_{k=1}^{+\infty} \frac{k z^{k-1}}{k!} = \sum_{k=1}^{+\infty} \frac{z^{k-1}}{(k-1)!} = \sum_{k=0}^{+\infty} \frac{z^k}{k!} = \exp(z).$$

$$\exp'(z) \equiv \exp(z).$$

Аналогичным вычислением получаем (проверьте это!), что

$$\sin'(z) \equiv \cos(z); \quad \cos'(z) \equiv -\sin(z).$$

3. Если существуют $f'_1(a)$, $f'_2(a)$, и $f = f_1 \cdot f_2$, то

$$\begin{aligned}
f'(a) &= \lim_{z \rightarrow a} \frac{f(z) - f(a)}{z - a} = \lim_{z \rightarrow a} \frac{f_1(z) \cdot f_2(z) - f_1(a) \cdot f_2(a)}{z - a} = \\
&= \lim_{z \rightarrow a} \frac{f_1(z) \cdot f_2(z) - f_1(a) \cdot f_2(z) + f_1(a) \cdot f_2(z) - f_1(a) \cdot f_2(a)}{z - a} = \\
&= \lim_{z \rightarrow a} \left(\frac{f_1(z) - f_1(a)}{z - a} \cdot f_2(z) \right) + \lim_{z \rightarrow a} \left(f_1(a) \cdot \frac{f_2(z) - f_2(a)}{z - a} \right) = \\
&= f'_1(a) \cdot f_2(a) + f_1(a) \cdot f'_2(a).
\end{aligned}$$

$$(f_1 \cdot f_2)' = f'_1 \cdot f_2 + f_1 \cdot f'_2.$$

4. Если существуют $f'_1(a)$, $f'_2(a)$, $f_2(a) \neq 0$ и $f = f_1/f_2$, то

$$\begin{aligned}
f'(a) &= \lim_{z \rightarrow a} \frac{f(z) - f(a)}{z - a} = \lim_{z \rightarrow a} \frac{f_1(z)/f_2(z) - f_1(a)/f_2(a)}{z - a} = \\
&= \lim_{z \rightarrow a} \frac{f_1(z) \cdot f_2(a) - f_1(a) \cdot f_2(z)}{f_2(z) \cdot f_2(a) \cdot (z - a)} = \\
&= \lim_{z \rightarrow a} \frac{\frac{f_1(z) - f_1(a)}{z - a} \cdot f_2(a) - f_1(a) \cdot \frac{f_2(z) - f_2(a)}{z - a}}{f_2(z) \cdot f_2(a)} = \\
&= \frac{f'_1(a) \cdot f_2(a) - f_1(a) \cdot f'_2(a)}{f_2^2(a)}.
\end{aligned}$$

$$(f_1/f_2)' = \frac{f'_1 \cdot f_2 - f_1 \cdot f'_2}{f_2^2}.$$

Пример.

$$\begin{aligned}
tg'(z) &= \left(\frac{\sin}{\cos} \right)'(z) = \frac{\sin'(z) \cdot \cos(z) - \sin(z) \cdot \cos'(z)}{\cos^2(z)} = \\
&= \frac{\cos^2(z) + \sin^2(z)}{\cos^2(z)} = 1 + \operatorname{tg}^2(z).
\end{aligned}$$

$$tg'(z) \equiv 1 + \operatorname{tg}^2(z) \equiv \frac{1}{\cos^2(z)}.$$

Аналогичным вычислением получаем (проверьте это!), что

$$ctg'(z) \equiv - (1 + ctg^2(z)) \equiv -\frac{1}{\sin^2(z)}.$$

5. Производная композиции (рис.8.2).

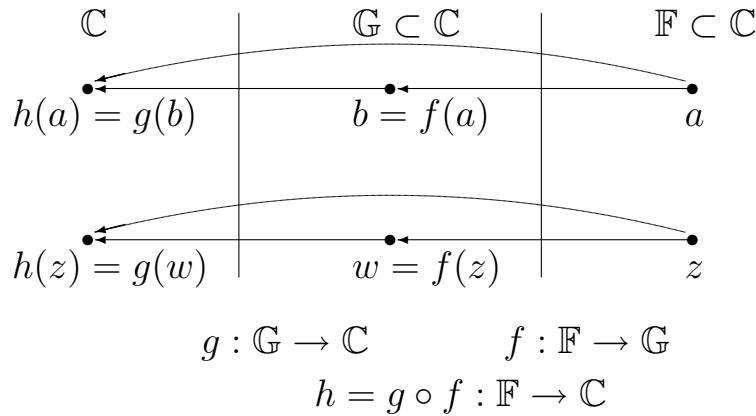


Рис.8.2

$$h'(a) = \lim_{z=a} \frac{h(z) - h(a)}{z - a} = \lim_{z=a} \frac{g(w) - g(b)}{w - b} \cdot \frac{f(z) - f(a)}{z - a}.$$

Поскольку из $z = a$ следует $w = b$, получаем

$$h'(a) = \lim_{w=b} \frac{g(w) - g(b)}{w - b} \cdot \lim_{z=a} \frac{f(z) - f(a)}{z - a} = g'(b) \cdot f'(a).$$

Производная композиции равна *произведению* производных функций, составляющих эту композицию.

Конечно, каждая из этих производных вычисляется в "своей" точке!

Пример. $w = f(z) = \sin(z)$, $g(w) = w^2$; $h(z) = (g \circ f)(z) = \sin^2(z)$.

$$h'(z) = g'(w) \cdot f'(z) = 2w \cdot \cos(z) = 2\sin(z) \cdot \cos(z).$$

6. Производная обратной функции (рис.8.3).

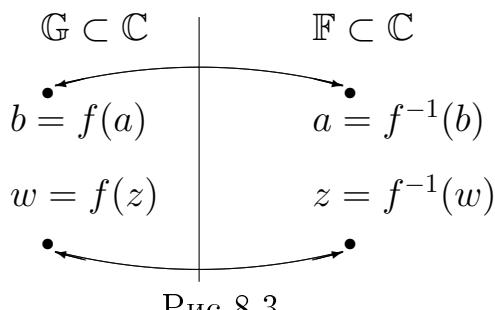


Рис.8.3

$$(f^{-1})'(b) = \lim_{w \rightarrow b} \frac{f^{-1}(w) - f^{-1}(b)}{w - b} = \lim_{z \rightarrow a} \frac{z - a}{f(z) - f(a)} =$$

$$= \lim_{z \rightarrow a} \frac{1}{\frac{f(z) - f(a)}{z - a}} = \frac{1}{f'(a)}.$$

Здесь предполагается, что $f'(a) \neq 0$.

$$(f^{-1})'(z) = \frac{1}{f'(f^{-1}(z))}.$$

Примеры. 1. $\ln = \exp^{-1}$.

$$w = \ln(z); \quad z = \exp(w); \quad \ln'(z) = \frac{1}{\exp'(w)} = \frac{1}{\exp(w)} = \frac{1}{z}.$$

$$\ln'(z) = \frac{1}{z}.$$

2. $\arctg = \tg^{-1}$.

$$w = \arctg(z); \quad z = \tg(w); \quad \arctg'(z) = \frac{1}{\tg'(w)} = \frac{1}{1 + \tg^2(w)} = \frac{1}{1 + z^2}.$$

$$\arctg'(z) = \frac{1}{1 + z^2}.$$

3. Найдем теперь производную степенной функции z^α , $z \neq 0$.

$$(z^\alpha)' = (\exp(\alpha \cdot \ln(z)))' = \exp(\alpha \cdot \ln(z)) \cdot \alpha \cdot \frac{1}{z} = \alpha \cdot z^{\alpha-1}.$$

$$(z^\alpha)' = \alpha \cdot z^{\alpha-1}.$$

8.3. Ряд Тейлора

Пусть аналитическая функция f определена в некотором круге с центром в точке p :

$$f(z) = a_0 + a_1(z - p) + a_2(z - p)^2 + a_3(z - p)^3 + \dots + a_n(z - p)^n + \dots$$

Полагая $z = p$, получим $f(p) = a_0$.

Продифференцируем f :

$$f'(z) = a_1 + 2a_2 \cdot (z - p) + 3a_3 \cdot (z - p)^2 + \dots + na_n \cdot (z - p)^{n-1} + \dots$$

Полагая $z = p$, получим $f'(p) = a_1$.

Далее,

$$f''(z) = 2 \cdot 1 \cdot a_2 + 3 \cdot 2a_3 \cdot (z - p) + \dots + n \cdot (n - 1)a_n \cdot (z - p)^{n-2} + \dots$$

Полагая $z = p$, получим $f''(p) = 2 \cdot 1 \cdot a_2$.

Нетрудно увидеть, что $f^{(n)}(p) = n!a_n$, откуда

$$a_n = \frac{f^{(n)}(p)}{n!}.$$

Эта формула верна и при $n = 0$ (полагают по определению $f^{(0)} = f$).

Мы получили представление аналитической функции в виде

$$f(z) = \sum_{n=0}^{+\infty} \frac{f^{(n)}(p)}{n!} (z - p)^n. \quad (8.3.1)$$

Ряд, стоящий в правой части формулы (8.3.1) называется *рядом Тейлора*²⁶ аналитической функции f в окрестности точки p .

Примеры. 1. Для любого натурального n $\exp^{(n)}(p) = \exp(p)$.
Поэтому

$$\exp(z) = \sum_{n=0}^{+\infty} \frac{\exp(p)}{n!} (z - p)^n = \exp(p) \cdot \sum_{n=0}^{+\infty} \frac{(z - p)^n}{n!}.$$

Заметим, что мы получили известное тождество

$$\exp(z) \equiv \exp(p) \cdot \exp(z - p).$$

2. $f(z) = \ln(1 + z)$, $f(0) = 0$. Последовательно дифференцируя, получаем

$$\begin{aligned} f'(z) &= (1 + z)^{-1}, & f'(0) &= 1; \\ f''(z) &= (-1) \cdot (1 + z)^{-2}, & f''(0) &= -1; \\ f'''(z) &= (-1) \cdot (-2) \cdot (1 + z)^{-3}, & f'''(0) &= 1 \cdot 2; \\ &\dots & &\dots \\ f^{(n)}(z) &= (-1)^{n-1}(n-1)! \cdot (1 + z)^{-n}, & f^{(n)}(0) &= (-1)^{n-1}(n-1)!. \end{aligned}$$

²⁶Брук ТЕЙЛОР (B. Taylor, 1685-1731) – английский математик, ученик секретарь Лондонского Королевского общества.

Итак, в некоторой окрестности нуля

$$\ln(1+z) = \sum_{n=1}^{+\infty} \frac{(-1)^{n-1}(n-1)!}{n!} z^n = \sum_{n=1}^{+\infty} \frac{(-1)^{n-1}}{n} z^n.$$

Найдем радиус сходимости этого ряда, используя признак Д'Аламбера:

$$D(z) = \lim \frac{|z|^k (k-1)}{k|z|^{k-1}} = |z| \cdot \lim \frac{k-1}{k} = |z| \implies R = 1.$$

Хотя функция $\ln(1+z)$ определена на всей комплексной плоскости, кроме точки $z = -1$, ее ряд Тейлора (в окрестности нуля) сходится только при $|z| < 1$. Это объясняется тем, что точка $z = -1$ "не дает" круга сходимости стать больше.

3. Пусть f – полином степени n : $f(z) = a_0 + a_1 z + \dots + a_n z^n$, $a_n \neq 0$.

$$f^{(n)}(z) = n! \cdot a_n \implies f^{(k)}(z) \equiv 0 \quad k > n.$$

Поэтому ряд (8.3.1) превращается в конечную сумму

$$f(z) = \sum_{k=0}^n \frac{f^{(k)}(p)}{k!} (z-p)^k.$$

Это тождество относительно z и p называют *формулой Тейлора для полинома* – мы получили еще одно представление полинома (см. формулу (3.1.2)).

Глава 9. ФУНКЦИИ из \mathbb{R} в \mathbb{R}

9.1. Примеры

В этой главе будут рассмотрены функции, заданные на \mathbb{R} или на части \mathbb{R} (например, на промежутке) и принимающие значения в \mathbb{R} . Их часто называют "вещественные функции вещественной переменной". Такие функции получаются прежде всего из рассмотренных нами ранее аналитических функций путем сужения их на ту часть \mathbb{R} , где они принимают вещественные значения. При этом окажется, что логарифм, например, определен только на положительной полуоси, а синус и косинус, "как в школе ограничены по модулю единицей".

Установим некоторые полезные свойства экспоненты и логарифма на вещественной оси.

1. $\exp : \mathbb{R} \rightarrow \mathbb{R}$;

$$\exp(x) = 1 + \frac{x}{1!} + \dots + \frac{x^n}{n!} + \frac{x^{n+1}}{(n+1)!} + \dots \quad (9.1.1)$$

Из формулы (9.1.1) видно, что при $x > 0$ $\exp(x) > x$ и, следовательно, экспонента на вещественной оси неограниченно возрастает. Этот факт записывают так:

$$\lim_{x \rightarrow +\infty} \exp(x) = +\infty.$$

Приведем точный смысл этого выражения: для любого положительного числа E можно указать такое число x_E , что при $x > x_E$ $\exp(x) > E$.

Отсюда вытекает, что

$$\lim_{x \rightarrow -\infty} \exp(x) = \lim_{x \rightarrow +\infty} \exp(-x) = \lim_{x \rightarrow +\infty} \frac{1}{\exp(x)} = \frac{1}{+\infty} = 0$$

(для любого положительного числа ε можно указать такое число x_ε , что при $x < x_\varepsilon$ $\exp(x) < \varepsilon$).

Далее, из (9.1.1) следует, что при $x > 0$

$$\exp(x) > \frac{x^{n+1}}{(n+1)!} \quad \text{или} \quad \frac{x^n}{\exp(x)} < \frac{(n+1)!}{x}.$$

Зафиксируем в последнем неравенстве n . Тогда получим

$$\lim_{x \rightarrow +\infty} \frac{x^n}{\exp(x)} = 0.$$

Экспонента растет на $+\infty$ быстрее,
чем любая положительная степень ее операнда.

2. $\ln :]0, +\infty[\rightarrow \mathbb{R}$.

Из возрастания экспоненты на вещественной оси (п.7.3, свойство 5) следует

$$x_1 > x_2 \iff \exp(x_1) > \exp(x_2).$$

Обозначив $y_1 = \exp(x_1)$, $y_2 = \exp(x_2)$, получим

$$y_1 > y_2 \iff \ln(y_1) > \ln(y_2),$$

т.е. логарифм возрастает на положительной вещественной полуоси.

Из $\lim_{x \rightarrow +\infty} \exp(x) = +\infty$ следует, что $\lim_{y \rightarrow +\infty} \ln(y) = +\infty$ – логарифм возрастает неограниченно.

Из $\lim_{x \rightarrow +\infty} \frac{x}{\exp(x)} = 0$ следует, что $\lim_{y \rightarrow +\infty} \frac{\ln(y)}{y} = 0$ и, наконец, при любом $\alpha > 0$ имеем

$$\lim_{x \rightarrow +\infty} \frac{\ln(x)}{x^\alpha} = \frac{1}{\alpha} \cdot \lim_{x \rightarrow +\infty} \frac{\ln(x^\alpha)}{x^\alpha} = \frac{1}{\alpha} \cdot \lim_{y \rightarrow +\infty} \frac{\ln(y)}{y} = 0.$$

Логарифм растет на $+\infty$ медленнее,
чем любая положительная степень его операнда.

9.2. Оценивание вещественных корней вещественных функций

Сформулируем без доказательства два важных свойства функций $\mathbb{R} \rightarrow \mathbb{R}$.

Теорема Вейерштрасса²⁷.

1. Множество значений непрерывной *на сегменте*, вещественной функции ограничено.

2. Среди значений непрерывной *на сегменте*, вещественной функции есть наибольшее и наименьшее (если Y – множество значений этой функции, то $\sup(Y) \in Y$, $\inf(Y) \in Y$).

²⁷Карл Теодор Вильгельм ВЕЙЕРШТРАСС (K.T.W. Weierstraß, 1815-1897) – немецкий математик, профессор Берлинского университета. Его учениками были многие известные математики из разных стран, в том числе С.В. Ковалевская.

Замечание. Обратите внимание на то, что здесь существенна не только *непрерывность* функции, но и ее задание *на сегменте*. Например, функция $f(x) = 1/x$ непрерывна на полуинтервале $]0, 1]$, но не ограничена на нем, хотя ограничена на любом сегменте, целиком лежащем в $]0, 1]$. Функция, $f(x) = x^2$ непрерывна и ограничена на $]0, 1]$, но среди ее значений нет наименьшего. Попробуйте придумать пример функции, непрерывной и ограниченной на $[0, 1]$, но не достигающей ни верхней, ни нижней грани своего множества значений.

Теорема Коши. Если вещественная функция непрерывна на *сегменте*, и множество ее значений Y содержит два различных числа $y_1 < y_2$ (т.е эта функция – не константа), то Y содержит и все "промежуточные" числа: $[y_1, y_2] \subset Y$.

Замечания. 1. Эта теорема кажется совершенно очевидной: она утверждает, что непрерывная функция не может ни "перепрыгнуть" через какое-нибудь число, ни "обогнать" его. На самом же деле этот факт – фундаментальное и довольно тонкое свойство множества \mathbb{R} .

Заметим также, что на комплексной плоскости множество значений непрерывной функции может иметь "дыры". Например, экспонента принимает все значения из \mathbb{C} , кроме нуля!

2. Теоремы Вейерштрасса и Коши часто формулируют в виде одного короткого утверждения: *непрерывная функция, действующая из \mathbb{R} в \mathbb{R} , отображает сегмент на сегмент*.

Отметим одно важное приложение теоремы Коши. Пусть f непрерывна на сегменте $[a, b]$, и $f(a) \cdot f(b) < 0$. Тогда множество значений f содержит нуль, т.е. найдется хотя бы одна такая точка $c \in]a, b[$, что $f(c) = 0$. Эту точку называют корнем уравнения $f(x) = 0$ или нулем функции f .

Поскольку условие $f(a) \cdot f(b) < 0$ гарантирует (при непрерывности f), что интервал $]a, b[$ накрывает хотя бы один корень уравнения $f(x) = 0$, назовем этот интервал *оценкой* корня. Качеством этой оценки естественно считать длину интервала.

Процесс построения оценки корня уравнения с заданным качеством (процесс "решения" уравнения) состоит из двух этапов:

- 1) поиск какой-нибудь оценки (поиск интервала, гарантированно накрывающего корень) – этот этап, вообще говоря, не алгоритмизируем;
- 2) уточнение имеющейся оценки.

Рассмотрим один шаг известного алгоритма уточнения оценки корня вещественной непрерывной функции – алгоритма *бисекции* ("половинного деления").

Пусть f непрерывна на $[a, b]$ и $f(a) \cdot f(b) < 0$.

Возьмем середину сегмента – точку $x = \frac{a+b}{2}$ и вычислим $f(x)$.

Возможны три случая:

- 1) $f(x) = 0$;
- 2) $f(x) \cdot f(a) < 0$;
- 3) $f(x) \cdot f(b) < 0$.

В случае (1) мы получили не оценку корня, а его значение.

В случае (2) оценкой корня становится интервал $]a, x[$, в случае (3) – интервал $]x, b[$.

Таким образом, мы либо нашли корень, либо вдвое сократили длину интервала-оценки.

Повторяя бисекцию, теоретически можно получить оценку любого наперед заданного качества. Практически же предел уточнению оценки кладет "машинная арифметика" (см. Приложение).

Серьезное предупреждение. При машинном счете находить середину сегмента по формуле $x = \frac{a+b}{2}$ нельзя: из-за вычислительных погрешностей точка может оказаться *вне сегмента!* Правильный результат дает формула $x = a + \frac{b-a}{2}$. Этот несложный пример подтверждает наш совет: пользователь не должен пытаться самостоятельно писать программы, реализующие вычислительные алгоритмы. Его удел – использование готовых библиотек.

9.3. Кусочно заданные функции

Уже давно нашли широкое применение кусочно заданные функции, простейшим примером которых является *кусочно полиномиальная функция*. Задается она так.

1. Строится *сетка*, т.е. упорядоченный по возрастанию конечный набор попарно различных вещественных чисел

$$x_0 < x_1 < \dots < x_n.$$

2. На каждом интервале $J_k =]x_{k-1}, x_k[, k = 1, \dots, n$, задается полином $f_k : J_k \rightarrow \mathbb{R}$.

3. Задаются значения функции в узлах сетки $f(x_0), \dots, f(x_n)$.

Примеры. 1. $sign(x)$ (рис.9.1)²⁸.

$$sign(x) = \begin{cases} -1 & \text{если } x < 0; \\ 0 & \text{если } x = 0; \\ 1 & \text{если } x > 0. \end{cases}$$

Эта функция есть простейшая модель зависимости силы *сухого трения* от скорости движения: если скорость равна нулю – трение отсутствует, если скорость отлична от нуля, сила трения определяется только направлением движения.

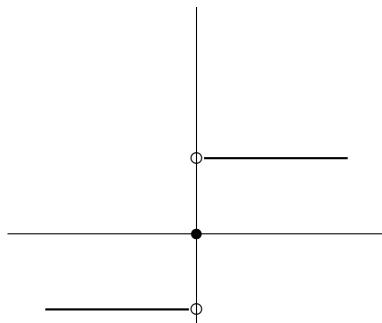


Рис.9.1

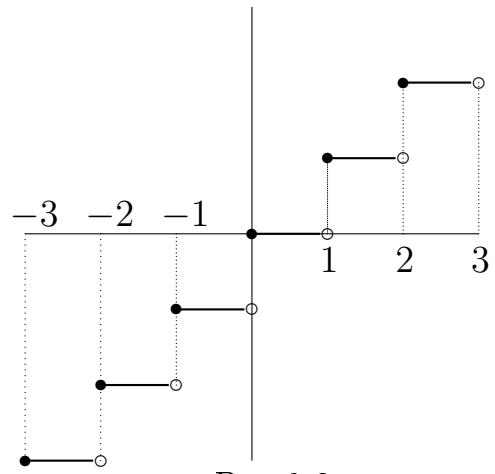


Рис.9.2

2. $entier(x)$ (см. п.5.1; на рис.9.2 приведен фрагмент ее графика).

Примеры **1** и **2** – это примеры *кусочно постоянных* функций.

3. $x - entier(x)$ – "дробная часть числа" (на рис.9.3 приведен фрагмент ее графика – на каждом интервале функция задана полиномом первой степени).

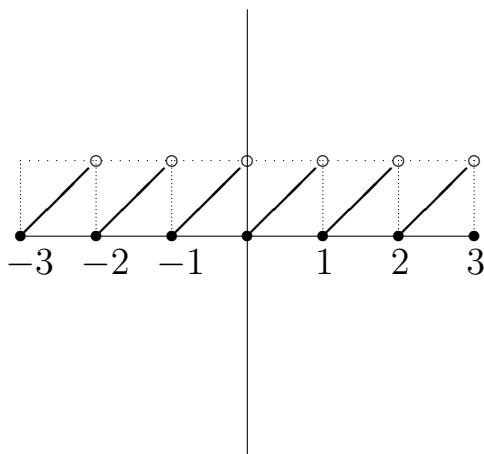


Рис.9.3

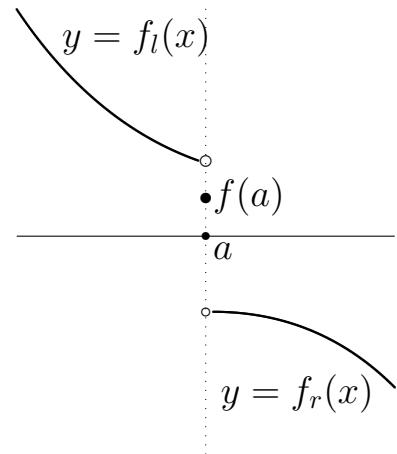


Рис.9.4

²⁸sign (англ.) – знак.

Рассмотрим теперь поведение кусочно полиномиальной функции в окрестности узла сетки (рис.9.4). Пусть этот узел $x = a$, и

$$f(x) = \begin{cases} f_l(x) & \text{при } x < a, \\ f_r(x) & \text{при } x > a, \end{cases}$$

где f_l и f_r – полиномы.

Воспользуемся тем, что полиномы f_l и f_r определены на всей оси, и вычислим $f_l(a)$ и $f_r(a)$.

Таким образом, в точке a определены:

- 1) $f_l(a)$ – вычисленное в точке a значение полинома f_l , задающего функцию f левее этой точки,
- 2) $f(a)$ – значение функции f в точке a ,
- 3) $f_r(a)$ – вычисленное в точке a значение полинома f_r , задающего функцию f правее этой точки.

Число $f_l(a)$ называют *левым пределом функции f в точке a* , число $f_r(a)$ – *правым пределом функции f в точке a* . Пишут

$$\lim_{x=a^-} f(x) = f_l(a), \quad \text{или} \quad f(a-0) = f_l(a), \quad \text{или} \quad f(a-) = f_l(a).$$

$$\lim_{x=a^+} f(x) = f_r(a), \quad \text{или} \quad f(a+0) = f_r(a), \quad \text{или} \quad f(a+) = f_r(a).$$

Число $f(a+) - f(a-)$ называется *скачком* функции f в точке a .

Замечания. 1. Если скачок функции в точке a равен нулю, т.е. левый и правый пределы равны между собой, мы будем считать эту функцию в точке a *непрерывной*, полагая $f(a) = f(a+) = f(a-)$. Тем самым мы исключаем из рассмотрения так называемые "устранимые разрывы считая, что если разрыв можно устраниТЬ, то его следует устранить. Эта точка зрения уже разъяснялась в п.4.5.

2. Если скачок функции в точке отличен от нуля, то значение функции в этой точке может совпадать с одним из ее односторонних пределов.

Если $f(a) = f(a+)$, то говорят, что функция f *непрерывна в точке a справа*. Если же $f(a) = f(a-)$, то говорят, что функция f *непрерывна в точке a слева*.

Очевидно, что непрерывность функции в точке равносильна ее непрерывности в этой точке справа и слева.

3. Если f определена в точке a , имеет в этой точке конечные односторонние пределы, и ее скачок в этой точке отличен от нуля, то

а называют точкой разрыва *первого рода* для функции f . Функция, все точки разрыва которой – первого рода, причем количество их на любой конечной части ее области определения конечно, называется *кусочно непрерывной*.

Примеры. 1. $\text{sign}(0-) = -1$, $\text{sign}(0+) = +1$, $\text{sign}(0) = 0$ (см. рис.9.1). sign – кусочно непрерывная функция (имеет одну точку разрыва первого рода). Скачок ее в нуле равен $\text{sign}(0+) - \text{sign}(0-) = 2$. Функция sign не является непрерывной в нуле ни слева, ни справа.

2. $\text{entier}(1-) = 0$, $\text{entier}(1) = 1$, $\text{entier}(1+) = 1$ (см. рис.9.2). Поэтому функция entier непрерывна в точке $x = 1$ справа. Точно также она непрерывна справа в любой точке $x \in \mathbb{Z}$. Функция entier также кусочно непрерывна (хотя количество ее точек разрыва первого рода и бесконечно, но на любой *конечной* части оси оно *конечно*).

Все понятия, введенные выше для *кусочно полиномиальных* функций, естественным образом распространяются на любые *кусочно аналитические* функции при условии, что каждая аналитическая функция f_k определена не только на интервале J_k , но и на его концах.

9.4. Полиномиальные сплайны

Кусочно полиномиальные функции принято называть *полиномиальными сплайнами*²⁹. Учитывая все возрастающую роль полиномиальных сплайнов в приложениях, рассмотрим кратко их основные конструкции.

Порядком полиномиального сплайна мы будем называть порядок образующих его полиномов, а полином f_k , задающий сплайн на интервале $J_k =]x_{k-1}, x_k[$, – *k-й порцией* сплайна.

Простейший полиномиальный сплайн – сплайн *первого* порядка – кусочно постоянная функция. Он имеет, вообще говоря, разрывы первого рода в узлах сетки (см. рис.9.1 и рис.9.2).

Пример сплайна *второго* порядка, имеющего разрывы во всех узлах сетки, приведен на рис.9.3. Однако при построении сплайна второго порядка уже можно обеспечить его непрерывность. Такой сплайн задается таблицей его значений в узлах сетки.

x	x_0	...	x_n
$f(x)$	y_0	...	y_n

²⁹spline (англ.) – рейка (приспособление, которое применяли чертежники для проведения гладких кривых через заданные точки).

Запишем его k -ю порцию в виде

$$f_k(x) = a_k(x - x_{k-1}) + b_k.$$

Коэффициенты a_k и b_k определяются из условий непрерывности в узлах:

$$f_k(x_{k-1}) = b_k = y_{k-1}; \quad f_k(x_k) = a_k(x_k - x_{k-1}) + b_k = y_k.$$

Отсюда

$$f_k(x) = y_{k-1} + \frac{y_k - y_{k-1}}{x_k - x_{k-1}} \cdot (x - x_{k-1}).$$

Геометрическая интерпретация непрерывного сплайна второго порядка – ломаная, соединяющая точки $(x_0, y_0), \dots, (x_n, y_n)$. На рис.9.5 изображен сплайн второго порядка, построенный по таблице 1.

Таблица 1

x	0	2	5	10	12
y	3	4	2	3	5

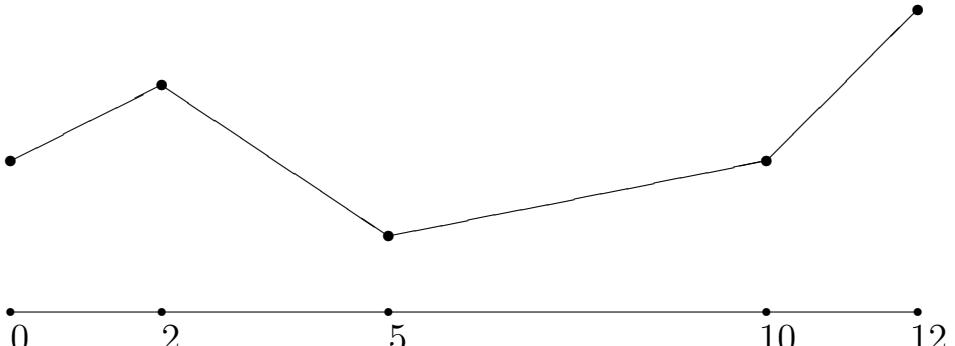


Рис.9.5

Производная непрерывного сплайна второго порядка, вообще говоря, не определена в узлах сетки.

Сплайн *третьего* порядка может не только быть непрерывным, но и обладать непрерывной первой производной. Запишем его k -ю порцию в виде

$$f_k(x) = a_k(x - x_{k-1})^2 + b_k(x - x_{k-1}) + c_k.$$

Фиксация значений порции сплайна на концах интервала даст два уравнения

$$f_k(x_{k-1}) = c_k = y_{k-1}, \tag{9.4.1}$$

$$f_k(x_k) = a_k(x_k - x_{k-1})^2 + b_k(x_k - x_{k-1}) + c_k = y_k. \tag{9.4.2}$$

Таких пар уравнений будет n (по числу порций сплайна). Потребовав непрерывности первой производной во внутренних узлах сетки ($k = 1, \dots, n - 1$), получим еще $n - 1$ уравнение

$$f'_k(x_k) = 2a_k(x_k - x_{k-1}) + b_k = b_{k+1} = f'_{k+1}(x_k). \quad (9.4.3)$$

Для определения $3n$ параметров сплайна ($a_k, b_k, c_k; k = 1, \dots, n$) мы получили $2n + (n - 1) = 3n - 1$ уравнений – система неопределенная!

Подставив (9.4.1) в (9.4.2), получим

$$\begin{cases} a_k(x_k - x_{k-1})^2 + b_k(x_k - x_{k-1}) &= y_k - y_{k-1} \\ 2a_k(x_k - x_{k-1}) + b_k &= b_{k+1} \end{cases}.$$

Переписав эту систему иначе

$$\begin{cases} a_k &= \frac{y_k - y_{k-1}}{(x_k - x_{k-1})^2} - \frac{b_k}{x_k - x_{k-1}}, & (k = 1, \dots, n - 1), \\ b_{k+1} &= 2a_k(x_k - x_{k-1}) + b_k \end{cases}$$

получим рекуррентные формулы для определения (при фиксированных значениях сплайна в узлах сетки) по *заданному произвольно* b_1 остальных параметров сплайна:

$$b_1 \rightarrow a_1 \rightarrow b_2 \rightarrow \dots$$

Произвол в выборе b_1 может быть использован как угодно (как и всякий другой произвол). На рис.9.6 изображены сплайны третьего порядка, построенные по той же таблице 1 при $b_1 = 0.3$ (жирная линия) и $b_1 = 1.2$ (тонкая линия).

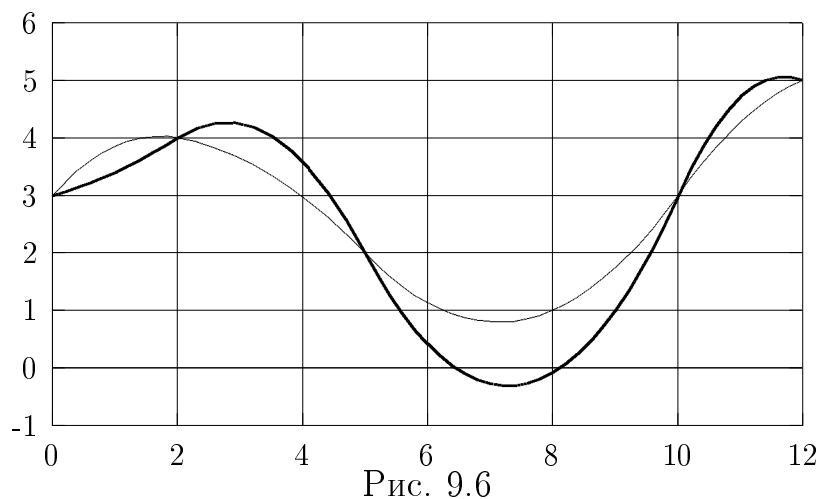


Рис. 9.6

Рассмотрим еще нашедший наибольшее применение полиномиальный сплайн четвертого порядка с двумя непрерывными производными. Его k -я порция имеет вид

$$f_k(x) = a_k(x - x_{k-1})^3 + b_k(x - x_{k-1})^2 + c_k(x - x_{k-1}) + d_k.$$

Подлежит определению $4n$ параметров сплайна ($a_k, b_k, c_k, d_k; k = 1, \dots, n$). Фиксируя значения порции сплайна на концах интервала, получим для каждой порции два уравнения

$$f_k(x_{k-1}) = d_k = y_{k-1}, \quad (9.4.4)$$

$$f_k(x_k) = a_k(x_k - x_{k-1})^3 + b_k(x_k - x_{k-1})^2 + c_k(x_k - x_{k-1}) + d_k = y_k. \quad (9.4.5)$$

Непрерывность первой производной во внутренних узлах сетки ($k = 1, \dots, n-1$) дает $n-1$ уравнение

$$f'_k(x_k) = 3a_k(x_k - x_{k-1})^2 + b_k(x_k - x_{k-1}) + c_k = c_{k+1} = f'_{k+1}(x_k). \quad (9.4.6)$$

И, наконец, непрерывность второй производной в тех же узлах дает еще $n-1$ уравнение

$$f''_k(x_k) = 6a_k(x_k - x_{k-1}) + 2b_k = 2b_{k+1} = f''_{k+1}(x_k). \quad (9.4.7)$$

Всего, таким образом, для определения $4n$ параметров сплайна имеем $2n + 2(n-1) = 4n - 2$ уравнений (2 параметра задаются произвольно). Подставив (9.4.4) в (9.4.5), получим после несложных преобразований систему рекуррентных соотношений

$$\begin{cases} a_k &= \frac{y_k - y_{k-1}}{(x_k - x_{k-1})^3} - \frac{b_k}{(x_k - x_{k-1})^2} - \frac{c_k}{x_k - x_{k-1}} \\ c_{k+1} &= 3a_k(x_k - x_{k+1})^2 + 2b_k(x_k - x_{k-1}) + c_k \\ b_{k+1} &= 3a_k(x_k - x_{k+1}) + b_k \end{cases}, \quad (k = 1, \dots, n-1),$$

которая позволяет (при фиксированных значениях сплайна в узлах сетки) по заданным произвольно c_1, b_1 определить остальные параметры сплайна.

$$c_1, b_1 \rightarrow a_1 \rightarrow c_2, b_2 \rightarrow \dots$$

Отметим, что множество сплайнов порядка ℓ , имеющих m непрерывных производных, образует линейное пространство. Можно показать, что его размерность равна $n \cdot (\ell - m - 1) + m + 1$, где n – количество порций сплайна.

Глава 10. ЛОКАЛЬНОЕ ИССЛЕДОВАНИЕ ГЛАДКИХ ФУНКЦИЙ $\mathbb{R} \rightarrow \mathbb{R}$

10.1. Теорема Ролля

Рассмотрим непрерывную функцию f , принимающую на концах сегмента $[a, b]$ одинаковые значения. Четыре примера таких функций изображены на рис.10.1.

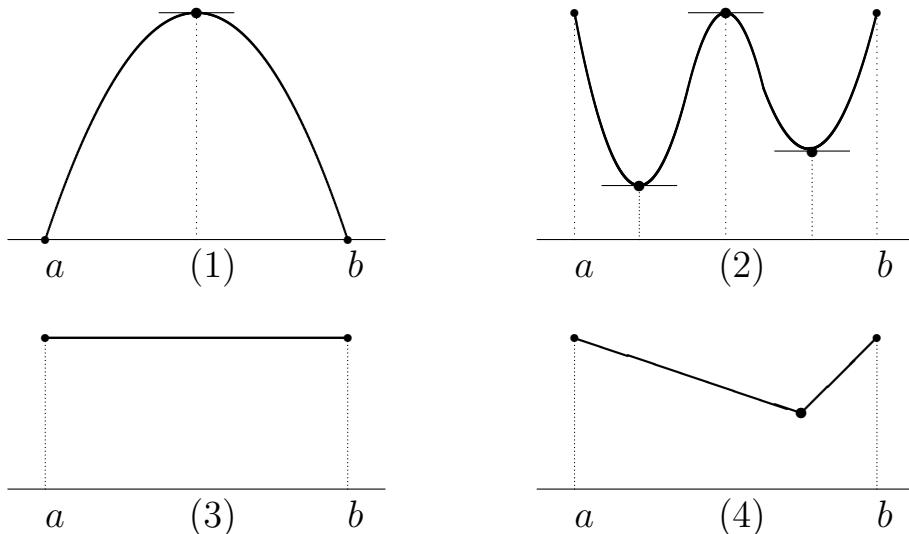


Рис.10.1

Видно, что в случае (1) на графике есть *одна* точка, в которой касательная горизонтальна. В случае (2) таких точек уже *три*. В случае (3) касательная горизонтальна в *каждой* *внутренней* точке графика. А вот в случае (4) на графике *нет* точек с горизонтальной касательной. Этот случай отличается от предыдущих тем, что касательную к графику можно провести не во всякой его точке (в отмеченной точке функция не имеет производной!).

Можно показать, что справедлива

Теорема Ролля³⁰. Пусть функция $f : [a, b] \rightarrow \mathbb{R}$:

- 1) непрерывна,
- 2) имеет производную в каждой точке интервала $]a, b[$,
- 3) $f(a) = f(b)$.

Тогда на $]a, b[$ найдется хотя бы *одна* точка, в которой производная функции f равна нулю.

³⁰Мишель РОЛЛЬ (M. Rolle, 1652-1719) – французский математик, член Парижской АН.

10.2. Формула Тейлора

Пусть f – аналитическая в r -окрестности точки $a \in \mathbb{R}$ функция:

$$f(x) = f(a) + \frac{f'(a)}{1!} \cdot (x - a) + \dots + \frac{f^{(n)}}{n!} \cdot (x - a)^n + \dots \quad (|x - a| < r).$$

При практических вычислениях работают не с рядом, а с его частной суммой, обрывая суммирование на некотором слагаемом и оценивая возникающую при этом погрешность.

Рассмотрим один из наиболее употребительных способов оценки погрешности, возникающей при замене *вещественной* функции частной суммой ее ряда Тейлора.

Пусть сперва частная "сумма" состоит из одного слагаемого:

$$f(x) \sim f(a) \quad (\sim \text{обозначает "заменяется на"}).$$

Зафиксируем точку $x \in]a, a + r[$ и введем число A по формуле

$$A = \frac{f(x) - f(a)}{x - a} \quad \left(f(x) = f(a) + A \cdot (x - a) \right).$$

Рассмотрим функцию $\psi : [a, x] \rightarrow \mathbb{R}$:

$$\psi(t) = f(x) - f(t) - A \cdot (x - t).$$

Эта функция непрерывна и дифференцируема на $[a, x]$, так как наследует свойства аналитической функции f и полинома. Кроме того, $\psi(a) = \psi(x) = 0$. Следовательно, по теореме Ролля на $]a, x[$ найдется такая точка ξ , что $\psi'(\xi) = 0$, или $-f'(\xi) + A = 0$, т.е. $A = f'(\xi)$. Тот же результат получится, если $x \in]a - r, a[$.

Итак, для каждого $x \in]a - r, a + r[$ найдется такая точка ξ , что

$$\frac{f(x) - f(a)}{x - a} = f'(\xi),$$

или

$$f(x) = f(a) + f'(\xi) \cdot (x - a). \quad (10.2.1)$$

Это утверждение называется *теоремой Лагранжа*³¹, а формула (10.2.1) – *формулой Лагранжа* или *формулой конечных приращений*.

³¹Жозеф Луи ЛАГРАНЖ (J.L. Lagrange, 1736-1813) – французский математик и механик, президент Берлинской АН, член многих академий мира, один из разработчиков метрической системы мер.

Если известна какая-нибудь оценка для производной функции f в рассматриваемой окрестности точки a (например, $|f'| \leq M_1$), то из формулы Лагранжа видно, что погрешность от замены $f(x)$ на $f(a)$ не превосходит $M_1 \cdot |x - a|$.

Формула Лагранжа имеет простую геометрическую интерпретацию: $\frac{f(x) - f(a)}{x - a}$ – угловой коэффициент хорды, соединяющей концы графика функции $y = f(t)$ на $[a, x]$; $f'(\xi)$ – угловой коэффициент касательной к этому графику в точке $\xi \in]a, x[$. Таким образом, теорема Лагранжа утверждает, что на графике функции имеется точка $(\xi, f(\xi))$, в которой *касательная параллельна хорде* (рис.10.2).

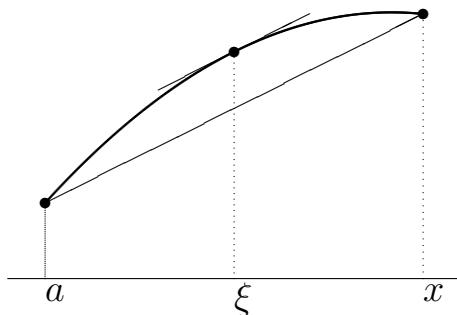


Рис.10.2

Серьезное предупреждение. Считаем необходимым подчеркнуть, что в формуле (10.2.1), так же как и во всех последующих аналогичных формулах, точка ξ зависит от x .

Заменим теперь функцию f суммой *двух* первых слагаемых определяющего ее ряда Тейлора (полиномом второго порядка):

$$f(x) \sim f(a) + \frac{f'(a)}{1!}(x - a).$$

Для оценивания погрешности, возникающей при такой замене, зафиксируем $x \in]a, a + r[$ и введем число A по формуле

$$A = \frac{f(x) - f(a) - \frac{f'(a)}{1!}(x - a)}{(x - a)^2} \quad \left(f(x) = f(a) + \frac{f'(a)}{1!}(x - a) + A \cdot (x - a)^2 \right).$$

Рассмотрим функцию $\psi : [a, x] \rightarrow \mathbb{R} :$

$$\psi(t) = f(x) - f(t) - \frac{f'(t)}{1!}(x - t) - A \cdot (x - t)^2.$$

Эта функция непрерывна и дифференцируема на $[a, x]$, так как наследует свойства аналитической функции f и полинома. Кроме того, $\psi(a) = \psi(x) = 0$. Следовательно, по теореме Ролля на $]a, x[$ найдется такая точка ξ , что $\psi'(\xi) = 0$. Вычисляем

$$\psi'(t) = -f'(t) + f'(t) - \frac{f''(t)}{1!}(x-t) + 2A \cdot (x-t),$$

$$\psi'(\xi) = -\frac{f''(\xi)}{1!}(x-\xi) + 2A \cdot (x-\xi) = 0.$$

Сокращая на $(x-\xi) \neq 0$, получим $A = \frac{f''(\xi)}{2!}$, откуда

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(\xi)}{2!}(x-a)^2.$$

Тот же результат получится, если $x \in]a-r, a[$.

Продолжение очевидно: заменим f первыми n слагаемыми определяющего ее ряда Тейлора:

$$f(x) \sim f(a) + \frac{f'(a)}{1!}(x-a) + \dots + \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1}$$

и оценим уже известным способом возникающую при этом погрешность. Получим утверждение:

Для всякого $x \in]a-r, a+r[$ между a и x найдется такая точка ξ , что

$$f(x) = \sum_{k=0}^{n-1} \frac{f^{(k)}(a)}{k!}(x-a)^k + \frac{f^{(n)}(\xi)}{n!}(x-a)^n.$$

Мы построили *формулу Тейлора порядка n* , которая представляет функцию f в виде суммы двух слагаемых:

1) *полином Тейлора* (порядка n)

$$T_f(n, a, x) = f(a) + \frac{f'(a)}{1!} \cdot (x-a) + \dots + \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1};$$

2) *остаточный член* формулы Тейлора

$$\frac{f^{(n)}(\xi)}{n!}(x-a)^n.$$

Замечания. 1. Обратите внимание на то, что остаточный член формулы Тейлора *похож* на слагаемые полинома Тейлора. Однако все значения производных функции f , содержащиеся в полиноме Тейлора, вычислены в известной точке a . Про точку же ξ , фигурирующую в остаточном члене, известно лишь то, что она лежит между a и x . Из сказанного следует, что значение полинома Тейлора можно *вычислить*, а значение остаточного члена – лишь *оценить*. Обычно бывает известна какая-нибудь оценка модуля n -й производной – $|f^{(n)}(x)| \leq M_n$. Тогда для остаточного члена имеем оценку

$$\left| \frac{f^{(n)}(x)}{n!} \cdot (x - a)^n \right| \leq \frac{M_n}{n!} \cdot |x - a|^n.$$

2. Анализ рассуждений, приведших к формуле Тейлора, показывает, что требование аналитичности функции f излишне. Справедливость равенства

$$f(x) = T_f(n, a, x) + \frac{f^{(n)}(\xi)}{n!} \cdot (x - a)^n$$

обеспечивается существованием непрерывной n -й производной функции f в рассматриваемой окрестности точки a .

Рассмотрим схему применения формулы Тейлора на примере вычисления значений экспоненты в окрестности нуля.

$$\exp(x) = T_{\exp}(n, 0, x) + \frac{\exp^{(n)}(\xi)}{n!} \cdot x^n.$$

1. Заменяем экспоненту ее полиномом Тейлора:

$$\exp(x) \sim T_{\exp}(n, 0, x) = 1 + \frac{x}{1!} + \dots + \frac{x^{n-1}}{(n-1)!}.$$

2. Оцениваем погрешность такой замены, т.е. модуль остаточного члена формулы Тейлора:

$$|\exp(x) - T_{\exp}(n, 0, x)| = \left| \frac{\exp^{(n)}(\xi)}{n!} \cdot x^n \right| = \frac{\exp(\xi)}{n!} \cdot |x|^n.$$

Пусть, например, замена осуществляется на сегменте $[0, 1]$. Тогда $0 < \xi < 1$, $\exp(\xi) < \exp(1) = e$. Поэтому

$$|\exp(x) - T_{\exp}(n, 0, x)| < e \cdot \frac{x^n}{n!} < \frac{e}{n!}.$$

Если $n = 7$, то $\frac{e}{7!} < 5.4 \cdot 10^{-4}$. Таким образом, замена

$$\exp(x) \sim 1 + x + \frac{x^2}{1.5} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120} + \frac{x^6}{720}$$

обеспечивает при $0 \leq x \leq 1$ *абсолютную* погрешность, меньшую, чем $5.4 \cdot 10^{-4}$.

10.3. Локальные экстремумы

Формула Тейлора служит основным инструментом локального исследования *гладких* функций (т.е. функций, имеющих достаточно производных для построения формулы Тейлора нужного порядка). Можно сказать, что гладкая функция в достаточно малой окрестности точки "ведет себя так же, как ее полином Тейлора". Это расплывчатое утверждение конкретизируется в п.п. 10.3 и 10.4.

Будем считать в этих пунктах, что функция f задана на некотором сегменте, и точка a не является концом этого сегмента.

Определение. Точка a называется *точкой локального максимума (минимума)* функции f , если у этой точки существует окрестность, для всех x из которой (кроме $x = a$) выполняется неравенство $f(x) < f(a)$ ($f(x) > f(a)$).

Точки локального максимума и точки локального минимума функции называют *точками ее локального экстремума*.

Теорема Ферма³². *Непрерывно дифференцируемая* функция не имеет экстремума в точках, где ее производная отлична от нуля.

Доказательство. Запишем в точке a формулу Тейлора второго порядка

$$f(x) = f(a) + \frac{f'(a)}{1!}(x - a) + \frac{f''(\xi)}{2!}(x - a)^2.$$

Пусть $f'(a) \neq 0$. Выберем настолько малую окрестность точки a , чтобы в ней знак приращения функции

$$f(x) - f(a) = \frac{f'(a)}{1!}(x - a) + \frac{f''(\xi)}{2!}(x - a)^2$$

³²Пьер ФЕРМА (P. Fermat, 1601-1665) – французский юрист и математик, один из основателей аналитической геометрии, автор основного принципа геометрической оптики. Одна из сформулированных им теорем – так называемая "великая теорема Ферма" – была доказана только в 1995 году.

определялся первым слагаемым этого приращения:

$$\operatorname{sign}(f(x) - f(a)) = \operatorname{sign}\left(\frac{f'(a)}{1!}(x - a)\right).$$

Это условие заведомо будет выполняться, если

$$\left|\frac{f'(a)}{1!}(x - a)\right| > \frac{M_2}{2!}(x - a)^2,$$

где M_2 – наибольшее значение модуля второй производной функции f на рассматриваемом сегменте.

Последнее неравенство выполняется при всех $x \neq a$, если $M_2 = 0$, и при $|x - a| < \frac{2|f'(a)|}{M_2}$, если $M_2 > 0$.

Если знак приращения функции совпадает со знаком произведения $\frac{f'(a)}{1!}(x - a)$, то очевидно, что приращение функции меняет знак в любой окрестности точки a и, следовательно, в этой точке нет экстремума. ■

Замечание. Мы доказали теорему, предполагая существование у функции второй производной. На самом деле теорема справедлива при наличии у функции лишь непрерывной *первой* производной.

Определение. Точки, в которых производная функции равна нулю, называются *стационарными точками* этой функции.

Пусть теперь a – стационарная точка функции f . Предположим, что $f''(a) \neq 0$, и исследуем поведение функции в малой окрестности точки a . Запишем формулу Тейлора третьего порядка

$$f(x) = f(a) + \frac{f''(a)}{2!}(x - a)^2 + \frac{f'''(\xi)}{3!}(x - a)^3$$

(напомним, что $f'(a) = 0$) и найдем такую окрестность точки a , где выполняется неравенство

$$\left|\frac{f''(a)}{2!}(x - a)^2\right| > \frac{M_3}{3!}|x - a|^3.$$

(Здесь M_3 – наибольшее значение модуля третьей производной функции на нашем сегменте).

Решение неравенства дает $|x - a| < \frac{3|f''(a)|}{M_3}$. В этой окрестности точки a

$$\operatorname{sign}(f(x) - f(a)) = \operatorname{sign}\left(\frac{f''(a)}{2!}(x - a)^2\right) = \operatorname{sign}(f''(a)).$$

Очевидно, что при $f''(a) < 0$ f имеет в точке a локальный максимум, а при $f''(a) > 0$ – локальный минимум.

Если a – стационарная точка и $f''(a) = 0$, то поведение гладкой функции f в окрестности точки $x = a$ определяется поведением полинома Тейлора более высокого порядка. Рассматривать этот случай мы не будем.

10.4. Направление выпуклости графика гладкой функции

Рассмотрим график функции $\sin : [-\pi, \pi] \rightarrow \mathbb{R}$ (рис.10.3):

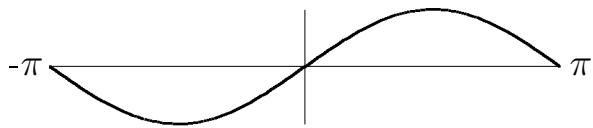


Рис. 10.3

В какой бы точке интервала $]-\pi, 0[$ ни провести касательную к графику, в некоторой окрестности этой точки график окажется *над этой касательной*. В какой бы точке интервала $]0, \pi[$ ни провести касательную к графику, в некоторой окрестности этой точки график окажется *под этой касательной*.

Определение. Говорят, что *на интервале* график функции *направлен выпуклостью вверх (вниз)*, если он лежит *под (над)* касательной, проведенной к нему в любой точке этого интервала. Функция f в этом случае называется выпуклой вверх (вниз) на интервале.

Теорема. На интервале, где $f'' < 0$ ($f'' > 0$), функция f выпукла вверх (соответственно, вниз).

Доказательство. Уравнение касательной, проведенной к графику функции $y = f(x)$ в точке $(a, f(a))$, имеет вид

$$Y = f(a) + f'(a) \cdot (x - a).$$

Найдем разность ординат графика и этой касательной в окрестности точки касания

$$y - Y = f(x) - f(a) - f'(a) \cdot (x - a).$$

Из формулы Тейлора следует, что между точками a и x найдется такая точка ξ , что

$$y - Y = f''(\xi) \frac{(x - a)^2}{2!}.$$

Поэтому при $f'' < 0$ $y - Y < 0$, т.е график функции лежит под касательной, а при $f'' > 0$ $y - Y > 0$ – график функции лежит над касательной. ■

Определение. Точка, разделяющая интервал, где функция выпукла вверх, и интервал, где функция выпукла вниз, называется *точкой перегиба* (точка 0 на рис.10.3).

Очевидно, что в точках перегиба вторая производная функции либо равна нулю, либо не существует. Обратное утверждение, вообще говоря, не верно. Например, если $f(x) = x^4$, то $f''(0) = 0$, однако точка 0 не является точкой перегиба для функции f (на всей оси эта функция выпукла вниз).

Терминологическое замечание. В математической литературе вместо "функция выпукла вниз" часто говорят просто "функция выпукла" а вместо "функция выпукла вверх" – "функция вогнута".

Глава 11. ФУНКЦИИ $\mathbb{R}^n \rightarrow \mathbb{R}^m$

11.1. Основные определения

Функцией нескольких вещественных переменных принято называть отображение части \mathbb{R}^n в \mathbb{R}^m ($n, m \in \mathbb{N}$, $n > 1$), т.е.

$$f : X \subset \mathbb{R}^n \rightarrow \mathbb{R}^m.$$

Такую функцию иначе называют *векторным полем*. Схематически векторное поле можно изобразить в виде "черного ящика" с n входами и m выходами (рис.11.1).

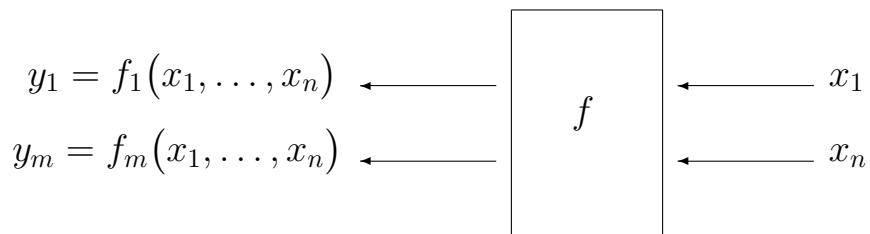


Рис.11.1

В частном случае ($m = 1$) получается *скалярное поле* (функционал)

$$f : X \subset \mathbb{R}^n \rightarrow \mathbb{R}.$$

Очевидно, что векторное поле – упорядоченный набор скалярных полей, каждое из которых может изучаться независимо от остальных. Поэтому мы начнем со случая скалярного поля.

Итак, пусть $f : X \subset \mathbb{R}^n \rightarrow \mathbb{R}$, и пусть $a = [a_1, \dots, a_n]^T$ – *внутренняя*³³ точка множества X .

Зафиксируем все координаты точки x , кроме k -й, полагая $x_j = a_j$ ($j \neq k$), и рассмотрим функцию одной "оставшейся" переменной x_k (сужение f на прямую, проходящую через точку a параллельно k -й оси координат)

$$\psi_k(x_k) = f(a_1, \dots, a_{k-1}, x_k, a_{k+1}, \dots, a_n).$$

³³Точка $a \in X \subset \mathbb{R}^n$ называется *внутренней* для X , если она имеет окрестность, состоящую *только* из точек X . Например, во множестве $V = \{x \mid x_1^2 + x_2^2 \leq 1\} \subset \mathbb{R}^2$ точка $[0, 0.5]^T$ – внутренняя, а точка $[0, 1]^T$ – нет (проверьте это!). Множество $X \subset \mathbb{R}^n$, все точки которого – внутренние, называется открытым в \mathbb{R}^n . Например, множество $U = \{x \mid x_1^2 + x_2^2 < 1\}$ – открыто в \mathbb{R}^2 (сравните его с множеством V).

Производная этой функции, вычисленная при $x_k = a_k$, называется *частной производной функции f в точке a по k -й переменной*.

$$D_k f(a) = \lim_{x_k=a_k} \frac{\psi_k(x_k) - \psi_k(a_k)}{x_k - a_k} = \\ = \lim_{x_k=a_k} \frac{f(a_1, \dots, a_{k-1}, x_k, a_{k+1}, \dots, a_n) - f(a_1, \dots, a_{k-1}, a_k, a_{k+1}, \dots, a_n)}{x_k - a_k}$$

(конечно, если этот предел существует).

Таким образом, в точке a определяются n чисел: $D_1 f(a), \dots, D_n f(a)$. Их записывают в виде строки, называют эту матрицу-строку *производной функции* (скалярного поля) f в точке a и обозначают $f'(a)$ или $Df(a)$:

$$f'(a) = Df(a) = [D_1 f(a), \dots, D_n f(a)].$$

Если производная существует во всех точках некоторого множества, то мы получаем заданную на этом множестве функцию-строку, которую называют *производной функцией* от функции f .

Пример. Если $f(x) = x_1 x_2^2 x_3^3$, $x \in \mathbb{R}^3$, то

$$D_1 f(x) = x_2^2 x_3^3, \quad D_2 f(x) = 2x_1 x_2 x_3^3, \quad D_3 f(x) = 3x_1 x_2^2 x_3^2, \\ f'(x) = Df(x) = [x_2^2 x_3^3, 2x_1 x_2 x_3^3, 3x_1 x_2^2 x_3^2].$$

Если транспонировать производную скалярного поля, то получится вектор (матрица-столбец), который называется *градиентом* скалярного поля. Пишут

$$\text{grad}(f) = \nabla f = (f')^T = [D_1 f, \dots, D_n f]^T.$$

Введение двух объектов – производной и градиента – оправдано тем, что в одних случаях удобно работать со строкой, а в других – со столбцом.

Определение. *Производной (матрицей Якоби³⁴)* векторного поля $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ называется $(m \times n)$ -матрица, строки которой – производные компонент этого поля (скалярных полей). Таким образом, если $f = [f_1, \dots, f_m]^T$, то

$$f' = Df = \begin{bmatrix} f'_1 \\ \vdots \\ f'_m \end{bmatrix} = \begin{bmatrix} D_1 f_1 & D_2 f_1 & \dots & D_n f_1 \\ \dots & \dots & \dots & \dots \\ D_1 f_m & D_2 f_m & \dots & D_n f_m \end{bmatrix}.$$

³⁴Карл Густав Якоб ЯКОБИ (K.G.J. Jacobi, 1804-1851) – немецкий математик, член Лондонского Королевского общества и многих академий Европы.

Если матрица Якоби квадратная, то ее определитель называется **якобианом** векторного поля.

Важное соглашение

Мы будем рассматривать в дифференциальном исчислении только гладкие функции нескольких переменных, т.е. будем считать, что частные производные существуют и непрерывны всюду в области определения функции.

Примеры. 1. Пусть A – постоянная $(m \times n)$ -матрица; $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $f(x) = Ax$. Покажите, что $f'(x) = A$.

В частности, если $m = 1$, то линейный функционал $f : \mathbb{R}^n \rightarrow \mathbb{R}$ можно записать в виде $f(x) = \langle x, a \rangle = a^T x$, где $a \in \mathbb{R}^n$. Соответственно, $f'(x) = a^T$, $\nabla f(x) = a$.

2. Пусть $f : [0, +\infty[\times [0, 2\pi[\rightarrow \mathbb{R}^2$,

$$x_1 = f_1(\rho, \varphi) = \rho \cdot \cos(\varphi); \quad x_2 = f_2(\rho, \varphi) = \rho \cdot \sin(\varphi).$$

Как известно (п.2.3), это отображение задает переход от полярных координат точки на плоскости к ее декартовым координатам. Вычислим матрицу Якоби и ее определитель (якобиан).

$$D_1 f_1(\rho, \varphi) = \cos(\varphi), \quad D_2 f_1(\rho, \varphi) = -\rho \cdot \sin(\varphi),$$

$$D_1 f_2(\rho, \varphi) = \sin(\varphi), \quad D_2 f_2(\rho, \varphi) = \rho \cdot \cos(\varphi);$$

$$f'(\rho, \varphi) = \begin{bmatrix} \cos(\varphi) & -\rho \cdot \sin(\varphi) \\ \sin(\varphi) & \rho \cdot \cos(\varphi) \end{bmatrix}; \quad \det(f'(\rho, \varphi)) = \rho.$$

11.2. Кривая и путь

Рассмотрим две содержательные интерпретации понятий, введенных в предыдущем пункте.

КРИВАЯ. Интуитивное представление о "кривой" (или "линии") можно получить так: возьмем резиновую нить, закрепим один ее конец в точке A , другой – в точке B , и растягивая нить, сделаем из нее нужную нам "кривую".

В соответствии с этим представлением естественно назвать кривой образ сегмента $[a, b]$ при *непрерывном* отображении

$$r : [a, b] \rightarrow \mathbb{R}^3 : \quad x_1 = r_1(t), \quad x_2 = r_2(t). \quad x_3 = r_3(t). \quad (11.2.1)$$

Однако вряд ли кто-нибудь согласится считать "линией" квадрат, а между тем существует непрерывное отображение сегмента $[0, 1]$ на квадрат $[0, 1] \times [0, 1]$ (так называемая *кривая Пеано*³⁵).

Имея в виду цели нашего курса, мы ограничимся рассмотрением *гладких* кривых, исключающих подобные патологии.

Определение. Пусть отображение $r : [a, b] \rightarrow \mathbb{R}^3$ непрерывно дифференцируемо и обладает следующими свойствами:

1. $r'(t) \neq \theta$ (производная отображения r *нигде* не обращается в нуль);
2. $r(t_1) = r(t_2) \implies t_1 = t_2$ (нет самопересечений).

Тогда множество значений этого отображения $\ell = r([a, b])$ называется *гладкой кривой* (см. рис.11.2).

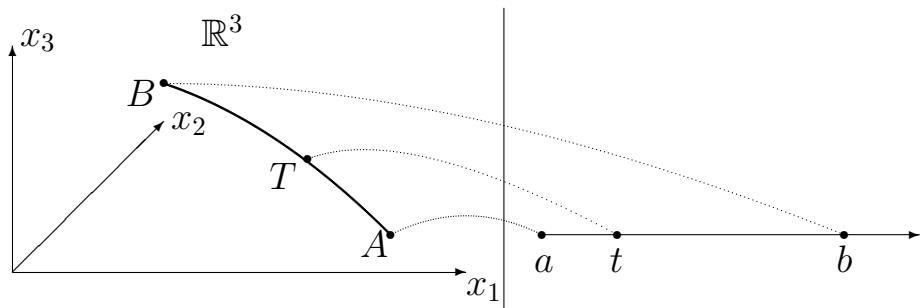


Рис.11.2

Пример. Кривая, заданная отображением

$$r(t) = a + (b - a) \cdot t, \quad a, b \in \mathbb{R}^3, \quad (a \neq b), \quad t \in [0, 1]$$

— это отрезок прямой, соединяющей точки a и b (рис.11.3).

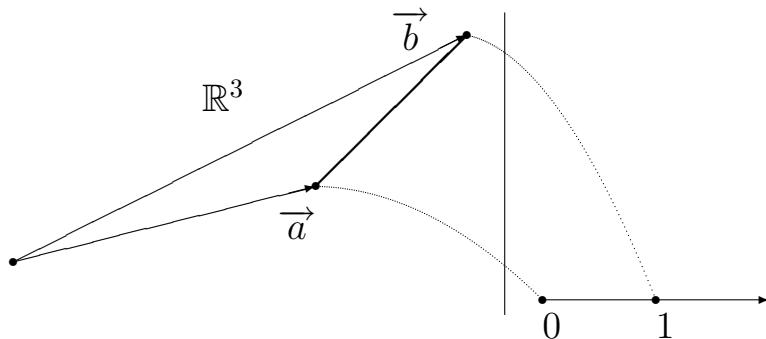


Рис.11.3

³⁵Джузеппе ПЕАНО (G. Peano, 1858-1932) — итальянский математик, член Турийской АН.

Замечания. 1. Точка $A = r(a)$ называется *началом* кривой, точка $B = r(b)$ – ее *концом*.

2. Условие взаимной однозначности отображения (условие 2 из определения) может нарушаться на концах сегмента $[a, b]$. В этом случае $A = r(a) = r(b) = B$, и такая гладкая кривая называется *замкнутой*.

3. Можно рассматривать отображение сегмента не в \mathbb{R}^3 , а в \mathbb{R}^2 (удовлетворяющее перечисленным в определении условиям). Множество значений такого отображения называется *плоской гладкой кривой*.

4. Уравнения (11.2.1) называются *параметрическими уравнениями* кривой, а переменная t – *параметром*.

Определение. Если гладкая кривая задана отображением $r : [a, b] \rightarrow \mathbb{R}^3$, и $t_0 \in [a, b]$, то ненулевой вектор $r'(t_0)$ называется *касательным вектором* к этой кривой в ее точке $x^{(0)} = r(t_0)$.

Пример. Рассмотрим график непрерывно дифференцируемой функции $f : [a, b] \rightarrow \mathbb{R}$. Это множество в \mathbb{R}^2 – образ сегмента $[a, b]$ при отображении

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = r(t) = \begin{bmatrix} t \\ f(t) \end{bmatrix}, \quad \text{причем } r'(t) = \begin{bmatrix} 1 \\ f'(t) \end{bmatrix} \neq \theta.$$

Поэтому такой график – гладкая плоская кривая.

Как известно, касательная к графику в точке $x^{(0)} = [t_0, f(t_0)]^T$ задается уравнением

$$x_2 - f(t_0) = f'(t_0) \cdot (x_1 - t_0) \quad \text{или} \quad \frac{x_2 - f(t_0)}{x_1 - t_0} = \frac{f'(t_0)}{1}.$$

Поэтому векторы $x - x^{(0)} = [x_1 - t_0, x_2 - f(t_0)]^T$ и $r'(t_0) = [1, f'(t_0)]^T$ линейно зависимы, т.е. касательная к кривой в точке $x^{(0)} = r(t_0)$ параллельна направленному отрезку $\overrightarrow{r'(t_0)}$ (рис.11.4). Можно показать, что этот факт имеет место и в общем случае.

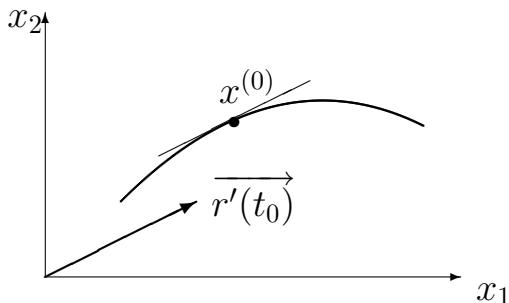


Рис.11.4

Замечания. 1. Условие $r'(t) \neq \theta$, $t \in [a, b]$ обеспечивает существование касательной к кривой во всех ее точках и, таким образом, оправдывает название "гладкая кривая". При нарушении этого условия у кривой может не быть касательной даже при непрерывно дифференцируемом отображении r . Проверьте, что кривая на рис.11.5 есть образ отображения $r(t) = [t \cdot |t|, t^2]^T$, $t \in [-1, 1]$ ($r'(0) = \theta$).

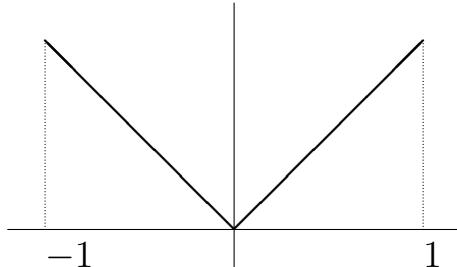


Рис.11.5

2. Наряду с гладкими кривыми в содержательных задачах встречаются *кусочно гладкие*. Мы будем называть кусочно гладкой кривой объединение конечного числа гладких кривых, которые

- 1) образуют упорядоченный набор,
- 2) не пересекают друг друга и
- 3) сцеплены между собой так, что конец предыдущего куска является началом следующего.

Если конец последнего куска совпадает с началом первого, то кусочно гладкая кривая называется *замкнутой*.

Пример. Граница квадрата $[0, 1] \times [0, 1]$ – кусочно гладкая (плоская, замкнутая) кривая, состоящая из четырех кусков (отрезков прямых).

ПУТЬ. В приложениях к механике естественно считать параметр t временем, а также отказаться от требования взаимной однозначности отображения $t \rightarrow r(t)$.

Определение. *Гладким путем* называется непрерывно дифференцируемое отображение сегмента в \mathbb{R}^3 , производная которого не обращается в нуль. Аналогично определим *плоский гладкий путь*.

Множество значений гладкого пути – множество точек в \mathbb{R}^3 (\mathbb{R}^2) – называется *носителем* этого пути. Удобно представлять себе носитель пути как "рельсы по которым движется точка". Например, три различных плоских пути

$$\begin{aligned} r^{(1)}(t) &= [\cos(t), \sin(t)]^T, & r^{(2)}(t) &= [\cos(2t), \sin(2t)]^T, \\ r^{(3)}(t) &= [\cos(t), -\sin(t)]^T & (t \in [0, 2\pi]) \end{aligned}$$

имеют один и тот же носитель – окружность единичного радиуса с центром в начале координат, но первая точка проходит эту окружность один раз, а вторая – дважды (за то же время), а третья – один раз, но в противоположном направлении.

Вектор $r'(t)$ есть мгновенная скорость движения точки. Проверьте, что в нашем примере $\|r^{(1)'}(t)\| = \|r^{(3)'}(t)\| = 1$, $\|r^{(2)'}(t)\| = 2$.

Замечания. 1. В математике невозможно "пройти один и тот же путь с разными скоростями"! Это будут разные пути (хотя и с общим носителем).

2. *Кусочно гладкий* путь определяется аналогично кусочно гладкой кривой.

3. Отображение, задающее гладкую кривую, является, очевидно, гладким путем. В этом случае вектор скорости совпадает с касательным вектором к кривой.

11.3. Поверхность

Придать точный смысл общему понятию "поверхность" еще сложнее, чем понятию "кривая". Мы ограничимся частным случаем.

Определение. Пусть $\Omega \subset \mathbb{R}^2$ – часть плоскости, ограниченная замкнутой кусочно гладкой кривой. *Гладкой поверхностью* $G = r(\Omega)$ называется образ множества Ω в \mathbb{R}^3 при отображении

$$r : \Omega \rightarrow \mathbb{R}^3; \quad x_1 = r_1(u), \quad x_2 = r_2(u), \quad x_3 = r_3(u), \quad (11.3.1)$$

обладающим следующими свойствами:

- 1) взаимная однозначность ($r(u) = r(v) \implies u = v$);
- 2) непрерывная дифференцируемость;
- 3) линейная независимость столбцов матрицы Якоби Dr .

Замечание. Эти условия могут нарушаться на границе Ω .

Примеры. 1. $r : [0, 2\pi] \times [0, H] \rightarrow \mathbb{R}^3$;

$$r(\varphi, z) = [R \cdot \cos(\varphi), R \cdot \sin(\varphi), z]^T, \quad Dr(\varphi, z) = \begin{bmatrix} -R \cdot \sin(\varphi) & 0 \\ R \cdot \cos(\varphi) & 0 \\ 0 & 1 \end{bmatrix}.$$

Проверьте, что это отображение задает боковую поверхность прямого кругового цилиндра высоты H и радиуса R . Заметьте, что $r(2\pi, z) \equiv r(0, z)$.

2. $r : [0, \pi] \times [0, 2\pi] \rightarrow \mathbb{R}^3$;

$$r(\vartheta, \varphi) = [R \cdot \sin(\vartheta) \cdot \cos(\varphi), R \cdot \sin(\vartheta) \cdot \sin(\varphi), R \cdot \cos(\vartheta)]^T.$$

Проверьте, что это отображение задает сферу радиуса R с центром в начале координат, причем $r(0, \varphi) \equiv [0, 0, R]^T$ – "северный полюс" $r(\pi, \varphi) \equiv [0, 0, -R]^T$ – "южный полюс" $r(\vartheta, 0) \equiv r(\vartheta, 2\pi)$ – "нулевой меридиан" $r(\pi/2, \varphi)$ – "экватор".

Убедитесь, что столбцы матрицы

$$Dr(\vartheta, \varphi) = \begin{bmatrix} R \cdot \cos(\vartheta) \cdot \cos(\varphi) & -R \cdot \sin(\vartheta) \cdot \sin(\varphi) \\ R \cdot \cos(\vartheta) \cdot \sin(\varphi) & R \cdot \sin(\vartheta) \cdot \cos(\varphi) \\ -R \cdot \sin(\vartheta) & 0 \end{bmatrix}$$

линейно независимы во всех точках сферы, кроме полюсов.

Пусть $a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$ – внутренняя точка Ω . Рассмотрим сужения отображения r на отрезки, проходящие через эту точку параллельно осям Ou_1 и Ou_2 : $\rho_1(u_1) = r(u_1, a_2)$, $\rho_2(u_2) = r(a_1, u_2)$ (рис.11.6). Их образы – кривые, проходящие через точку $A = r(a)$ и лежащие на поверхности G . Легко видеть, что касательные векторы к этим кривым в точке A – столбцы матрицы Якоби отображения r в точке a :

$$\rho'_1(a_1) = D_1r(a), \quad \rho'_2(a_2) = D_2r(a).$$

В силу линейной независимости векторов $D_1r(a)$ и $D_2r(a)$ соответствующие им направленные отрезки неколлинеарны и определяют единственную плоскость, проходящую через точку A . Ее называют *касательной плоскостью* к поверхности G в точке A .

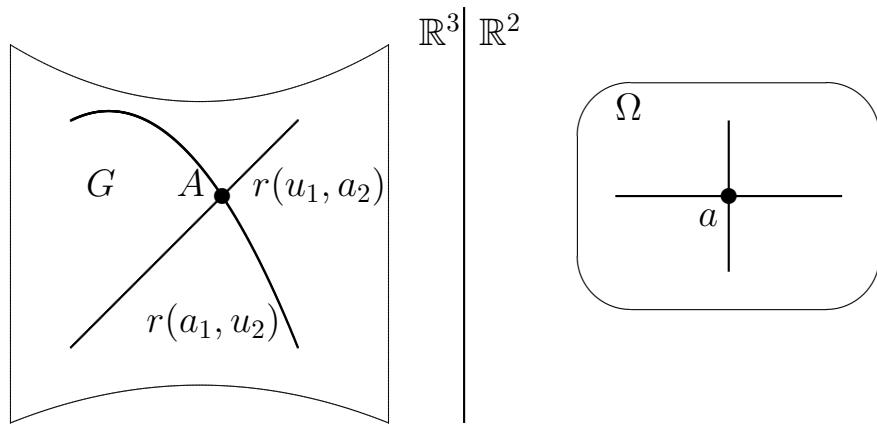


Рис.11.6

Пример. Пусть G – график непрерывно дифференцируемой функции $f : \Omega \rightarrow \mathbb{R}$. Тогда поверхность G задается отображением

$$r(u_1, u_2) = [u_1, u_2, f(u_1, u_2)]^T,$$

причем

$$Dr(a) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ D_1f(a) & D_2f(a) \end{bmatrix}, \quad a \in \Omega.$$

Очевидно, что столбцы матрицы $Dr(a)$ линейно независимы, и вектор $v = [D_1f(a), D_2f(a), -1]^T$ ортогонален им обоим. Следовательно, соответствующий направленный отрезок перпендикулярен касательной плоскости к поверхности G в ее точке $A = r(a)$. Из курса линейной алгебры известно, что уравнение этой плоскости имеет вид $\langle x - A, v \rangle = 0$, или

$$x_3 - f(a_1, a_2) = D_1f(a_1, a_2) \cdot (x_1 - a_1) + D_2f(a_1, a_2) \cdot (x_2 - a_2). \quad (11.3.2)$$

Пример. Пусть $f(x_1, x_2) = x_1^2/4 + x_2^2/9$ (график этой функции – эллиптический параболоид). Убедитесь, что точка $(2, 3, 2)$ лежит на параболоиде. Проведем в этой точке касательную плоскость к параболоиду (напишем уравнение этой плоскости). Здесь

$$Df(x_1, x_2) = \left[\frac{2x_1}{4}, \frac{2x_2}{9} \right], \quad D_1f(2, 3) = 1, \quad D_2f(2, 3) = 2/3.$$

Уравнение касательной плоскости имеет вид

$$x_3 - 2 = 1 \cdot (x_1 - 2) + 2/3 \cdot (x_2 - 3) \quad \text{или} \quad x_1 + 2/3 \cdot x_2 - x_3 = 2.$$

Замечания. 1. Условие линейной независимости векторов D_1r и D_2r обеспечивает наличие касательной плоскости во всех точках поверхности и, таким образом, оправдывает название "гладкая поверхность".

2. Можно рассматривать *кусочно гладкие* поверхности, т.е. поверхности, состоящие из нескольких гладких кусков, сцепленных своими краями. Простейший пример кусочно гладкой поверхности – многогранник.

11.4. Композиция функций и ее производная

Пусть заданы векторные поля $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ и $g : \mathbb{R}^m \rightarrow \mathbb{R}^k$. Тогда на \mathbb{R}^n определена их композиция – векторное поле $h = g \circ f : \mathbb{R}^n \rightarrow \mathbb{R}^k$.

В соответствии с **Важным соглашением** (п.11.1) будем предполагать, что f и g непрерывно дифференцируемы. Тогда их производные (матрицы Якоби) имеют размеры $k \times m$ и $m \times n$ соответственно.

Известно, что в случае функций одной переменной непрерывная дифференцируемость функций f и g влечет за собой непрерывную дифференцируемость их композиции $h = g \circ f$, причем (п.8.2)

$$h'(x) = (g \circ f)'(x) = g'(f(x)) \cdot f'(x).$$

Можно показать, что в многомерном случае имеет место аналогичная

Теорема. Если функции f и g имеют непрерывные производные, то их композиция $h = g \circ f$ также непрерывно дифференцируема, причем

$$h' = (g \circ f)' = (g' \circ f) \cdot f'. \quad (11.4.1)$$

Проверьте согласованность размеров перемножаемых матриц Якоби!

Примеры. 1. Пусть даны числовые матрицы A (размера $m \times n$) и B (размера $k \times m$); $f(x) = Ax$, $g(y) = By$. Тогда

$$h(x) = (g \circ f)(x) = B(Ax) = (BA)x \quad (BA - \text{матрица размера } k \times n).$$

Из примера 1 (п.11.1) следует, что

$$f'(x) = A, \quad g'(y) = B, \quad h'(x) = BA.$$

2. Пусть гладкая кривая ℓ задана отображением $r : [a, b] \rightarrow \mathbb{R}^3$, а непрерывно дифференцируемая функция f взаимно однозначно отображает $[\alpha, \beta]$ на $[a, b]$, причем f' не обращается в нуль. Тогда композиция $\rho = r \circ f$ задает ту же кривую ℓ . При этом по формуле (11.4.1)

$$\rho'(t) = r'(f(t)) \cdot f'(t), \quad t \in [\alpha, \beta],$$

откуда видно, что касательные векторы $\rho'(t)$ и $r'(f(t))$ различны, но линейно зависимы. Поэтому соответствующие им направленные отрезки коллинеарны, и, следовательно, задают одну и ту же касательную к кривой.

Матричное соотношение (11.4.1) может быть записано в координатной форме:

$$D_j(g \circ f)_i(x) = [D_1 g_i(f(x)), \dots, D_m g_i(f(x))] \cdot \begin{bmatrix} D_j f_1(x) \\ \vdots \\ D_j f_m(x) \end{bmatrix}, \quad \begin{array}{l} i = 1, \dots, k; \\ j = 1, \dots, n. \end{array}$$

Оперировать *матричным* соотношением (11.4.1) гораздо проще, однако при "ручном" счете может возникнуть потребность и в его *координатной* форме.

11.5. Понятие о конформном отображении

В этом пункте координаты точки на плоскости обозначаются x и y , $z = x + iy$.

Напомним, что аналитическая функция была введена нами в п.7.1 как сумма степенного ряда. Она определена в круге сходимости этого ряда и имеет там производные всех порядков. В частности,

$$f'(z_0) = \lim_{z=z_0} \frac{f(z) - f(z_0)}{z - z_0}. \quad (11.5.1)$$

Функция f порождает пару вещественных функций, заданных внутри круга сходимости:

$$F_1(x, y) = \operatorname{Re}(f(z)), \quad F_2(x, y) = \operatorname{Im}(f(z)).$$

Используя эти функции, можно переписать (11.5.1) в виде

$$f'(z_0) = \lim_{\substack{x=x_0 \\ y=y_0}} \frac{F_1(x, y) - F_1(x_0, y_0) + i(F_2(x, y) - F_2(x_0, y_0))}{x - x_0 + i(y - y_0)}.$$

При одном порядке вычисления пределов получим

$$\begin{aligned} f'(z_0) &= \lim_{y=y_0} \left(\lim_{x=x_0} \frac{F_1(x, y) - F_1(x_0, y_0) + i(F_2(x, y) - F_2(x_0, y_0))}{x - x_0 + i(y - y_0)} \right) = \\ &= \lim_{y=y_0} \frac{F_1(x_0, y) - F_1(x_0, y_0) + i(F_2(x_0, y) - F_2(x_0, y_0))}{i(y - y_0)} = \\ &= D_2 F_2(x_0, y_0) - i \cdot D_2 F_1(x_0, y_0). \end{aligned} \quad (11.5.2)$$

Изменение порядка даст

$$\begin{aligned} f'(z_0) &= \lim_{x=x_0} \left(\lim_{y=y_0} \frac{F_1(x, y) - F_1(x_0, y_0) + i(F_2(x, y) - F_2(x_0, y_0))}{x - x_0 + i(y - y_0)} \right) = \\ &= \lim_{x=x_0} \frac{F_1(x, y_0) - F_1(x_0, y_0) + i(F_2(x, y_0) - F_2(x_0, y_0))}{x - x_0} = \\ &= D_1 F_1(x_0, y_0) + i \cdot D_1 F_2(x_0, y_0). \end{aligned} \quad (11.5.3)$$

Из (11.5.2) и (11.5.3) вытекают тождества, известные как *условия Коши–Римана*³⁶:

$$\boxed{D_1 F_1(x, y) \equiv D_2 F_2(x, y) \equiv \operatorname{Re}(f'(z)), \\ D_1 F_2(x, y) \equiv -D_2 F_1(x, y) \equiv \operatorname{Im}(f'(z)).}$$

Матрица Якоби отображения $F = [F_1, F_2]^T$ принимает теперь вид

$$F'(x, y) = \begin{bmatrix} \operatorname{Re}(f'(z)) & -\operatorname{Im}(f'(z)) \\ \operatorname{Im}(f'(z)) & \operatorname{Re}(f'(z)) \end{bmatrix}.$$

Матрица Грама для $F'(x, y)$ равна

$$(F'(x, y))^* \cdot F'(x, y) = |f'(z)|^2 \cdot I_2,$$

и, следовательно, если $f'(z) \neq 0$, то матрица $\frac{1}{|f'(z)|} \cdot F'(x, y)$ – ортогональная.

Рассмотрим теперь пару плоских кривых $z = r(t)$ и $z = s(\tau)$, проходящих через точку $z_0 = r(t_0) = s(\tau_0)$. Углом между кривыми в точке их пересечения называют угол между направленными отрезками, соответствующими касательным векторам к кривым в этой точке (рис.11.7а).

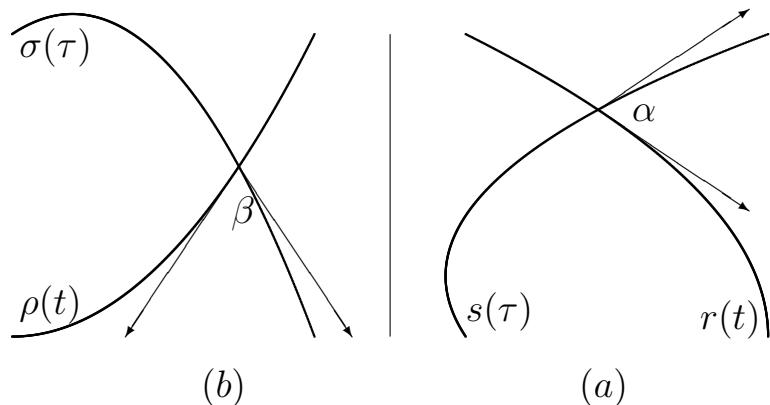


Рис.11.7

Из курса линейной алгебры известно, что

$$\cos(\alpha) = \frac{\langle r'(t_0), s'(\tau_0) \rangle}{\|r'(t_0)\| \cdot \|s'(\tau_0)\|}.$$

³⁶Георг Фридрих Бернгард РИМАН (G.F.B. Riemann, 1826-1866) – немецкий математик, один из основателей неевклидовой геометрии. Автор многих фундаментальных результатов в различных разделах математики.

Образы наших кривых при отображении F задаются композициями $\rho = F \circ r$ и $\sigma = F \circ s$ соответственно.

Из теоремы о производной композиции имеем

$$\rho'(t_0) = F'(x_0, y_0) \cdot r'(t_0), \quad \sigma'(\tau_0) = F'(x_0, y_0) \cdot s'(\tau_0).$$

Следовательно, косинус угла между этими образами в точке $\rho(t_0) = \sigma(\tau_0)$ равен (рис.11.7б)

$$\cos(\beta) = \frac{\langle \rho'(t_0), \sigma'(\tau_0) \rangle}{\|\rho'(t_0)\| \cdot \|\sigma'(\tau_0)\|}.$$

Но для любых двух векторов $a, b \in \mathbb{R}^2$

$$\begin{aligned} \langle F'(x_0, y_0) \cdot a, F'(x_0, y_0) \cdot b \rangle &= \langle (F'(x_0, y_0))^* \cdot F'(x_0, y_0) \cdot a, b \rangle = \\ &= |f'(z_0)|^2 \cdot \langle a, b \rangle; \end{aligned}$$

$$\|F'(x_0, y_0) \cdot a\| = (\langle F'(x_0, y_0) \cdot a, F'(x_0, y_0) \cdot a \rangle)^{1/2} = |f'(z_0)| \cdot \|a\|.$$

Поэтому, если $f'(z_0) \neq 0$, то

$$\cos(\beta) = \frac{|f'(z_0)|^2 \cdot \langle r'(t_0), s'(\tau_0) \rangle}{|f'(z_0)| \cdot \|r'(t_0)\| \cdot |f'(z_0)| \cdot \|s'(\tau_0)\|} = \cos(\alpha).$$

Мы показали, что в точках, где $f'(z) \neq 0$, отображение F сохраняет углы между гладкими кривыми. Поэтому отображение, порождаемое аналитической функцией, производная которой не обращается в нуль, называется *конформным*.

Пример. Поскольку $\exp'(z) = \exp(z) \neq 0$, экспонента порождает конформное отображение. Проверьте, что ортогональные друг другу семейства прямых $Re(z) = const$ и $Im(z) = const$ переходят при отображении $w = \exp(z)$ в ортогональные друг другу семейство окружностей $|w| = const$ и семейство лучей $\arg(w) = const$ соответственно.

11.6. Производные высших порядков

Скалярное поле $f : \mathbb{R}^n \rightarrow \mathbb{R}$ порождает n скалярных полей $D_1 f, \dots, D_n f : \mathbb{R}^n \rightarrow \mathbb{R}$. В силу **Важного соглашения** (п.11.1) мы рассматриваем только ситуации, в которых все они непрерывно дифференцируемы. Поэтому каждое из них, в свою очередь, порождает n скалярных полей

$$\begin{aligned} D_1(D_1 f), \dots, D_n(D_1 f), \\ \dots \dots \dots \dots \dots \dots \\ D_1(D_n f), \dots, D_n(D_n f), \end{aligned}$$

которые называют *вторыми частными производными* функции f . Каждая вторая частная производная порождает n *третьих частных производных* и т.д.

Вторые частные производные сводятся в квадратную матрицу порядка n . Она называется *второй производной* скалярного поля f или его *матрицей Гессе*³⁷:

$$f'' = \begin{bmatrix} D_1(D_1 f) & \dots & D_n(D_1 f) \\ \dots & \dots & \dots \\ D_1(D_n f) & \dots & D_n(D_n f) \end{bmatrix}.$$

Пример. $f(x_1, x_2) = x_1^3 x_2^2$.

$$\begin{aligned} D_1 f(x_1, x_2) &= 3x_1^2 x_2^2, & D_2 f(x_1, x_2) &= 2x_1^3 x_2, \\ D_1(D_1 f)(x_1, x_2) &= 6x_1 x_2^2, & D_2(D_1 f)(x_1, x_2) &= 6x_1^2 x_2, \\ D_1(D_2 f)(x_1, x_2) &= 6x_1^2 x_2, & D_2(D_2 f)(x_1, x_2) &= 2x_1^3. \end{aligned}$$

Матрица Гессе имеет вид

$$f''(x_1, x_2) = \begin{bmatrix} 6x_1 x_2^2 & 6x_1^2 x_2 \\ 6x_1^2 x_2 & 2x_1^3 \end{bmatrix}.$$

Обратите внимание на то, что в нашем примере $D_1 D_2 f = D_2 D_1 f$. *Можно показать*, что из непрерывности вторых смешанных частных производных $D_i D_j f$ и $D_j D_i f$ следует их равенство и, следовательно, симметричность матрицы Гессе.

Замечание. В отличие от функции одной переменной мы не можем написать $f'' = (f')'$ (так как f' – строка, а не столбец!). Проверьте, что верна формула

$$f'' = ((f')^T)' = (\nabla f)'.$$

Пример. Пусть $f(x) = \langle Ax, x \rangle$ и $A = A^T$. Покажите, что

$$\nabla f(x) = 2Ax, \quad f''(x) = 2A.$$

11.7. Формула Тейлора второго порядка

Пусть $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Рассмотрим приращение функции f при переходе из точки $a \in \mathbb{R}^n$ в точку $a + h \in \mathbb{R}^n$:

$$f(a + h) - f(a).$$

³⁷Людвиг Отто ГЕССЕ (L.O. Hesse, 1811-1874) – немецкий математик.

Пусть этот переход происходит по "отрезку прямой" в \mathbb{R}^n

$$r(t) = a + t \cdot e, \quad 0 \leq t \leq \|h\|,$$

где $e = h/\|h\|$ – единичный вектор, сонаправленный с h .

Построим функцию

$$\phi = f \circ r; \quad \phi(t) = f(a + t \cdot e). \quad (11.7.1)$$

Тогда $f(a + h) - f(a) = \phi(\|h\|) - \phi(0)$. По формуле Тейлора второго порядка для функции ϕ

$$\phi(\|h\|) - \phi(0) = \phi'(0) \cdot \|h\| + \phi''(\xi) \cdot \frac{\|h\|^2}{2}, \quad (11.7.2)$$

где ξ – некоторая точка из интервала $]0, \|h\||$.

По теореме о производной композиции

$$\phi'(t) = f'(a + t \cdot e) \cdot e, \quad \phi'(0) = f'(a) \cdot e. \quad (11.7.3)$$

Далее,

$$\begin{aligned} \phi''(t) &= (f'(a + t \cdot e) \cdot e)' = (e^T \cdot \nabla f(a + t \cdot e))' = \\ &= e^T \cdot (\nabla f(a + t \cdot e))' = e^T \cdot f''(a + t \cdot e) \cdot e. \end{aligned} \quad (11.7.4)$$

Подставляя значения функции ϕ и ее производных в (11.7.2), получим *формулу Тейлора второго порядка для скалярного поля* (сравните с (10.2.2))

$$f(a + h) = f(a) + f'(a) \cdot h + \frac{1}{2} \cdot h^T \cdot f''(a + \xi \cdot e) \cdot h. \quad (11.7.5)$$

Заметьте, что $f'(a)$, h^T – строки, $f''(a + \xi \cdot e)$ – квадратная матрица, а h – столбец, поэтому оба слагаемых в правой части (11.7.5) – числа.

Формулу (11.7.5) можно переписать иначе, используя понятие градиента:

$$\begin{aligned} f(a + h) &= f(a) + \langle \nabla f(a), h \rangle + \frac{1}{2} \cdot \langle f''(a + \xi \cdot e) \cdot h, h \rangle = \\ &= f(a) + \langle \nabla f(a), e \rangle \cdot \|h\| + \langle f''(a + \xi \cdot e) \cdot e, e \rangle \cdot \frac{\|h\|^2}{2}. \end{aligned} \quad (11.7.6)$$

Если $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ – векторное поле, то, объединив формулы (11.7.5) для всех его компонент, получим формулу Тейлора

$$F(a + h) = F(a) + F'(a) \cdot h + \alpha \cdot \frac{\|h\|^2}{2}, \quad (11.7.7)$$

где F' – $(m \times n)$ -матрица Якоби, α – m -мерный вектор, компоненты которого выражаются через значения вторых производных компонент поля F .

Формула Тейлора позволяет доказать теорему о локальных экстремумах скалярного поля (сравните ее с теоремой Ферма).

Теорема. Если $\nabla f(a) \neq \theta$, то в точке a скалярное поле f не имеет экстремума.

Доказательство. Положим в формуле (11.7.6) $h = t \cdot \nabla f(a)$:

$$f(a + h) - f(a) = \|\nabla f(a)\|^2 \left(1 + \langle f''(a + \xi \cdot e) \cdot e, e \rangle \cdot \frac{t}{2} \right) \cdot t.$$

Обозначим

$$M_2 = \sup_{\xi \in [-\varepsilon, \varepsilon]} |\langle f''(a + \xi \cdot e) \cdot e, e \rangle|, \quad \text{где } \varepsilon = \|\nabla f(a)\|.$$

Тогда при достаточно малых значениях $|t|$ $\left(|t| < \frac{2}{M_2}\right)$ выражение в скобках положительно и, следовательно,

$$\operatorname{sign}(f(a + h) - f(a)) = \operatorname{sign}(t).$$

Последнее равенство говорит о том, что приращение функции меняет знак в любой окрестности точки a , т.е. в этой точке нет экстремума. ■

Замечания. 1. Точки, в которых $\nabla f = \theta$, называются *стационарными* точками функции f .

2. Мы доказали теорему в предположении существования *второй* производной у функции f . Однако она справедлива и при наличии у f лишь *первой непрерывной* производной.

11.8. Формула Тейлора третьего порядка.

Исследование гладкого функционала в окрестности стационарной точки

Запишем для функции ϕ (см. (11.7.1)) формулу Тейлора третьего порядка

$$\phi(\|h\|) - \phi(0) = \phi'(0) \cdot \|h\| + \phi''(0) \cdot \frac{\|h\|^2}{2!} + \phi^{(3)}(\xi) \cdot \frac{\|h\|^3}{3!}$$

и вычислим с ее помощью приращение скалярного поля:

$$f(a + h) - f(a) = \phi(\|h\|) - \phi(0).$$

Согласно (11.7.3) и (11.7.4) имеем

$$\phi'(0) = f'(a)e, \quad \phi''(0) = e^T f''(a)e$$

(напомним, что $e = \frac{h}{\|h\|}$ – единичный вектор, сонаправленный с h).

В силу **Важного соглашения** f имеет непрерывные частные производные третьего порядка, т.е. $\phi^{(3)}$ существует и непрерывна.

Имеем

$$f(a + h) - f(a) = f'(a)h + \frac{1}{2!}h^T \cdot f''(a) \cdot h + \phi^{(3)}(\xi) \frac{\|h\|^3}{3!} \quad (11.8.1)$$

или

$$f(a + h) = f(a) + \langle \nabla f(a), h \rangle + \frac{1}{2!} \langle f''(a) \cdot h, h \rangle + \phi^{(3)}(\xi) \frac{\|h\|^3}{3!}$$

(здесь по-прежнему ξ – некоторая точка на интервале $]0, \|h\|[$).

Формула (11.8.1) называется *формулой Тейлора третьего порядка для скалярного поля*.

Исследуем с помощью этой формулы вопрос о наличии экстремума функции f в стационарной точке a . Так как $\nabla f(a) = \theta$, то

$$f(a + h) - f(a) = \frac{1}{2} \langle f''(a) \cdot h, h \rangle + \phi^{(3)}(\xi) \frac{\|h\|^3}{6}.$$

Из курса линейной алгебры известно, что квадратичная форма удовлетворяет неравенству

$$\lambda_{min} \|h\|^2 \leq \langle f''(a)h, h \rangle \leq \lambda_{max} \|h\|^2,$$

где λ_{min} и λ_{max} – наименьшее и наибольшее собственные числа симметричной матрицы Гессе $f''(a)$.

Если матрица $f''(a)$ положительно определена, то $\lambda_{min} > 0$ и

$$f(a + h) - f(a) > \frac{\lambda_{min}}{2} \cdot \|h\|^2 + \phi^{(3)}(\xi) \frac{\|h\|^3}{6} = \left(\frac{\lambda_{min}}{2} + \phi^{(3)}(\xi) \frac{\|h\|}{6} \right) \|h\|^2.$$

Можно взять вектор h настолько малым по норме, чтобы обеспечить выполнение неравенства $\frac{\lambda_{min}}{2} + \phi^{(3)}(\xi) \frac{\|h\|}{6} > 0$ в окрестности точки a . В

этой окрестности $f(a + h) - f(a) > 0$, если $h \neq \theta$, т.е. функция f имеет в *стационарной* точке a локальный минимум. Аналогично доказывается, что в случае *отрицательно определенной* матрицы Гессе $f''(a)$ функция f имеет в *стационарной* точке a локальный максимум.

Если матрица Гессе знакопеременна, т.е. среди ее собственных чисел есть хотя бы одно отрицательное (обозначим его λ_m) и хотя бы одно положительное (обозначим его λ_p), то, выбрав в качестве h собственный вектор h_p этой матрицы, соответствующий λ_p , получим

$$f(a + h_p) - f(a) = \frac{\lambda_p}{2} \|h_p\|^2 + \phi^{(3)}(\xi) \frac{\|h_p\|^3}{6} = \left(\frac{\lambda_p}{2} + \phi^{(3)}(\xi) \frac{\|h_p\|}{6} \right) \|h_p\|^2.$$

Взяв вектор h_p достаточно малым по норме, получим

$$f(a + h_p) - f(a) > 0.$$

Выбрав в качестве h собственный вектор h_m матрицы Гессе, соответствующий λ_m , получим

$$f(a + h_m) - f(a) = \left(\frac{\lambda_m}{2} + \phi^{(3)}(\xi) \frac{\|h_m\|}{6} \right) \|h_m\|^2.$$

Взяв вектор h_m достаточно малым по норме, получим

$$f(a + h_m) - f(a) < 0.$$

Мы показали, что приращение функции f меняет знак в любой окрестности точки a . Следовательно, в этой точке локального экстремума нет.

Если матрица Гессе полуопределенная (имеет хотя бы одно собственное число $\lambda_0 = 0$, а все ненулевые – одного знака), то выбрав в качестве h ее собственный вектор h_0 , соответствующий λ_0 , получим

$$f(a + h_0) - f(a) = \phi^{(3)}(\xi) \frac{\|h_0\|^3}{6}.$$

Видно, что формула Тейлора *третьего* порядка в этом случае не дает возможности судить о знаке приращения функции в окрестности стационарной точки. Использование же формулы Тейлора более высоких порядков выходит за рамки нашего курса.

Глава 12. ТЕОРЕМА О НЕЯВНО ЗАДАННОЙ ФУНКЦИИ И ЕЕ ПРИЛОЖЕНИЯ

12.1. Неявно заданные функции

Начнем с примера. Пусть задано *одно* уравнение с *двумя вещественными* переменными

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

Будем задавать произвольные значения x и решать получившееся уравнение относительно y (напоминаем, что нас интересуют только *вещественные* решения!). Возможны три ситуации:

1. При данном x уравнение *не имеет* решений. Например, при $x = 2$ получаем уравнение $y^2 = -3$.
2. При данном x уравнение имеет *несколько* решений. Например, при $x = 0$ получаем уравнение $y^2 = 1$, т.е. $y_1 = 1$, $y_2 = -1$.
3. При данном x уравнение имеет *ровно одно* решение. Например, при $x = 1$ получаем уравнение $y^2 = 0$, т.е. $y = 0$.

Определение. Пусть на открытом прямоугольнике $]a, b[\times]c, d[$ задана *вещественная* функция F . Если для всякого $x \in]a, b[$ уравнение $F(x, y) = 0$ имеет *ровно одно* решение $y \in]c, d[$, то говорят, что это уравнение *неявно задает* на интервале $]a, b[$ функцию *со значениями из интервала* $]c, d[$. Если обозначить эту функцию f , то при всех $x \in]a, b[$ $F(x, f(x)) = 0$.

Замечание. Отличительной особенностью неявного задания функции является наличие в алгоритме вычисления ее значения операции "решить уравнение" (естественно, без указания способа фактического выполнения этой операции).

Рассмотрим для начала *линейное* уравнение

$$a \cdot (x - x_0) + b \cdot (y - y_0) = 0, \quad (12.1.1)$$

где a, b, x_0, y_0 – заданные числа, и $b \neq 0$.

Очевидно, что это уравнение при любом x однозначно разрешимо относительно y , т.е. задает *неявно* функцию $f : \mathbb{R} \rightarrow \mathbb{R}$. Нетрудно получить явное выражение для этой функции:

$$y = f(x) = y_0 - b^{-1}a \cdot (x - x_0) \quad (\text{заметьте, что } f(x_0) = y_0).$$

Теперь перейдем к уравнению общего вида $F(x, y) = 0$ и предположим, что пара чисел (x_0, y_0) есть решение этого уравнения, т.е. $F(x_0, y_0) = 0$.

Используя для функции F формулу Тейлора³⁸ (11.7.5) и полагая

$$a = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}, \quad h = \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}, \quad a + h = \begin{bmatrix} x \\ y \end{bmatrix},$$

перепишем наше уравнение в виде

$$F(x, y) = F(x_0, y_0) + F'(x_0, y_0) \cdot \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix} + \alpha(x, y) = 0,$$

или

$$D_x F(x_0, y_0) \cdot (x - x_0) + D_y F(x_0, y_0) \cdot (y - y_0) + \alpha(x, y) = 0, \quad (12.1.2)$$

где $\alpha(x, y) = \frac{1}{2} \cdot h^T \cdot F''(a + \xi e) \cdot h$ – остаточный член формулы Тейлора.

Мы видим, что уравнение (12.1.2) отличается от уравнения (12.1.1) только остаточным членом формулы Тейлора. Поскольку при $F'(x_0, y_0) \neq \theta$ и (x, y) , близких к (x_0, y_0) , остаточный член $\alpha(x, y)$ мал по сравнению с первыми слагаемыми, можно ожидать, что при $D_y F(x_0, y_0) \neq 0$ уравнение (12.1.2), как и (12.1.1) будет однозначно разрешимо относительно y при любом x хотя бы в малой окрестности точки (x_0, y_0) .

Действительно, можно показать, что справедлива

Теорема. Пусть на открытом прямоугольнике $J =]a, b[\times]c, d[$ задана непрерывно дифференцируемая функция $F : J \rightarrow \mathbb{R}$. Пусть $(x_0, y_0) \in J$, причем $F(x_0, y_0) = 0$, а $D_y F(x_0, y_0) \neq 0$.

Тогда существуют такие $J_x \subset]a, b[$ (окрестность точки x_0) и $J_y \subset]c, d[$ (окрестность точки y_0), что

1) при любом $x \in J_x$ уравнение $F(x, y) = 0$ имеет единственное решение $y \in J_y$ (обратите внимание: не вообще единственное решение, а единственное решение, лежащее в J_y !). Таким образом, уравнение $F(x, y) = 0$ неявно задает функцию $f : J_x \rightarrow J_y$;

2) $F(x, f(x)) = 0$ при всех $x \in J_x$;

3) $f(x_0) = y_0$;

4) f непрерывно дифференцируема на J_x .

³⁸В соответствии с **Важным соглашением** мы считаем, что функция имеет все нужные нам производные.

Геометрический смысл этой теоремы очень прост: в некоторой окрестности точки (x_0, y_0) множество решений уравнения $F(x, y) = 0$ есть график гладкой функции f .

Примеры. 1. (рис.12.1а). Пусть $F(x, y) = x^2 + y^2 - 1$, и (x_0, y_0) – некоторая точка, лежащая на верхней дуге единичной окружности, т.е. $F(x_0, y_0) = 0, y_0 > 0$. Тогда $D_y F(x_0, y_0) = 2y_0 \neq 0$. Из рисунка видно, что в некоторой окрестности точки (x_0, y_0) множество решений уравнения $x^2 + y^2 - 1 = 0$ представляет собой график функции $y = f(x)$, т.е. для каждого x , достаточно близкого к x_0 , существует ровно один y , близкий к y_0 , такой, что $x^2 + y^2 - 1 = 0$. Видно, что если не требовать близости y к y_0 , то нашлось бы не одно, а два решения уравнения: y и $-y$. Фрагмент графика функции $y = f(x)$ изображен на рис.12.1б.

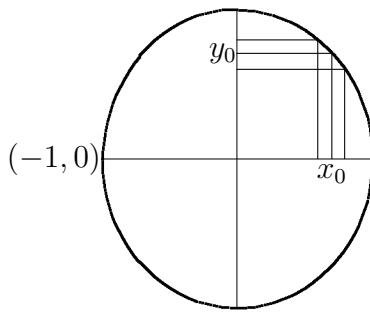


Рис.12.1а

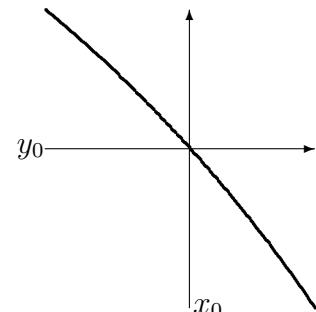


Рис.12.1б

К сожалению, теорема не дает сведений об окрестностях J_x и J_y . Очевидно, что на рисунке они взяты не наибольшими из возможных.

Заметим, что в нашем примере нетрудно найти явное выражение функции f : $f(x) = +\sqrt{1 - x^2}$ (здесь учтено, что $y_0 > 0$).

В точке же $(-1, 0)$ $D_y F = 0$. Из рисунка видно, что в любой окрестности этой точки множество решений уравнения не является графиком функции.

2. Пусть $F(x, y) = x^7 - xy + y^6 - 1$, $x_0 = 0$, $y_0 = 1$. Тогда $F(0, 1) = 0$, $D_y F(0, 1) \neq 0$. В силу сформулированной выше теоремы уравнение $x^7 - xy + y^6 - 1 = 0$ определяет в некоторой окрестности точки $(0, 1)$ функцию $y = f(x)$, однако нетрудно видеть, что ее явное задание получить не удастся. Фрагмент графика этой функции изображен на рис.12.2.

Замечание. В случае $D_y F(x_0, y_0) = 0$ теорема не дает никаких заключений о структуре множества решений уравнения в окрестности точки (x_0, y_0) . Приведем несколько простых примеров (во всех $x_0 = y_0 = 0$ и $D_y F(0, 0) = 0$).

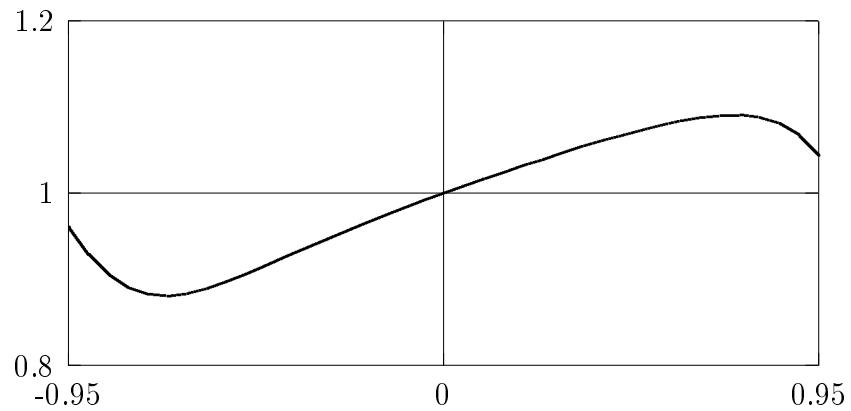


Рис.12.2

1. $x - y^2 = 0$ (рис.12.3). Здесь при $x > 0$ уравнение имеет два решения, при $x = 0$ – одно, а при $x < 0$ – ни одного.

2. $x - y^3 = 0$ (рис.12.4). Здесь при всех $x \in \mathbb{R}$ уравнение имеет одно решение.

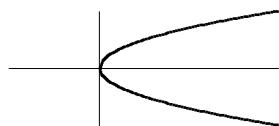


Рис.12.3

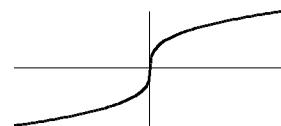


Рис.12.4

3. $x^2 - y^2 = 0$ (рис.12.5). Здесь при всех $x \neq 0$ уравнение имеет два решения.

4. $x^2 + y^2 = 0$ (рис.12.6). Здесь при любом $x \neq 0$ уравнение не имеет ни одного решения.

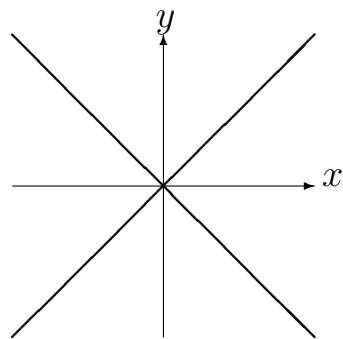


Рис.12.5

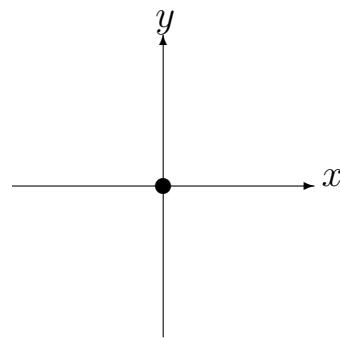


Рис.12.6

Обратите внимание на то, что множества решений уравнений в примерах 1, 3, 4 не являются графиками функций *ни в каком*

открытом прямоугольнике, содержащем начало координат, в то время как в примере 2 уравнение задает неявно функцию без каких-либо ограничений на переменные (нетрудно найти ее явное выражение: $f(x) = \text{sign}(x) \cdot |x|^{1/3}$).

Пусть теперь условия теоремы выполнены, и уравнение $F(x, y) = 0$ задает неявно в некоторой окрестности точки x_0 функцию $y = f(x)$ со значениями из окрестности точки y_0 . Получим формулу для вычисления производной неявно заданной функции f (существование этой производной гарантировано теоремой).

Подставив в уравнение $F(x, y) = 0$ его решение $y = f(x)$, получим тождество

$$(F \circ \varphi)(u) = F(u, f(u)) \equiv 0, \quad (12.1.3)$$

где отображение φ (см. рис.12.7) определено на J_x .

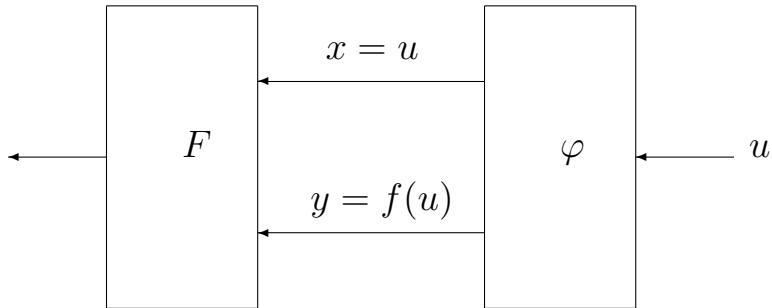


Рис.12.7

Дифференцируем тождество (12.1.3):

$$(F' \circ \varphi)(u) \cdot \varphi'(u) \equiv 0.$$

Подставив сюда

$$F'(x, y) = [D_x F(x, y), D_y F(x, y)], \quad \varphi'(u) = \begin{bmatrix} 1 \\ f'(u) \end{bmatrix},$$

получим

$$D_x F(u, f(u)) + D_y F(u, f(u)) \cdot f'(u) \equiv 0,$$

откуда

$$f'(u) = -\left(D_y F(u, f(u))\right)^{-1} \cdot D_x F(u, f(u)) = -\frac{D_x F(u, f(u))}{D_y F(u, f(u))}. \quad (12.1.4)$$

Замечания. 1. В формуле (12.1.4) деление на $D_y F(u, f(u))$ возможно, так как по условию теоремы $D_y F$ отлична от нуля в точке (x_0, y_0) , а следовательно, по непрерывности, и в некоторой ее окрестности (вне которой эта формула, естественно, не работает).

2. Для фактического вычисления значения производной неявно заданной функции f по формуле (12.1.4), необходимо предварительно найти значение этой функции, т.е. решить (например, численно) уравнение $F(x, y) = 0$.

Вернемся к нашим примерам. Если $F(x, y) = x^2 + y^2 - 1$, $x > 0$, $y > 0$, то $D_x F(x, y) = 2x$, $D_y F(x, y) = 2y$. Поэтому $f'(x) = -\frac{2x}{2y} = -\frac{x}{f(x)}$.

Здесь можно найти явное выражение для функции: $f(x) = +\sqrt{1 - x^2}$. Поэтому $f'(x) = -\frac{x}{\sqrt{1 - x^2}}$, что легко проверить непосредственно.

Если же $F(x, y) = x^7 - xy + y^6 - 1$, то

$$D_x F(x, y) = 7x^6 - y, \quad D_y F(x, y) = -x + 6y^5, \quad f'(x) = -\frac{7x^6 - y}{-x + 6y^5}.$$

Здесь получить явное выражение для функции невозможно, однако при $x = 0$ имеем $y = 1$ и

$$f'(0) = -\left. \frac{7x^6 - y}{-x + 6y^5} \right|_{x=0, y=1} = \frac{1}{6}.$$

Обратимся теперь к неявному заданию *векторного поля*. Пусть задана система из n уравнений с $n + m$ переменными:

$$F(x, y) = \theta_n. \tag{12.1.5}$$

где $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$, $F(x, y) = [F_1(x, y), \dots, F_n(x, y)]^T$ – функция из $\mathbb{R}^m \times \mathbb{R}^n$ в \mathbb{R}^n .

Если существуют такие множества $U \subset \mathbb{R}^m$ и $V \subset \mathbb{R}^n$, что для всякого $x \in U$ система (12.1.5) имеет ровно одно решение $y \in V$, то говорят, что эта система задает *неявно* функцию $f : U \rightarrow V$. Очевидно, что при всех $x \in U$ $F(x, f(x)) = \theta_n$.

Вновь начнем с рассмотрения *линейной* системы

$$A \cdot (x - x^{(0)}) + B \cdot (y - y^{(0)}) = \theta_n, \tag{12.1.6}$$

где A и B – матрицы размеров $n \times m$ и $n \times n$ соответственно.

Очевидно, что при $\det(B) \neq 0$ система (12.1.6) однозначно разрешима относительно y и задает неявно функцию $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$. Нетрудно получить явное задание этой функции:

$$y = f(x) = y^{(0)} - B^{-1}A \cdot (x - x^{(0)}); \quad f(x^{(0)}) = y^{(0)}.$$

Пусть теперь дана система общего вида (12.1.5), и известно ее решение – упорядоченная пара векторов $(x^{(0)}, y^{(0)}) : F(x^{(0)}, y^{(0)}) = \theta_n$.

Используя формулу Тейлора (11.7.7), запишем систему (12.1.5) в виде

$$F'(x^{(0)}, y^{(0)}) \cdot \begin{bmatrix} x - x^{(0)} \\ y - y^{(0)} \end{bmatrix} + \alpha(x, y) = \theta_n, \quad (12.1.7)$$

где

$F'(x^{(0)}, y^{(0)}) = [D_x F(x^{(0)}, y^{(0)}), D_y F(x^{(0)}, y^{(0)})]$ – матрица Якоби размера $n \times (m + n)$ (ее блоки – матрицы $D_x F$ и $D_y F$, составленные из частных производных компонент F по координатам векторов x и y соответственно, – имеют размеры $n \times m$ и $n \times n$);

$\begin{bmatrix} x - x^{(0)} \\ y - y^{(0)} \end{bmatrix}$ – столбец высоты $m + n$;

$\alpha(x, y)$ – остаточный член формулы Тейлора – столбец высоты n .

Расписав формулу (12.1.7) подробнее:

$$D_x F(x^{(0)}, y^{(0)}) \cdot (x - x^{(0)}) + D_y F(x^{(0)}, y^{(0)}) \cdot (y - y^{(0)}) + \alpha(x, y) = \theta_n,$$

видим, что она отличается от (12.1.6) только остаточным членом $\alpha(x, y)$. Поэтому можно ожидать, что при

$$\det(B) = \det(D_y F(x^{(0)}, y^{(0)})) \neq 0$$

система при заданном значении x будет однозначно разрешима относительно y в достаточно малой окрестности точки $(x^{(0)}, y^{(0)})$.

Сформулируем соответствующую теорему.

Теорема. Пусть

$J_m =]a_1, b_1[\times]a_2, b_2[\times \dots \times]a_m, b_m[$ – параллелепипед в \mathbb{R}^m ;

$J_n =]c_1, d_1[\times]c_2, d_2[\times \dots \times]c_n, d_n[$ – параллелепипед в \mathbb{R}^n ;

$J_{mn} = J_m \times J_n$ – параллелепипед в \mathbb{R}^{m+n} ;

$F : J_{mn} \rightarrow \mathbb{R}^n$ – непрерывно дифференцируемая вектор-функция;

$(x^{(0)}, y^{(0)}) \in J_{mn}$, $F(x^{(0)}, y^{(0)}) = \theta_n$, $\det(D_y F(x^{(0)}, y^{(0)})) \neq 0$.

Тогда существуют такие $U \subset J_m$ – окрестность точки $x^{(0)}$ и $V \subset J_n$ – окрестность точки $y^{(0)}$, что

- 1) для каждого $x \in U$ существует единственный $y \in V$ – решение системы $F(x, y) = \theta_n$, т.е. эта система неявно задает функцию $f : U \rightarrow V$;
- 2) $F(x, f(x)) = \theta_n$ при всех $x \in U$;
- 3) $f(x^{(0)}) = y^{(0)}$;
- 4) f непрерывно дифференцируема на U .

Для вычисления матрицы Якоби неявно заданной функции f рассмотрим вновь рис.12.7 (отображение φ при этом определено на U). Дифференцируя тождество $(F \circ \varphi)(u) = F(u, f(u)) \equiv \theta_n$, выполняющееся на U , получим

$$(F \circ \varphi)'(u) = (F' \circ \varphi)(u) \cdot \varphi'(u) = \Theta_{n \times m}$$

$(\Theta_{n \times m}$ – нуль-матрица размера $n \times m$).

Подставляя сюда

$$\varphi'(u) = \begin{bmatrix} u \\ f(u) \end{bmatrix}' = \begin{bmatrix} I_m \\ f'(u) \end{bmatrix}$$

$(I_m$ – единичная матрица порядка m), получим

$$D_x F(u, f(u)) + D_y F(u, f(u)) \cdot f'(u) = \Theta_{n \times m}$$

(проверьте согласованность размеров матриц!), откуда

$$f'(u) = - \left(D_y F(u, f(u)) \right)^{-1} \cdot D_x F(u, f(u)). \quad (12.1.8)$$

Замечания. 1. Поскольку $\det(D_y F)$ не равен нулю в точке (x_0, y_0) , по непрерывности он отличен от нуля и в некоторой ее окрестности, в которой, следовательно, обратима матрица $D_y F$ и применима формула (12.1.9). Остается также в силе замечание 2 на стр.128.

2. В реальной задаче переменные, конечно, не поделены изначально на независимые ("иксы") и зависимые ("игреки"). Постановщик задачи сам выбирает те переменные, относительно которых он хочет разрешить систему, и проверяет выполнение условий теоремы.

Пример. Пусть

$$m = n, \quad x, y \in \mathbb{R}^n, \quad g : \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad F(x, y) = x - g(y).$$

Тогда уравнением $x - g(y) = \theta_n$ задается неявно функция g^{-1} , обратная к функции g .

Поскольку $D_y F = -g'$, теорему о неявно заданной функции можно переформулировать в этом частном случае так:

Пусть $J_n =]c_1, d_1[\times \dots \times]c_n, d_n[$ – параллелепипед в \mathbb{R}^n ; $g : J_n \rightarrow \mathbb{R}^n$, $y^{(0)} \in J_n$, $\det(g'(y^{(0)})) \neq 0$.

Тогда в некоторой окрестности точки $y^{(0)}$ функция g обратима, обратная ей функция g^{-1} непрерывно дифференцируема, и

$$(g^{-1})' = [g' \circ g^{-1}]^{-1}. \quad (12.1.9)$$

В случае $n = 1$ эта формула дает известный результат: $(g^{-1})' = \frac{1}{g' \circ g^{-1}}$.

Замечание. В формуле (12.1.9) имеет место "коллизия символов": g^{-1} обозначает функцию, обратную функции g , а $[g' \circ g^{-1}]^{-1}$ – матрицу, обратную матрице $[g' \circ g^{-1}]$.

12.2. Задание гладкой поверхности в \mathbb{R}^3 уравнением

Рассмотрим уравнение

$$F(x) = c, \quad (12.2.1)$$

где $F : \mathbb{R}^3 \rightarrow \mathbb{R}$ – непрерывно дифференцируемый функционал, а c – число. Напомним, что множество решений этого уравнения называется *множеством уровня* функционала F .

Предположим, что точка a удовлетворяет уравнению (12.2.1), и $\nabla F(a) \neq \theta$. Тогда, не умаляя общности, можно считать, например, что $D_3 F(a) \neq 0$. Применяя к уравнению теорему о неявно заданной функции (здесь $m = 2$, $n = 1$), получаем, что в некоторой окрестности точки a уравнение можно *однозначно* разрешить относительно x_3 , т.е. множество уровня функционала в этой окрестности представляет собой график функции $x_3 = f(x_1, x_2)$, и

$$F(x_1, x_2, f(x_1, x_2)) - c \equiv 0, \quad f(a_1, a_2) = a_3.$$

Функция f непрерывно дифференцируема, и, следовательно (см. п.11.3), ее график есть гладкая поверхность. Проведем в точке a касательную плоскость к этой поверхности.

Из формулы (12.1.8) получаем

$$f'(a_1, a_2) = -D_3^{-1} F(a_1, a_2, a_3) \cdot [D_1 F(a_1, a_2, a_3), D_2 F(a_1, a_2, a_3)].$$

Отсюда

$$D_1 f(a_1, a_2) = -\frac{D_1 F(a)}{D_3 F(a)}, \quad D_2 f(a_1, a_2) = -\frac{D_2 F(a)}{D_3 F(a)}.$$

Подставляя результат в формулу (11.3.2), получим

$$x_3 - a_3 = -\frac{D_1 F(a)}{D_3 F(a)} \cdot (x_1 - a_1) - \frac{D_2 F(a)}{D_3 F(a)} \cdot (x_2 - a_2)$$

или, после умножения на $D_3 F(a) \neq 0$,

$$D_1 F(a) \cdot (x_1 - a_1) + D_2 F(a) \cdot (x_2 - a_2) + D_3 F(a) \cdot (x_3 - a_3) = 0,$$

или

$$F'(a) \cdot (x - a) = 0,$$

или, наконец,

$$\langle \nabla F(a), (x - a) \rangle = 0. \quad (12.2.2)$$

В силу симметрии этого уравнения относительно координат оно будет выполняться и в том случае, когда не третья, а какая-нибудь другая компонента $\nabla F(a)$ отлична от нуля.

Итак, если $F(a) = c$ и $\nabla F(a) \neq \theta$, то в некоторой окрестности точки a множество уровня функционала есть гладкая поверхность. Если же градиент отличен от нуля во всех точках множества уровня функционала, то естественно сказать, что все это множество есть гладкая поверхность (ее называют *поверхностью уровня функционала F*).

Примеры. 1. $F(x_1, x_2, x_3) \equiv x_1^2 + x_2^2 + x_3^2 = 1$.

$\nabla F(x) = 2x$, очевидно, не обращается в нуль ни в одной точке *множества уровня*. Как известно, это множество есть сфера – гладкая поверхность (см. пример 2 п.11.3).

2. $x_1^2 + x_2^2 + x_3^2 = 0$.

Это – другое множество уровня того же функционала. Очевидно, оно состоит из одной точки θ , причем $\nabla F(\theta) = \theta$.

3. $F(x_1, x_2, x_3) \equiv x_1^2 + x_2^2 - x_3^2 = -1$.

$\nabla F(x) = 2[x_1, x_2, -x_3]^T$. Градиент, очевидно, не обращается в нуль ни в одной точке *множества уровня*. Как известно из курса линейной алгебры, это множество – двухполостный гиперболоид – поверхность, "гладкая" и в нематематическом, обыденном смысле.

$$2. \ x_1^2 + x_2^2 - x_3^2 = 0.$$

Это – другое множество уровня того же функционала. Теперь начало координат (точка, в которой ∇F обращается в нуль) принадлежит множеству уровня. Теорема о неявно заданной функции не работает в окрестности этой точки. Как известно из курса линейной алгебры, рассматриваемое множество уровня – конус, "негладкий" в вершине. Во всех остальных своих точках он является гладкой поверхностью – как в смысле нашего определения, так и в обыденном смысле.

Единичный направленный отрезок, перпендикулярный касательной плоскости, называется *нормалью* к поверхности в точке касания. Из (12.2.2) видно, что $\overrightarrow{\nabla F(a)}$ перпендикулярен касательной плоскости, т.е. коллинеарен нормали. Говорят, что *градиент функционала ортогонален поверхности уровня этого функционала*.

12.3. Условный экстремум

В приложениях часто возникают задачи о нахождении экстремума функционала при некоторых дополнительных условиях. Одну из таких задач мы рассмотрим в этом пункте.

Пусть $F, G : \mathbb{R}^3 \rightarrow \mathbb{R}$ – непрерывно дифференцируемые функционалы. Требуется найти локальные экстремумы функционала F на гладкой поверхности уровня функционала G , заданной уравнением $G(x) = c$.

Пусть точка a лежит на этой поверхности уровня. Вследствие гладкости поверхности $\nabla G(a) \neq 0$, и мы можем разрешить уравнение $G(x) = c$ в окрестности точки a относительно хотя бы одной из координат. Пусть, например, $D_3 G(a) \neq 0$. Тогда поверхность $G(x) = c$ в окрестности точки a представляет собой график функции $x_3 = f(x_1, x_2)$. Поэтому наличие в точке a локального экстремума у функционала F на поверхности уровня функционала G равносильно наличию локального экстремума у функционала

$$\phi(x_1, x_2) = F(x_1, x_2, f(x_1, x_2))$$

в точке $(a_1, a_2) \in \mathbb{R}^2$.

Как известно, точка локального экстремума гладкого функционала ϕ обязана быть его стационарной точкой. Таким образом $\nabla \phi(a_1, a_2) = 0$.

Из рис.12.8 видно, что $\phi' = (F \circ \varphi)' = (F' \circ \varphi) \cdot \varphi'$.

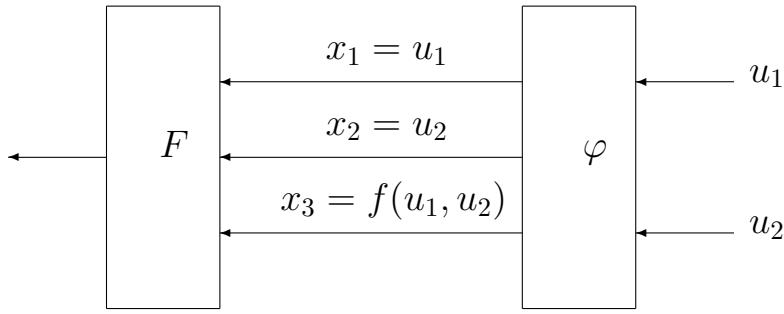


Рис.12.8

Поскольку $(F' \circ \varphi)(u_1, u_2) = F'(u_1, u_2, f(u_1, u_2))$ и

$$\varphi'(u) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ D_1f(u) & D_2f(u) \end{bmatrix},$$

имеем

$$\begin{aligned} \nabla\phi(a_1, a_2) &= (\phi'(a_1, a_2))^T = \\ &= \begin{bmatrix} D_1F(a) \cdot 1 + D_2F(a) \cdot 0 + D_3F(a) \cdot D_1f(a_1, a_2) \\ D_1F(a) \cdot 0 + D_2F(a) \cdot 1 + D_3F(a) \cdot D_2f(a_1, a_2) \end{bmatrix}. \end{aligned}$$

Отсюда

$$D_1F(a) + D_3F(a) \cdot D_1f(a_1, a_2) = D_2F(a) + D_3F(a) \cdot D_2f(a_1, a_2) = 0$$

Вспоминая формулы (12.2.2), получаем

$$D_1F(a) - D_3F(a) \cdot \frac{D_1G(a)}{D_3G(a)} = D_2F(a) - D_3F(a) \cdot \frac{D_2G(a)}{D_3G(a)} = 0$$

или

$$[D_1F(a), D_2F(a), D_3F(a)] = \frac{D_3F(a)}{D_3G(a)} \cdot [D_1G(a), D_2G(a), D_3G(a)],$$

или

$$\nabla F(a) = \lambda \cdot \nabla G(a), \quad (12.3.1)$$

где $\lambda = \frac{D_3F(a)}{D_3G(a)}$ – число.

Если предположить, что отлична от нуля другая компонента вектора $\nabla G(a)$, то после аналогичных преобразований придем опять к соотношению (12.3.1).

Таким образом, если векторы $\nabla F(a)$ и $\nabla G(a)$ линейно независимы, то в точке a функционал F не может иметь локального экстремума на поверхности уровня функционала G . Точки поверхности $G(x) = c$, в которых выполнено условие (12.3.1), называются *стационарными точками* функционала F на этой поверхности.

Можно показать, что аналогичный результат имеет место в пространстве любой размерности: если $F, G : \mathbb{R}^n \rightarrow \mathbb{R}$ – непрерывно дифференцируемые функционалы, а уравнение $G(x) = c$ задает гладкую "поверхность" (т.е. $\nabla G \neq 0$ на множестве $\{x \mid G(x) = c\}$), то в точках этой поверхности, где градиенты ∇F и ∇G линейно независимы, функционал F не может иметь локального экстремума на поверхности уровня функционала G .

Это утверждение называется *теоремой Эйлера*.

По историческим причинам рассмотренную задачу обычно называют задачей об экстремуме функционала F при дополнительном условии $G(x) = c$, или задачей об *условном экстремуме*. Важно понимать, что это просто задача об экстремуме нового функционала (изменяется *область определения* функционала F)!

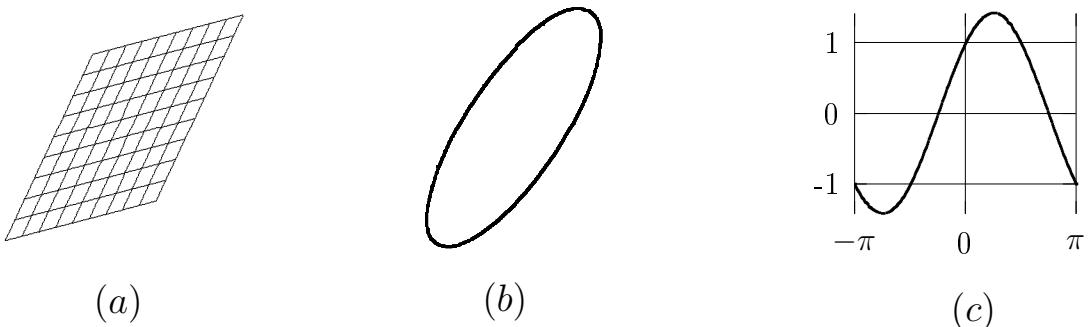


Рис.12.9

Пример (см. рис.12.9). Известно, что линейный функционал

$$F : \mathbb{R}^2 \rightarrow \mathbb{R}; \quad F(x_1, x_2) = x_1 + x_2$$

(фрагмент его графика – наклонной плоскости – изображен на рис.12.9a) не имеет локальных экстремумов. Однако если рассматривать поведение этого функционала не на всей координатной плоскости, а на окружности $G(x) = x_1^2 + x_2^2 = 1$, то экстремумы появляются.

Действительно, параметрические уравнения нашей окружности

$$x_1 = \cos(t), \quad x_2 = \sin(t); \quad t \in [-\pi, \pi],$$

график сужения F на нее (пространственная кривая) изображен на рис.12.9б, а его развертка – на рис.12.9с.

Теорема Эйлера имеет в \mathbb{R}^2 и \mathbb{R}^3 простую геометрическую интерпретацию. Ранее (п.11.7) было показано, что функционал F не может иметь локального ("безусловного") экстремума в точке a , в которой его градиент отличен от нуля, так как F возрастает в направлении градиента $\nabla F(a)$ и убывает в направлении антиградиента $-\nabla F(a)$. В задаче об условном экстремуме ситуация осложняется. Рассмотрим ситуации, схематически изображенные на рис.12.10.

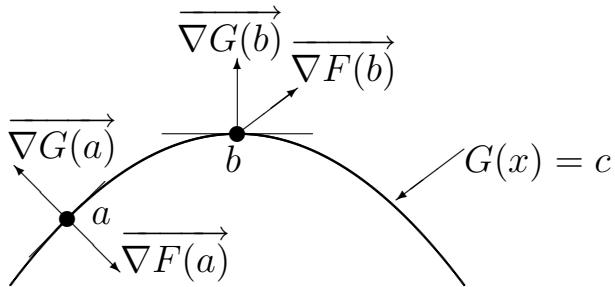


Рис.12.10

Мы не можем "выйти" из точки a в произвольном направлении, а можем "двигаться" только по поверхности $G(x) = c$. Поэтому градиент функционала не обязан обращаться в нуль в точке условного экстремума этого функционала. Но если (как в точке b на рис.12.10) проекция $\overrightarrow{\nabla F(b)}$ на касательную плоскость не обращается в нуль, то можно "пойти" из точки b по поверхности в направлении этой проекции, и функция будет возрастать, а в противоположном направлении – убывать. Следовательно, в окрестности точки b функционал F принимает на поверхности уровня значения и большие и меньшие, чем $F(b)$, т.е. условного экстремума в этой точке нет. Обращение же (как в точке a) в нуль проекции $\overrightarrow{\nabla F(a)}$ на касательную плоскость означает ортогональность $\overrightarrow{\nabla F(a)}$ поверхности уровня функционала G . Поскольку, как известно (п.12.2), $\overrightarrow{\nabla G(a)}$ заведомо ортогонален этой поверхности, получаем, что $\overrightarrow{\nabla F(a)}$ коллинеарен $\overrightarrow{\nabla G(a)}$, что и приводит к (12.3.1). Подчеркнем, что это рассуждение – не доказательство теоремы Эйлера, а иллюстрация к ней.

Соотношение (12.3.1) может быть переписано так:

$$\nabla(F - \lambda G) = \theta,$$

т.е. стационарные точки функционала F на поверхности уровня функционала G – это в точности стационарные точки функционала $F - \lambda G$ (с дополнительной переменной λ , называемой *множителем Лагранжа*).

Замечание. Для поиска стационарных точек функционала F , как известно (п.11.7), следует решить систему из n уравнений с n переменными: $\nabla F(x) = \theta_n$. Поиск стационарных точек функционала F на поверхности уровня функционала G приводит к системе из n уравнений

$$\nabla(F - \lambda G)(x) = \theta_n \quad (12.3.2)$$

с $n + 1$ переменными: $x_1, \dots, x_n; \lambda$.

Кажется, что у нас не хватает одного уравнения. Однако соотношение (12.3.2) должно выполняться в точке условного экстремума, независимо от того, на какой именно поверхности уровня функционала G мы ищем экстремум функционала F . Поэтому к системе (12.3.2) мы должны добавить исходное условие $G(x) = c$, которое и будет "недостающим" уравнением.

Примеры. 1. Найти расстояние от точки $x^{(0)} \in \mathbb{R}^3$ до плоскости Π : $\langle x, a \rangle = c$, $a \neq \theta_3$, не проходящей через эту точку.

По определению, $dist(x^{(0)}, \Pi) = \min_{x \in \Pi} \|x^{(0)} - x\|$. Минимизируя, как обычно, квадрат нормы, получим задачу на условный экстремум: найти минимум функционала $\|x^{(0)} - x\|^2$ при условии $\langle x, a \rangle = c$.

Составим уравнение (12.3.1):

$$\nabla \|x^{(0)} - x\|^2 = \lambda \cdot \nabla \langle x, a \rangle, \quad \text{или} \quad 2 \cdot (x - x^{(0)}) = \lambda \cdot a. \quad (12.3.3)$$

Как известно, $\overrightarrow{a} \perp \Pi$. Отсюда и из (12.3.3) следует, что $\overrightarrow{x - x^{(0)}} \perp \Pi$, и мы получили известный результат: из всех отрезков, соединяющих данную точку с точками плоскости, наименьшую длину имеет перпендикуляр. (На самом деле мы доказали только, что *никакой другой отрезок не может дать минимума*; то, что перпендикуляр дает минимум, требует еще проверки!).

2. Пусть A – вещественная симметричная матрица порядка n . Найти экстремумы квадратичной формы $\langle Ax, x \rangle$ на сфере $\|x\|^2 = r^2$.

Поскольку (см. пример в п.11.6) $\nabla \langle Ax, x \rangle = 2Ax$, $\nabla \|x\|^2 = 2x$, система (12.3.2) приобретает вид $(A - \lambda I_n)x = \theta_n$.

Таким образом, стационарные точки квадратичной формы *на сфере* – собственные векторы матрицы этой квадратичной формы, а множители Лагранжа – соответствующие собственные числа. Из курса линейной алгебры известно, что *глобальные* экстремумы квадратичной формы на сфере действительно достигаются на собственных векторах, соответствующих наибольшему и наименьшему собственным числам. *Можно показать, что* собственные векторы, соответствующие остальным собственным числам, не дают квадратичной форме даже локальных условных экстремумов.

Глава 13. ВЫЧИСЛИТЕЛЬНЫЕ АЛГОРИТМЫ, ИСПОЛЬЗУЮЩИЕ ПРОИЗВОДНЫЕ

13.1. Решение нелинейных уравнений и систем методом простой итерации

Рассмотрим уравнение, записанное в специальном виде

$$x = f(x), \quad (13.1.1)$$

где f – непрерывно дифференцируемая функция из \mathbb{R} в \mathbb{R} .

Теорема 1. Если $|f'(x)| \leq q < 1$ при всех $x \in \mathbb{R}$, то

- 1) уравнение (13.1.1) имеет единственное решение (обозначим его \tilde{x});
- 2) для любого $x_0 \in \mathbb{R}$ последовательность

$$x_1 = f(x_0), \dots, x_k = f(x_{k-1}), \dots \quad (13.1.2)$$

сходится к \tilde{x} ;

- 3) $|x_k - \tilde{x}| \leq q^k \cdot |x_0 - \tilde{x}|$.

Доказательство. Заметим для начала, что по формуле Лагранжа (п.10.2) для всех $x, y \in \mathbb{R}$

$$|f(x) - f(y)| = |f'(\xi) \cdot (x - y)| \leq q \cdot |x - y| \quad (13.1.3)$$

(здесь ξ – некоторая точка между x и y).

Докажем *существование* решения. Если $f(0) = 0$, то доказывать нечего. Если $f(0) \neq 0$, то из (13.1.3) имеем $|f(x) - f(0)| \leq q \cdot |x|$, или

$$f(0) - q \cdot |x| - x \leq f(x) - x \leq f(0) + q \cdot |x| - x. \quad (13.1.4)$$

При $\hat{x}_1 = |f(0)|/(1 - q) > 0$ правая часть неравенства (13.1.4) даст

$$f(\hat{x}_1) - \hat{x}_1 \leq f(0) + (q - 1) \cdot |f(0)|/(1 - q) = f(0) - |f(0)| \leq 0.$$

При $\hat{x}_2 = -|f(0)|/(1 - q) < 0$ левая часть неравенства (13.1.4) даст

$$f(\hat{x}_2) - \hat{x}_2 \geq f(0) + (1 - q) \cdot |f(0)|/(1 - q) = f(0) + |f(0)| \geq 0.$$

Теорема Коши (п.9.2) гарантирует существование хотя бы одного решения уравнения $f(x) - x = 0$ на сегменте $[\hat{x}_2, \hat{x}_1]$.

Теперь докажем *единственность* решения. Если \tilde{x}_1 и \tilde{x}_2 – два решения уравнения (13.1.1), то

$$|\tilde{x}_1 - \tilde{x}_2| = |f(\tilde{x}_1) - f(\tilde{x}_2)| \leq q \cdot |\tilde{x}_1 - \tilde{x}_2| \implies |\tilde{x}_1 - \tilde{x}_2| \cdot (1 - q) \leq 0,$$

откуда $\tilde{x}_1 = \tilde{x}_2$.

Далее, для любого $x_0 \in \mathbb{R}$ значения последовательности (13.1.2) удовлетворяют неравенству

$$|x_k - \tilde{x}| = |f(x_{k-1}) - f(\tilde{x})| \leq q \cdot |x_{k-1} - \tilde{x}| \leq \dots \leq q^k \cdot |x_0 - \tilde{x}|,$$

т.е. итерации сходятся к единственному решению уравнения не медленнее, чем геометрическая прогрессия со знаменателем q . ■

Замечания. 1. Отображение, удовлетворяющее условию (13.1.3), называется *сжимающим отображением* или *сжатием*. Из доказательства теоремы видно, что ее утверждение верно для любого сжимающего отображения $\mathbb{R} \rightarrow \mathbb{R}$. Выполнение условия $|f'(x)| \leq q < 1$ гарантирует, что f – сжатие. Решение уравнения (13.1.1) называется *неподвижной точкой* отображения f .

2. Условие $|f'(x)| \leq q < 1$ в теореме нельзя заменить условием $|f'(x)| < 1$. Приведем простой пример.

Пусть $f(x) = (x^2 + 1)^{1/2}$. Тогда $f'(x) = \frac{x}{(x^2 + 1)^{1/2}}$. Поскольку $|x| < (x^2 + 1)^{1/2}$, имеем $|f'(x)| < 1$. Однако уравнение $x = (x^2 + 1)^{1/2}$, очевидно, не имеет решений.

Практическая ценность доказанной теоремы невелика – ограничения, накладываемые на функцию f , слишком жесткие. Поэтому на практике удобной оказывается следующая *локальная* ее модификация.

Теорема 2. Пусть \tilde{x} – решение уравнения (13.1.1), причем $|f'(\tilde{x})| < 1$. Тогда существует такая окрестность точки \tilde{x} , что при x_0 , взятом из этой окрестности, последовательность (13.1.2) сходится к \tilde{x} .

Доказательство. Пусть $|f'(\tilde{x})| = q < 1$. В силу непрерывности f' найдется такая окрестность точки \tilde{x} , в которой $|f'(x)| \leq \tilde{q} = \frac{1+q}{2} < 1$. Если взять точку x_0 из этой окрестности и построить последовательность (13.1.2), то

$$|x_1 - \tilde{x}| = |f(x_0) - f(\tilde{x})| \leq \tilde{q} \cdot |x_0 - \tilde{x}|.$$

Отсюда следует, что x_1 лежит в этой же окрестности. Поэтому

$$|x_2 - \tilde{x}| \leq \tilde{q} \cdot |x_1 - \tilde{x}| \leq \tilde{q}^2 \cdot |x_0 - \tilde{x}|, \text{ и т.д.}$$

Таким образом, если $|f'(\tilde{x})| < 1$, то итерационный процесс (13.1.2), начатый из любой точки, лежащей в некоторой окрестности точки \tilde{x} (решения уравнения), сходится к этому решению. Неподвижную точку \tilde{x} отображения f называют в этом случае *точкой притяжения* итерационного процесса (13.1.2). ■

Если же \tilde{x} – решение уравнения (13.1.1), но $|f'(\tilde{x})| > 1$, то в некоторой окрестности точки \tilde{x} имеем $|f'(x)| > 1$, и взяв $x_0 \neq \tilde{x}$ из этой окрестности, получим

$$|x_1 - \tilde{x}| = |f(x_0) - f(\tilde{x})| = |f'(\xi) \cdot (x_0 - \tilde{x})| > |x_0 - \tilde{x}|.$$

Теперь последовательность итераций "выталкивается" из окрестности точки \tilde{x} (решения уравнения). Такую неподвижную точку отображения f называют *точкой отталкивания* итерационного процесса (13.1.2).

Пример. Пусть $f(x) = \exp(ax)$, $a > 0$. Уравнение $x = \exp(ax)$ имеет при $a < 1/e$ два решения (рис.13.1). При этом $0 < f'(x_1) < 1$, $f'(x_2) > 1$, т.е. x_1 – точка притяжения, а x_2 – точка отталкивания итерационного процесса. В этом нетрудно убедиться, выполнив несколько шагов (хотя бы с помощью микрокалькулятора). Проделайте это и найдите x_1 для какого-нибудь конкретного значения a .

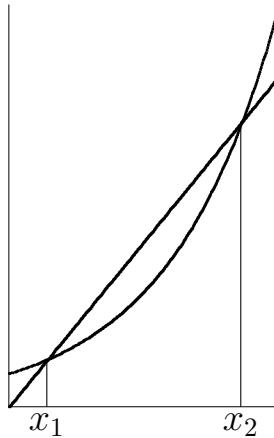


Рис.13.1

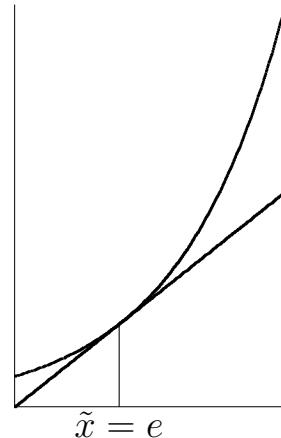


Рис.13.2

Если функция f имеет обратную, то неподвижные точки f являются и неподвижными точками f^{-1} . Из формулы для производной обратной функции (п.8.2) следует, что

$$(f^{-1})'(\tilde{x}) = \frac{1}{f'(f^{-1}(\tilde{x}))} = \frac{1}{f'(\tilde{x})}.$$

Видно, что точки отталкивания функции f будут точками притяжения для функции f^{-1} . Найдите в нашем примере x_2 , преобразовав уравнение $x = \exp(ax)$ к виду $x = \frac{\ln(x)}{a}$.

Замечание. При $a = 1/e$ решения уравнения сливаются (рис.13.2), причем $f'(\tilde{x}) = 1$. В этом случае теорема не работает. *Можно показать, что* в данном примере при $x_0 < \tilde{x} = e$ последовательность (13.1.2) сходится к \tilde{x} , а при $x_0 > \tilde{x} = e$ – нет. При $a > 1/e$ уравнение решений не имеет.

По аналогии с уравнением (13.1.1) рассмотрим систему уравнений специального вида

$$x = F(x), \quad x \in \mathbb{R}^n, \quad F : \mathbb{R}^n \rightarrow \mathbb{R}^n. \quad (13.1.5)$$

Можно показать, что справедлив следующий аналог теоремы 1:

Теорема 1'. Пусть $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ – непрерывно дифференцируемое векторное поле, и для любого $x \in \mathbb{R}^n$ $\|F'(x)\| \leq q < 1$ (здесь $\|\cdot\|$ – норма матрицы). Тогда

- 1) система (13.1.5) имеет единственное решение (обозначим его \tilde{x});
- 2) для любого $x^{(0)} \in \mathbb{R}^n$ последовательность

$$x^{(1)} = F(x^{(0)}), \dots, x^{(k)} = F(x^{(k-1)}), \dots \quad (13.1.6)$$

сходится к \tilde{x} ;

$$3) \|x^{(k)} - \tilde{x}\| \leq q^k \cdot \|x^{(0)} - \tilde{x}\|.$$

Пример. Пусть $F(x) = Ax + b$, A – квадратная матрица порядка n , $b \in \mathbb{R}^n$. Поскольку $F'(x) = A$, теорема в этом случае совпадает с известной из курса линейной алгебры теоремой о сходимости метода простой итерации.

Замечание. Условия теоремы гарантируют, что отображение F является сжатием. *Можно показать, что* любое сжимающее отображение \mathbb{R}^n в себя имеет единственную неподвижную точку. Это утверждение справедливо и для любого конечномерного нормированного пространства (т.е. пространства, в котором введена норма).

Приведем также аналог теоремы 2 для систем:

Теорема 2'. Пусть непрерывно дифференцируемое отображение $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ имеет неподвижную точку \tilde{x} , и $\|F'(\tilde{x})\| < 1$. Тогда \tilde{x} – точка притяжения итерационного процесса (13.1.6), т.е. существует такая

окрестность точки \tilde{x} , что для любого $x^{(0)}$ из нее последовательность (13.1.6) сходится к этой точке.

Замечания. 1. Теоремы **2** и **2'** обладают двумя общими недостатками: во-первых, необходимо знать заранее, что неподвижная точка есть; во-вторых, неизвестна окрестность, в которой следует брать начальное приближение.

2. В многомерном случае нет простого способа превратить точку отталкивания в точку притяжения (переходом к обратной функции). Хотя матрица $(F^{-1})'(\tilde{x})$ обратна матрице $F'(\tilde{x})$, но из того, что $\|A\| > 1$, вовсе не следует, что $\|A^{-1}\| < 1$.

13.2. Метод Ньютона для решения нелинейных уравнений и систем

Начнем с рассмотрения одного уравнения

$$f(x) = 0. \quad (13.2.1)$$

Пусть \tilde{x} – решение этого уравнения, $x_0 \in \mathbb{R}$ – произвольная точка, и $f'(x_0) \neq 0$.

Запишем левую часть уравнения (13.2.1) по формуле Тейлора:

$$f(x_0) + f'(x_0) \cdot (x - x_0) + \frac{f''(\xi)}{2} \cdot (x - x_0)^2 = 0$$

и отбросим остаточный член. Получим *линейное* уравнение

$$f(x_0) + f'(x_0) \cdot (x - x_0) = 0, \quad (13.2.2)$$

решение которого обозначим x_1 :

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

Если точка \tilde{x} лежит достаточно близко к x_0 , то остаточный член формулы Тейлора мал по сравнению с ее первыми слагаемыми, и можно ожидать, что решения уравнений (13.2.2) и (13.2.1), т.е. точки x_1 и \tilde{x} , будут мало отличаться друг от друга. Это наводит на мысль, что последовательность

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}, \dots, x_k = x_{k-1} - \frac{f(x_{k-1})}{f'(x_{k-1})}, \dots \quad (13.2.3)$$

должна сходиться к \tilde{x} .

Действительно, рассмотрим уравнение

$$x = \phi(x) \equiv x - \frac{f(x)}{f'(x)},$$

которое равносильно (13.2.1), если $f'(\tilde{x}) \neq 0$.

Так как

$$\phi'(x) = 1 - \frac{(f')^2(x) - f(x) \cdot f''(x)}{(f')^2(x)} = \frac{f(x) \cdot f''(x)}{(f')^2(x)},$$

то $\phi'(\tilde{x}) = 0$, и по теореме 2 из п.13.1 \tilde{x} – точка притяжения итерационного процесса (13.2.3), т.е. последовательность (13.2.3) действительно сходится к решению уравнения, *если начальное приближение взято достаточно близко к решению.*

Алгоритм, основанный на итерациях по формулам (13.2.3), называют алгоритмом Ньютона³⁹.

Замечания. 1. Мы доказали сходимость алгоритма Ньютона, предполагая наличие у функции второй производной. *Можно показать, что для сходимости достаточно существования лишь непрерывной первой производной и отличия ее от нуля в точке-решении.*

2. Естественно, знание того факта, что искомое решение \tilde{x} есть точка притяжения итерационного процесса (13.2.3), ничего нам не говорит об окрестности, в которой этот процесс сходится. Однако алгоритм Ньютона обладает приятным свойством: он сходится не всегда, но если уж сходится, то обычно очень быстро. Это дает возможность проводить машинный эксперимент: брать различные начальные приближения x_0 и смотреть, сходится ли последовательность итераций.

Пример. Решим уравнение $\arctg(x) = 0$. Итерационная формула принимает вид

$$x_k = x_{k-1} - \frac{\arctg(x_{k-1})}{\arctg'(x_{k-1})} = x_{k-1} - (1 + x_{k-1}^2) \cdot \arctg(x_{k-1}).$$

Попробуйте выполнить несколько итераций, начиная с $x_0 = 1$, а затем – с $x_0 = 1.5$.

³⁹Исаак НЬЮТОН (I. Newton, 1643-1727) – крупнейший английский математик, физик и астроном, президент Лондонского Королевского общества, член Парижской АН, один из основателей (наряду с Г.В. Лейбницем) математического анализа. Ньютону принадлежит современная формулировка законов классической механики и закона всемирного тяготения, а также важнейшие открытия в оптике.

Геометрическая интерпретация метода Ньютона очень проста: уравнение $y = f(x_0) + f'(x_0) \cdot (x - x_0)$ определяет касательную к графику функции f в точке x_0 . Поэтому замена уравнения (13.2.1) уравнением (13.2.2) означает, что мы вместо точки пересечения с осью абсцисс графика функции берем точку пересечения с осью абсцисс касательной к этому графику (рис.13.3). Поэтому метод Ньютона называют также *методом касательных*.

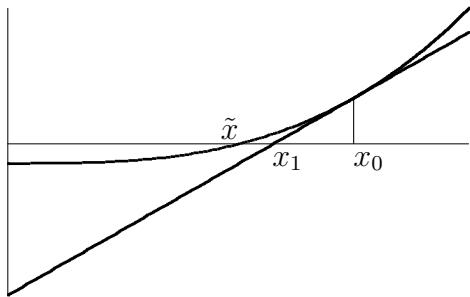


Рис.13.3

Перейдем теперь к рассмотрению системы уравнений

$$F(x) = \theta_n, \quad (13.2.4)$$

где $x \in \mathbb{R}^n$, $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ — векторное поле. Будем действовать аналогично случаю одного уравнения.

Пусть \tilde{x} — решение системы (13.2.4). Если матрица $F'(\tilde{x})$ не вырождена, то в некоторой окрестности точки \tilde{x} система (13.2.4) равносильна системе

$$x = \psi(x) \equiv x - (F'(x))^{-1} \cdot F(x).$$

Как и в случае одного уравнения, можно показать, что $\psi'(\tilde{x}) = \Theta_n$ (нуль-матрица размера $n \times n$). Поэтому $\|\psi'(\tilde{x})\| = 0$, и по теореме 2' из п.13.1 решение \tilde{x} является точкой притяжения итерационного процесса

$$x^{(k)} = x^{(k-1)} - (F'(x^{(k-1)}))^{-1} \cdot F(x^{(k-1)}). \quad (13.2.5)$$

Замечания. 1. Конечно, не следует на каждом шаге алгоритма обращать матрицу $F'(x^{(k-1)})$. Следует просто решить линейную систему $F'(x^{(k-1)}) \cdot y = F(x^{(k-1)})$ и положить $x^{(k)} = x^{(k-1)} - y$.

2. Матрица Якоби ψ' существует лишь при наличии *второй* производной (матрицы Гессе) у векторного поля F . Однако, как и

в случае одного уравнения, для сходимости итерационного процесса (13.2.5) достаточно существования непрерывной первой производной этого поля (при условии, конечно, что начальное приближение выбрано достаточно близко к решению). Остается также в силе замечание 2 к одномерной задаче.

13.3. Производная скалярного поля по направлению. Метод градиентного спуска

Определение. Пусть $f : U \subset \mathbb{R}^n \rightarrow \mathbb{R}$ – непрерывно дифференцируемый функционал (скалярное поле), $e \in \mathbb{R}^n$, $\|e\| = 1$. *Производной скалярного поля f в точке $a \in U$ по направлению вектора e* называется число

$$D_e f(a) = \lim_{t=0+} \frac{f(a + te) - f(a)}{t}.$$

Запишем формулу Тейлора первого порядка для функции $\phi(t) = f(a + te)$:

$$\phi(t) - \phi(0) = \phi'(\xi)t = f'(a + \xi e) \cdot et \quad (\xi \in]0, t[).$$

Отсюда

$$D_e f(a) = \lim_{t=0+} \frac{f'(a + \xi e) \cdot et}{t} = f'(a) \cdot e = \langle \nabla f(a), e \rangle.$$

Таким образом, производная скалярного поля по направлению некоторого вектора равна проекции градиента этого поля на это направление.

Замечание. Если $e = e^{(k)}$ – вектор из стандартного базиса, то $D_e f(a) = D_k f(a)$ – частная производная функционала f в точке a по переменной x_k .

По неравенству Коши–Буняковского–Шварца (см. п.7.1 раздела "Линейная алгебра")

$$|\langle \nabla f(a), e \rangle| \leq \|\nabla f(a)\| \cdot \|e\| = \|\nabla f(a)\|.$$

Равенство достигается, когда \vec{e} коллинеарен $\overrightarrow{\nabla f(a)}$. При этом, если \vec{e} сонаправлен $\overrightarrow{\nabla f(a)}$, то $D_e f(a) = \|\nabla f(a)\|$, если же \vec{e} противонаправлен $\overrightarrow{\nabla f(a)}$, то $D_e f(a) = -\|\nabla f(a)\|$.

Таким образом, скалярное поле "быстрее всего возрастает" в направлении своего градиента и "быстрее всего убывает" в направлении своего антиградиента, т.е. вектора $-\nabla f(a)$. Производная скалярного поля по направлению, ортогональному градиенту, равна нулю.

Опишем теперь так называемый *метод градиентного спуска* – вычислительный алгоритм, позволяющий находить локальные экстремумы функционала. Для определенности будем считать, что мы ищем точки локального минимума (для поиска максимума следует просто заменить f на $(-f)$). Также предположим, как обычно, что функционал непрерывно дифференцируем на \mathbb{R}^n .

Алгоритм. 1. Взять произвольную точку $x^{(0)} \in \mathbb{R}^n$.

2. Вычислить $\nabla f(x^{(0)})$.

3. Если $\nabla f(x^{(0)}) = \theta_n$, то $x^{(0)}$ – стационарная точка. Если она является точкой минимума, то *работа алгоритма закончена*. Иначе следует взять новую начальную точку и перейти к п.2.

Если $\nabla f(x^{(0)}) \neq \theta_n$, то построить сужение f на луч с началом в точке $x^{(0)}$ и направляющим вектором $(-\nabla f(x^{(0)}))$:

$$\psi(t) = f\left(x^{(0)} - t \cdot \nabla f(x^{(0)})\right), \quad t \geq 0.$$

4. Выйдя из точки $x^{(0)}$, двигаться по лучу $x = x^{(0)} - t \cdot \nabla f(x^{(0)})$ до тех пор, пока функционал убывает, т.е. следует найти t_0 – *наименьший положительный* корень уравнения

$$\psi'(t) = 0 \quad \text{или} \quad -f'\left(x^{(0)} - t \cdot \nabla f(x^{(0)})\right) \cdot \nabla f(x^{(0)}) = 0.$$

Если это уравнение не имеет положительных корней, то *работа алгоритма закончена* – минимум не найден.

5. Заменить $x^{(0)}$ на $x^{(0)} - t_0 \cdot \nabla f(x^{(0)})$ и перейти к п.2.

Идея алгоритма чрезвычайно проста: на каждом его шаге "выходят" из точки $x^{(0)}$ в направлении антиградиента (направлении наискорейшего убывания функционала в точке $x^{(0)}$) и движутся до тех пор, пока функционал в этом направлении убывает.

Конечно, алгоритм не гарантирует нахождение локального минимума, так как, во-первых, этот минимум может просто отсутствовать, а во-вторых, мы можем его "не там искать". Однако обсуждение технических проблем выходит за рамки нашего курса.

Глава 14. ИНТЕГРАЛ РИМАНА

14.1. Суммы Дарбу. Определение интеграла

В этом пункте мы будем рассматривать вещественные, кусочно непрерывные функции, заданные на некотором сегменте. Напомним, что функция называется *кусочно непрерывной* на сегменте, если она

- или непрерывна на этом сегменте,
- или имеет на этом сегменте *конечное* число точек разрыва и во всех этих точках существуют *конечные* односторонние пределы.

Мы будем говорить о *разбиении* сегмента $[a, b]$, если на этом сегменте задана сетка, содержащая концы сегмента – точки a и b . Записывать разбиение P мы будем так:

$$P = \{x_0, x_1, \dots, x_n\}.$$

Здесь $a = x_0 < x_1 < \dots < x_n = b$.

Итак, пусть на сегменте $[a, b]$ задана кусочно непрерывная вещественная функция f . Возьмем какое-нибудь разбиение P этого сегмента. Обозначим

$$J_k = [x_{k-1}, x_k]; \quad m_k = \inf_{x \in J_k} \{f(x)\}; \quad M_k = \sup_{x \in J_k} \{f(x)\}; \quad k = 1, \dots, n.$$

Назовем *нижней суммой Дарбу*⁴⁰ функции f при разбиении P число

$$L(f, P) = m_1(x_1 - x_0) + \dots + m_n(x_n - x_{n-1}) = \sum_{k=1}^n m_k(x_k - x_{k-1}),$$

а *верхней суммой Дарбу* функции f при разбиении P число

$$U(f, P) = M_1(x_1 - x_0) + \dots + M_n(x_n - x_{n-1}) = \sum_{k=1}^n M_k(x_k - x_{k-1}).$$

Пример. (Рис.14.1). $f : [0, 3] \rightarrow \mathbb{R}$:

$$f(x) = \begin{cases} x + 1 & \text{при } 0 \leq x < 1; \\ x & \text{при } 1 \leq x \leq 2; \\ 3(x - 2) & \text{при } 2 < x \leq 3. \end{cases}$$

⁴⁰Жан Гастон ДАРБУ (J.G. Darboux, 1842-1917) – французский математик, член Парижской АН, член-корр. Петербургской АН.

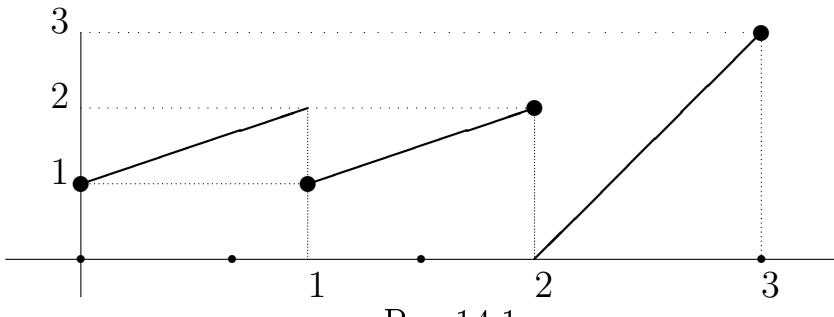


Рис.14.1

Возьмем разбиение $P = \{0, \frac{2}{3}, \frac{3}{2}, 3\}$.

На сегменте $J_1 = [0, \frac{2}{3}]$ имеем $m_1 = f(0) = 1$, $M_1 = f(\frac{2}{3}) = \frac{5}{3}$.

На сегменте $J_2 = [\frac{2}{3}, \frac{3}{2}]$ наибольшего значения нет, но $\sup_{x \in J_2} \{f(x)\} = \lim_{x=1-} f(x) = 2$; поэтому $M_2 = 2$. Далее, $m_2 = f(1) = 1$.

На сегменте $J_3 = [\frac{3}{2}, 3]$ наименьшего значения нет, но $\inf_{x \in J_3} \{f(x)\} = \lim_{x=2+} f(x) = 0$; поэтому $m_3 = 0$. Далее, $M_3 = f(3) = 3$.

Вычисляем суммы Дарбу:

$$L(f, P) = 1 \cdot \left(\frac{2}{3} - 0\right) + 1 \cdot \left(\frac{3}{2} - \frac{2}{3}\right) + 0 \cdot \left(3 - \frac{3}{2}\right) = \frac{3}{2},$$

$$U(f, P) = \frac{5}{3} \cdot \left(\frac{2}{3} - 0\right) + 2 \cdot \left(\frac{3}{2} - \frac{2}{3}\right) + 3 \cdot \left(3 - \frac{3}{2}\right) = \frac{131}{18}.$$

Установим некоторые свойства сумм Дарбу.

1. $L(f, P) \leq U(f, P)$.

Доказательство. Очевидно из построения.

2. При добавлении к сетке новой точки верхняя сумма Дарбу не увеличивается, а нижняя – не уменьшается.

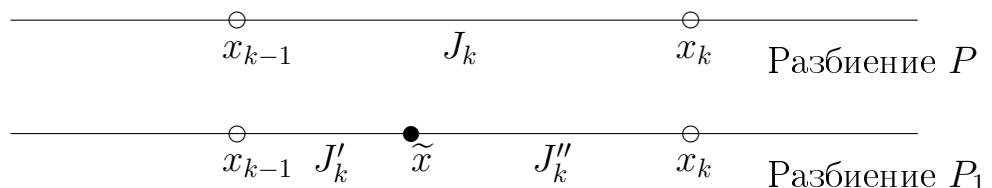


Рис.14.2

Доказательство. На рис.14.2 изображены разбиение P и полученное из него добавлением точки \tilde{x} разбиение P_1 . Очевидно, что верхние суммы Дарбу для этих разбиений отличаются лишь тем, что $U(f, P_1)$ вместо

одного слагаемого $M_k(x_k - x_{k-1})$ содержит два: $M'_k(\tilde{x} - x_{k-1}) + M''_k(x_k - \tilde{x})$ (здесь $M'_k = \sup_{x \in J'_k} \{f(x)\}$, $M''_k = \sup_{x \in J''_k} \{f(x)\}$).

Поэтому

$$\begin{aligned} U(f, P) - U(f, P_1) &= M_k(x_k - x_{k-1}) - (M''_k(x_k - \tilde{x}) + M'_k(\tilde{x} - x_{k-1})) = \\ &= M_k((x_k - \tilde{x}) + (\tilde{x} - x_{k-1})) - M''_k(x_k - \tilde{x}) - M'_k(\tilde{x} - x_{k-1}) = \\ &= (M_k - M''_k) \cdot (x_k - \tilde{x}) + (M_k - M'_k) \cdot (\tilde{x} - x_{k-1}). \end{aligned}$$

Сегменты J'_k и J''_k суть части сегмента J_k . Поэтому $M'_k \leq M_k$ и $M''_k \leq M_k$. Отсюда $U(f, P) - U(f, P_1) \geq 0$, т.е. $U(f, P_1) \leq U(f, P)$.

Аналогично доказывается, что $L(f, P_1) \geq L(f, P)$. ■

Замечание. Очевидно, что это свойство выполняется при добавлению к разбиению *любого конечного количества точек*.

3. Для *любых* разбиений P_1 и P_2

$$L(f, P_1) \leq U(f, P_2).$$

Доказательство. Если $P_1 = P_2$, то неравенство очевидно (по построению). Если $P_1 \neq P_2$, то построим разбиение P , содержащее все точки P_1 и все точки P_2 . Тогда на основании свойства 2

$$L(f, P_1) \leq L(f, P), \quad U(f, P) \leq U(f, P_2).$$

Учитывая, что $L(f, P) \leq U(f, P)$, получим $L(f, P_1) \leq U(f, P_2)$. ■

Подведем итоги. Наименьшее количество точек в сетке – две. Соответствующее "разбиение" сегмента $[a, b]$ содержит единственный сегмент $J_1 = [a, b]$. Этому разбиению соответствуют *наименьшая нижняя* сумма Дарбу $m(b - a)$ и *наибольшая верхняя* сумма Дарбу $M(b - a)$ (здесь $m = \inf_{x \in [a, b]} \{f(x)\}$, $M = \sup_{x \in [a, b]} \{f(x)\}$).

Если исключить тривиальный случай функции-константы, когда $m = M$, то множество всех (как верхних, так и нижних) сумм Дарбу содержится в сегменте $[m(b - a), M(b - a)]$ и потому ограничено. Пусть $L(f) = \sup\{L(f, P)\}$, $U(f) = \inf\{U(f, P)\}$. Возможны два случая:

1. На сегменте $[m(b - a), M(b - a)]$ есть "пустой" промежуток, разделяющий верхние и нижние суммы Дарбу (рис.14.3а).
2. На сегменте $[m(b - a), M(b - a)]$ есть единственная точка, разделяющая верхние и нижние суммы Дарбу (рис.14.3б).

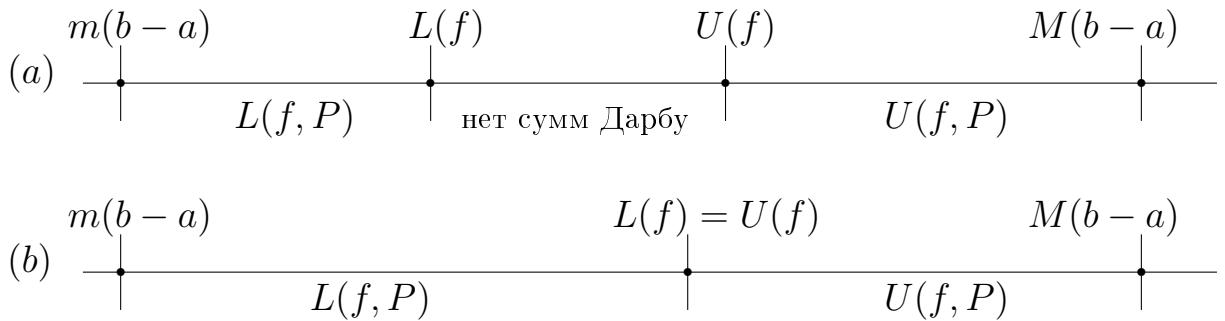


Рис.14.3

Можно показать, что справедлива следующая

Теорема. Если f – вещественная, кусочно непрерывная на $[a, b]$ функция, то существует единственное число $L(f) = U(f)$, удовлетворяющее неравенству

$$L(f, P) \leq L(f) = U(f) \leq U(f, P)$$

при любом разбиении P .

Это число называют *интегралом Римана* функции f по сегменту $[a, b]$ и обозначают

$$\int_a^b f.$$

Не имея возможности в рамках нашего курса доказать эту теорему, покажем, что требование кусочной непрерывности функции существенно. Рассмотрим один "патологический" пример – так называемую функцию Дирихле:

$$f : [0, 1] \rightarrow \mathbb{R}, \quad f(x) = \begin{cases} 1, & \text{если } x \text{ – рациональное число;} \\ 0, & \text{если } x \text{ – иррациональное число.} \end{cases}$$

График этой функции можно представить себе так: следует "вынуть" из отрезка $[0, 1]$ числовой оси все точки с рациональными координатами и поднять эти точки вверх на одну единицу длины. Получается два "дырявых" отрезка (рис.14.4).



Рис.14.4

Возьмем произвольное разбиение сегмента $[0, 1]$. На любом элементарном сегменте J_k этого разбиения найдутся и рациональные, и иррациональные точки, т.е. множество значений функции Дирихле на J_k будет состоять из двух чисел – нуля и единицы. Поэтому $m_k \equiv 0$, $M_k \equiv 1$.

Итак, все нижние суммы Дарбу для функции Дирихле равны нулю, а все верхние – единице. Для этой функции интеграл Римана не существует!

Замечания. 1. Символ

$$\int_a^b f$$

содержит все необходимые сведения об интеграле Римана. Функцию f обычно называют *подынтегральной функцией*, а числа a и b – концы сегмента – *пределами интегрирования* (a – *нижним*, b – *верхним*). Отметим (на всякий случай), что в этом контексте слово "предел" ничего общего с понятием "предел функции" не имеет!

По традиции символ интеграла Римана (или, как его еще до сих пор называют, "определенного интеграла") записывают так:

$$\int_a^b f(x) dx,$$

а букву x называют "переменная интегрирования"⁴¹. Очевидно, не хуже выглядят записи

$$\int_a^b f(y) dy, \quad \int_a^b f(\bar{y}) d\bar{y}, \quad \int_a^b f(\cdot) d(\cdot), \quad \dots$$

Отметим, что в случае функции нескольких переменных такое обозначение оказывается целесообразным. Например, в выражении

$$\int_a^b f(x, y) dy$$

⁴¹ Встречались даже "теоремы" о независимости интеграла от обозначения переменной интегрирования.

видно, что переменная x фиксирована, и f рассматривается как функция одной переменной — y .

2. Первоначально определение интеграла Римана было другим — через *предел интегральных сумм*.

Назовем *рангом разбиения* P число $\lambda(P) = \max_k(x_k - x_{k-1})$ — наибольшую из длин элементарных сегментов этого разбиения, а интегральной суммой — число

$$S(f, P) = \sum_{k=1}^n f(\xi_k) \cdot (x_k - x_{k-1}),$$

где ξ_k — какая-нибудь точка на сегменте $[x_{k-1}, x_k]$.

Обратите внимание на то, что при заданном разбиении можно построить сколько угодно интегральных сумм, варьируя точки ξ_k , в которых вычисляются значения функции.

Можно показать, что имеет место

Теорема. Пусть f кусочно непрерывна на сегменте. Тогда для всякого положительного числа ε можно указать такое положительное число δ , что при любом разбиении, ранг которого меньше, чем δ , все интегральные суммы отличаются от интеграла Римана меньше, чем на ε , т.е.

$$\lambda(P) < \delta \implies |S(f, P) - \int_a^b f| < \varepsilon.$$

Этот любопытный факт мы в дальнейшем использовать не будем.

3. В математике существуют другие конструкции, в названии которых присутствует слово "интеграл" (интеграл Лебега, интеграл Радона...). Поскольку в нашем курсе рассматривается только интеграл Римана, мы будем иногда позволять себе говорить просто "интеграл подразумевая интеграл Римана. Кроме того, если не оговорено противное, мы будем считать по умолчанию, что подынтегральная функция — вещественная.

4. Если $f = \text{const}$, то при любом разбиении P сегмента $[a, b]$ имеем $L(f, P) = U(f, P) = \text{const} \cdot (b - a)$ и, следовательно,

$$\int_a^b \text{const} = \text{const} \cdot (b - a).$$

14.2. Физические и геометрические интерпретации интеграла Римана

В этом пункте мы рассмотрим три содержательные задачи, приводящие к интегралу Римана.

Задача 1. Электрический заряд распределен вдоль стержня длины L с линейной плотностью ρ (кулонов на метр). Найти полный заряд стержня.

Совместим начало координат с началом стержня. Тогда конец стержня будет иметь координату $x = L$, и по условию задачи на сегменте $[0, L]$ будет определена вещественная функция $x \rightarrow \rho(x)$. Будем считать ее кусочно непрерывной (нам не удалось придумать реальную физическую задачу, где это условие не выполнено).

Возьмем произвольное разбиение сегмента $[0, L]$

$$P = \{0 = x_0, \dots, x_{k-1}, x_k, \dots, x_n = L\}.$$

Обозначим

$$J_k = [x_{k-1}, x_k]; \quad m_k = \inf_{x \in J_k} \{\rho(x)\}; \quad M_k = \sup_{x \in J_k} \{\rho(x)\}; \quad k = 1, \dots, n.$$

Верхняя сумма Дарбу $U(\rho, P) = \sum_{k=1}^n M_k(x_k - x_{k-1})$ может рассматриваться как полный заряд стержня при кусочно постоянной плотности распределения (на каждом элементарном сегменте J_k плотность постоянна и равна M_k). Аналогично, нижнюю сумму Дарбу можно рассматривать как полный заряд стержня с кусочно постоянной плотностью распределения, но теперь на каждом элементарном сегменте плотность распределения заряда равна m_k .

Очевидно (для физика) неравенство, которому должно удовлетворять число Q (заряд стержня) при любом разбиении P :

$$L(\rho, P) \leq Q \leq U(\rho, P).$$

Однако известно, что существует *единственное* число, обладающее таким свойством, и это число – интеграл от функции ρ по сегменту $[0, L]$. Таким образом,

$$Q = \int_0^L \rho.$$

Задача 2. Плоская фигура ограничена осью абсцисс, прямыми $x = a$ и $x = b$ и графиком неотрицательной непрерывной функции f (рис.14.5). Найти площадь этой "криволинейной трапеции".

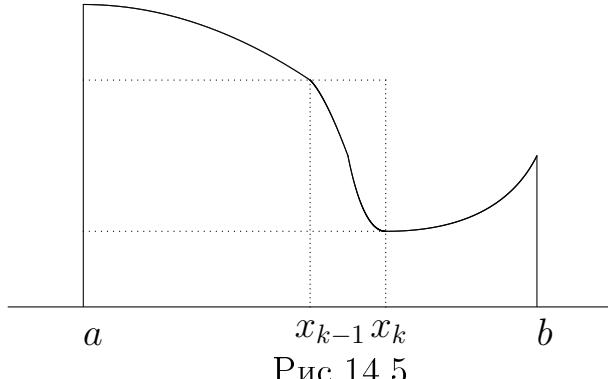


Рис.14.5

Возьмем произвольное разбиение сегмента $[a, b]$

$$P = \{a = x_0, \dots, x_{k-1}, x_k, \dots, x_n = b\}.$$

Верхняя сумма Дарбу функции f для этого разбиения может быть истолкована как площадь "ступенчатой фигуры составленной из прямоугольников, основаниями которых служат элементарные сегменты разбиения, а высотами – наибольшие ординаты графика на этих сегментах. Аналогично, нижняя сумма Дарбу – площадь ступенчатой фигуры, составленной из прямоугольников с теми же основаниями, но высоты теперь – наименьшие ординаты графика.

Очевидно (для геометра) неравенство, которому должно удовлетворять число S (площадь фигуры) при любом разбиении P :

$$L(f, P) \leq S \leq U(f, P).$$

Однако известно, что существует единственное число, обладающее таким свойством, и это число – интеграл от функции f по сегменту $[a, b]$. Таким образом,

$$S = \int_a^b f.$$

Задача 3. Задан кусочно гладкий путь $r : [a, b] \rightarrow \mathbb{R}^3$. Найти длину этого пути.

Возьмем произвольное разбиение сегмента $[a, b]$

$$P = \{a = t_0, \dots, t_{k-1}, t_k, \dots, t_n = b\}.$$

Если бы на $J_k = [t_{k-1}, t_k]$ скорость r' была постоянной, то, очевидно, длина пути, пройденного с момента t_{k-1} до момента t_k , была бы равна $\|r'\| \cdot (t_k - t_{k-1})$. Далее, очевидно (для физика), что истинная длина пути, пройденного с момента t_{k-1} до момента t_k , должна удовлетворять неравенству

$$\inf_{t \in J_k} \{\|r'(t)\|\} \cdot (t_k - t_{k-1}) \leq S_k \leq \sup_{t \in J_k} \{\|r'(t)\|\} \cdot (t_k - t_{k-1}).$$

Складывая такие неравенства по всем $k = 1, \dots, n$, получим неравенство

$$L(\|r'\|, P) \leq S \leq U(\|r'\|, P),$$

которое должно выполняться для любого разбиения P .

Поскольку для *кусочно непрерывной* (по определению кусочно гладкого пути) функции $\|r'\|$ этому неравенству удовлетворяет только интеграл Римана, мы приходим к формуле

$$S = \int_a^b \|r'\| = \int_a^b \sqrt{(r'_1)^2 + (r'_2)^2 + (r'_3)^2}. \quad (14.2.1)$$

Замечания. 1. В случае, когда путь плоский, выражение под корнем в (14.2.1), естественно, содержит два слагаемых.

2. Если отображение r задает гладкую кривую, длина пути называется также *длиной кривой*.

Пример. Как известно, график непрерывно дифференцируемой функции $f : [a, b] \rightarrow \mathbb{R}$ – гладкая кривая (см. пример в п.11.2). Для этого случая формула (14.2.1) перепишется так:

$$S = \int_a^b \sqrt{1 + (f')^2}.$$

Рассмотренные примеры демонстрируют, как различные по содержанию *прикладные задачи* сводятся к одной и той же *математической модели* – интегралу Римана.

14.3. Простейшие свойства интеграла Римана

Из определения интеграла следует, что нижний предел интегрирования (левый конец сегмента) должен быть *меньше* верхнего предела

(правого конца сегмента). Чтобы освободиться от этого стеснительного условия, при любой функции f полагают *по определению*

$$\int_a^a f = 0 \quad (14.3.1)$$

(заряд стержня нулевой длины естественно считать равным нулю).

Если $a > b$, то также *по определению* полагают

$$\int_a^b f = - \int_b^a f. \quad (14.3.2)$$

Доказательства следующих двух свойств не приводятся из-за их технической сложности, но мы рекомендуем читателю интерпретировать эти свойства физически (на примере заряженного стержня) и геометрически (на примере площади криволинейной трапеции).

1. Если функция f кусочно непрерывна на сегменте $[a, c]$, а точка b лежит на этом сегменте, то

$$\int_a^c f = \int_a^b f + \int_b^c f. \quad (14.3.3)$$

("Если заряженный стержень разрезать на части, то заряд всего стержня равен сумме зарядов частей". Не примите эту фразу за доказательство!).

Учитывая определения (14.3.1) и (14.3.2), можно распространить формулу (14.3.3) на случай *произвольного* расположения точек на числовой оси (лишь бы f была кусочно непрерывна на всех участвующих в равенстве сегментах).

2. Если функции f_1 и f_2 кусочно непрерывны на $[a, b]$, то при любых числах α_1 и α_2

$$\int_a^b (\alpha_1 f_1 + \alpha_2 f_2) = \alpha_1 \cdot \int_a^b f_1 + \alpha_2 \cdot \int_a^b f_2. \quad (14.3.4)$$

Это равенство позволяет назвать интеграл Римана *линейным* функционалом, заданном на множестве функций, кусочно непрерывных на $[a, b]$.

3. Интеграл не изменится, если произвольно изменить значения подынтегральной функции в *конечном* числе точек сегмента.

Доказательство. Рассмотрим функцию ϕ_c , равную единице в точке c сегмента $[a, b]$, а в остальных точках этого сегмента равную нулю. При любом разбиении сегмента наименьшее значение ϕ_c на каждом элементарном сегменте будет равно нулю, т.е. будут равны нулю все нижние суммы Дарбу. Поэтому будет равна нулю и верхняя грань множества нижних сумм Дарбу (она же – интеграл Римана от функции ϕ_c). Итак,

$$\int_a^b \phi_c = 0.$$

Пусть теперь f – произвольная функция, кусочно непрерывная на $[a, b]$. Изменим ее значение в точке $c \in [a, b]$ на величину A . Получим новую функцию $g = f + A \cdot \phi_c$. Далее, вследствие линейности интеграла,

$$\int_a^b g = \int_a^b f + A \int_a^b \phi_c = \int_a^b f.$$

Распространение доказательства на случай изменения значений функции в нескольких точках очевидно. ■

Замечание. Это свойство дает основание интегрировать кусочно непрерывные функции, *не определенные в конечном числе точек сегмента* (результат, очевидно, не зависит от способа доопределения).

14.4. Среднее значение функции на сегменте

Определение. *Средним значением* кусочно непрерывной функции f на сегменте $[a, b]$ называется число

$$f_{cp} = \frac{1}{b-a} \int_a^b f.$$

Пример. Пусть $f(x) = \begin{cases} 1 & \text{при } 0 \leq x < 1; \\ 2 & \text{при } 1 \leq x \leq 2. \end{cases}$ Тогда

$$f_{cp} = \frac{1}{2-0} \cdot \int_0^2 f = \frac{1}{2} \cdot \left(\int_0^1 1 + \int_1^2 2 \right) = \frac{1}{2}(1+2) = \frac{3}{2}.$$

Отметим, что в этом примере среднее значение функции не содержится во множестве ее значений!

Ситуация меняется, если функция непрерывна.

Теорема. Если функция f непрерывна на сегменте $[a, b]$, то ее среднее значение является ее значением. Точнее, найдется такая точка $\xi \in [a, b]$, что $f(\xi) = f_{cp}$.

Доказательство. По теореме Вейерштрасса среди значений непрерывной на $[a, b]$ функции f есть наибольшее (M) и наименьшее (m). По определению интеграла

$$m(b-a) \leq \int_a^b f \leq M(b-a), \quad \text{или} \quad m \leq \frac{1}{b-a} \int_a^b f \leq M,$$

т.е. среднее значение непрерывной функции лежит на сегменте $[m, M]$. По теореме Коши все числа этого сегмента – значения функции f . Следовательно, среднее значение непрерывной функции является ее значением. ■

Эту теорему обычно называют *теоремой о среднем* и пишут

$$\int_a^b f = f(\xi)(b-a) \quad (\xi – некоторая точка на $[a, b]$).$$

Замечания. 1. Мы выделили слово "некоторая чтобы подчеркнуть, что теорема о среднем лишь утверждает существование такой точки, но не дает алгоритма ее отыскания.

2. В доказательстве предполагалось, что $a < b$. Однако легко видеть, что утверждение теоремы сохраняется и при $a > b$.

14.5. Интегрирование неравенств

Теорема. Если f кусочно непрерывна на $[a, b]$, и $f \geq 0$, то $\int_a^b f \geq 0$.

Доказательство. Если $f \geq 0$, то $m = \inf_{x \in [a, b]} \{f(x)\} \geq 0$, и по определению интеграла

$$\int_a^b f \geq m(b-a) \geq 0. \quad ■$$

Следствия. 1. Если f_1 и f_2 – функции, кусочно непрерывные на $[a, b]$, $f_1 \geq f_2$ и $a < b$, то $\int_a^b f_1 \geq \int_a^b f_2$.

Доказательство.

$$f_1 \geq f_2 \implies f_1 - f_2 \geq 0 \implies \int_a^b (f_1 - f_2) \geq 0 \implies \int_a^b f_1 \geq \int_a^b f_2. \blacksquare$$

2. Если вещественная функция f кусочно непрерывна на $[a, b]$, то

$$\left| \int_a^b f \right| \leq \int_a^b |f|.$$

Доказательство. Заметим, что функция $|f|$ также кусочно непрерывна на $[a, b]$. Интегрируя очевидное неравенство $-|f| \leq f \leq |f|$, получим

$$-\int_a^b |f| \leq \int_a^b f \leq \int_a^b |f|, \quad \text{или} \quad \left| \int_a^b f \right| \leq \int_a^b |f|. \blacksquare$$

Замечание. В этом пункте предполагается, что $a < b$. Если нижний предел интегрирования больше верхнего, то при интегрировании неравенства его знак меняется!

14.6. Интеграл от комплекснозначной функции

Всякую функцию $f : [a, b] \rightarrow \mathbb{C}$ можно записать в виде $f = f_1 + i \cdot f_2$, где f_1 и f_2 – вещественные функции:

$$f_1(x) = \operatorname{Re}(f(x)), \quad f_2(x) = \operatorname{Im}(f(x)).$$

Эти функции естественно обозначать $\operatorname{Re}(f)$ и $\operatorname{Im}(f)$ соответственно.

Если $\operatorname{Re}(f)$ и $\operatorname{Im}(f)$ кусочно непрерывны на $[a, b]$, положим по определению

$$\int_a^b f = \int_a^b \operatorname{Re}(f) + i \int_a^b \operatorname{Im}(f).$$

Докажем, что для комплекснозначной функции f справедливо неравенство (сравните со следствием 2 в п.14.5)

$$\left| \int_a^b f \right| \leq \int_a^b |f|. \tag{14.6.1}$$

Обозначим $A = \int_a^b Re(f)$, $B = \int_a^b Im(f)$. Тогда $\int_a^b f = A + i \cdot B$.

$$\left| \int_a^b f \right|^2 = A^2 + B^2 = A \cdot \int_a^b Re(f) + B \cdot \int_a^b Im(f) = \int_a^b (A \cdot Re(f) + B \cdot Im(f)).$$

По неравенству Коши–Буняковского–Шварца

$$\begin{aligned} A \cdot Re(f(x)) + B \cdot Im(f(x)) &\leq \sqrt{A^2 + B^2} \cdot \sqrt{Re^2(f(x)) + Im^2(f(x))} \leq \\ &\leq \sqrt{A^2 + B^2} \cdot |f(x)|. \end{aligned}$$

Поэтому

$$\left| \int_a^b f \right|^2 \leq \int_a^b (\sqrt{A^2 + B^2} \cdot |f|).$$

Сокращая на $\left| \int_a^b f \right| = \sqrt{A^2 + B^2}$, получим $\left| \int_a^b f \right| \leq \int_a^b |f|$. ■

14.7. Интегралы с переменными пределами. Первообразная функция

Если f кусочно непрерывна на $[a, b]$, то она кусочно непрерывна на любом сегменте, содержащемся в $[a, b]$. Зафиксируем точку $c \in [a, b]$ и определим при всех $x \in [a, b]$ новую функцию

$$F(x) = \int_c^x f, \tag{14.7.1}$$

называемую *интеграл с переменным верхним пределом*.

Установим некоторые свойства интеграла с переменным верхним пределом (14.7.1).

1. F непрерывна на $[a, b]$.

Доказательство. Пусть $x, y \in [a, b]$ и $x < y$. Тогда

$$F(y) - F(x) = \int_c^y f - \int_c^x f = \int_x^y f.$$

Если $M = \sup_{x \in [a,b]} \{|f(x)|\}$, то $|f| \leq M$. Поэтому

$$|F(y) - F(x)| = \left| \int_x^y f \right| \leq \int_x^y |f| \leq \int_x^y M = M \cdot (y - x)$$

(здесь использовано то, что $x < y$).

Пусть теперь ε – произвольное положительное число. Взяв $\delta = \frac{\varepsilon}{M}$, получим

$$|y - x| = y - x < \delta \implies |F(y) - F(x)| \leq M \cdot \delta = \varepsilon.$$

Доказана непрерывность *справа* функции F в любой точке $x \in [a, b[$ и непрерывность ее *слева* в любой точке $x \in]a, b]$, т.е. непрерывность F на $[a, b]$. ■

2. Если f непрерывна в точке $x \in]a, b[$, то $F'(x) = f(x)$ (это утверждение называют теоремой Барроу⁴²).

Доказательство. Кусочно непрерывная функция f по определению может иметь на $[a, b]$ лишь конечное множество точек разрыва первого рода. Пусть точка y выбрана так, что на $[x, y]$ (или на $[y, x]$) f непрерывна. Тогда по теореме о среднем между x и y найдется такая точка ξ , что $\int_x^y f = f(\xi) \cdot (y - x)$. Следовательно,

$$F(y) - F(x) = \int_x^y f = f(\xi) \cdot (y - x), \quad \text{или} \quad \frac{F(y) - F(x)}{y - x} = f(\xi).$$

Переходим к пределу, учитывая, что ξ лежит между x и y , а f непрерывна в точке x :

$$F'(x) = \lim_{y \rightarrow x} \frac{F(y) - F(x)}{y - x} = \lim_{y \rightarrow x} f(\xi) = f(x). \quad ■$$

Итак, если f кусочно непрерывна на $[a, b]$, то функция F , определенная равенством (14.7.1), *непрерывна во всех точках* $[a, b]$, а *в точках, где f непрерывна*, F имеет производную, причем $F'(x) = f(x)$.

Определение. Если f *кусочно непрерывна* на некотором промежутке, то всякая *непрерывная* на этом промежутке функция, производная

⁴²Исаак БАРРОУ (I. Barrow, 1630-1677) – английский математик и богослов. Учитель И. Ньютона.

которой в точках непрерывности f совпадает с f , называется *первообразной функцией* для функции f .

Примеры. 1. Как было показано выше, интеграл с переменным верхним пределом есть первообразная (обычно говорят не "первообразная функция а "первообразная") для подынтегральной функции.

2. Функция

$$abs(x) = \begin{cases} -x & \text{при } x < 0 \\ x & \text{при } x \geq 0 \end{cases}$$

– первообразная для функции $sign$, ибо она всюду непрерывна, и $abs'(x) = sign(x)$ при $x \neq 0$.

Заметим, что вообще первообразная для полиномиального сплайна сама является полиномиальным сплайном.

Легко видеть, что первообразная у кусочно непрерывной функции не единственна. Так, прибавляя к первообразной F функцию-константу, мы получаем новую первообразную, так как $F + const$ непрерывна и $(F + const)' = F'$.

Однако произвол в выборе первообразной этим и ограничивается.

Теорема. Разность двух первообразных кусочно непрерывной функции есть функция-константа.

Доказательство. Пусть F_1 и F_2 – первообразные для f на некотором промежутке. Рассмотрим их разность $F = F_1 - F_2$ и покажем сначала, что она постоянна на интервалах непрерывности f .

Действительно, для любых двух точек $x_1 < x_2$ из интервала непрерывности f формула конечных приращений (п.10.2) дает

$$F(x_2) - F(x_1) = F'(\xi) \cdot (x_2 - x_1) = (F'_1(\xi) - F'_2(\xi)) \cdot (x_2 - x_1) = f(\xi) - f(\xi) = 0$$

(здесь ξ – некоторая точка из $]x_1, x_2[$).

Итак, F *кусочно постоянна* на промежутке. Но поскольку она непрерывна (как разность непрерывных функций), она *постоянна!* ■

Пример. Пусть f кусочно непрерывна на $[a, b]$. Зафиксируем точку $d \in [a, b]$ и определим на $[a, b]$ *интеграл с переменным нижним пределом*

$$\Phi(x) = \int_x^d f.$$

Легко видеть, что функция $-\Phi$ является интегралом с переменным верхним пределом от f . Поэтому она удовлетворяет свойствам **1** и **2** и, следовательно, является первообразной для f . Согласно доказанной теореме,

$$-\Phi = F + \text{const}, \quad \text{или} \quad F(x) + \Phi(x) = \text{const}.$$

Но это очевидно, так как $\int\limits_c^x f + \int\limits_x^d f = \int\limits_c^d f$.

Интегралы с переменными пределами часто используются как способ задания функций.

Примеры. 1. "Функция ошибок" $\text{erf} : \mathbb{R} \rightarrow \mathbb{R}$

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \cdot \int\limits_0^x \exp(-t^2) dt.$$

2. "Интегральный синус" $\text{Si} : \mathbb{R} \rightarrow \mathbb{R}$

$$\text{Si}(x) = \int\limits_0^x \frac{\sin(t)}{t} dt.$$

3. Интегралы Френеля⁴³ $S, C : \mathbb{R} \rightarrow \mathbb{R}$

$$S(x) = \int\limits_0^x \sin\left(\frac{\pi}{2}t^2\right) dt, \quad C(x) = \int\limits_0^x \cos\left(\frac{\pi}{2}t^2\right) dt.$$

Замечание. Приведенные в этих примерах функции не выражаются через "школьные" (которые принято называть "элементарными") с помощью конечного числа арифметических операций и композиций. По этой причине их уважительно именуют "специальными функциями". Такое деление функций на элементарные и специальные не представляется сегодня оправданным. На самом деле все функции можно разделить на две группы:

- 1) полиномы и рациональные дроби, непосредственно вычисляемые компьютером;
- 2) функции, аппроксимируемые полиномами и рациональными дробями.

⁴³Огюстен Жан ФРЕНЕЛЬ (A.-J. Fresnel, 1788-1827) – французский инженер, физик и математик, член Парижской АН.

Отношение конкретного пользователя к функциям из второй группы определяется его конкретной "вооруженностью": знанием свойств функции и наличием программных средств для эффективного вычисления ее значений. При таком подходе интегральный синус ничем не хуже обычного "школьного" синуса.

Мы рекомендуем читателю ознакомиться с книгой под редакцией М. Абрамовица и И. Стиган "Справочник по специальным функциям М.: Наука, 1979. Такие среды конечного пользователя, как MAPLE, MATHEMATICA "знают" все приведенные в этой книге функции, умеют вычислять их значения с заданной пользователем точностью и даже выполняют над ними многие "аналитические" операции (вычисляют пределы, дифференцируют и т.п.)

14.8. Теорема Ньютона–Лейбница. Формальное интегрирование

Теорема Ньютона–Лейбница. Если f кусочно непрерывна на $[a, b]$ и Ψ – какая-нибудь ее первообразная, то

$$\int_a^b f = \Psi(b) - \Psi(a) \quad (14.8.1)$$

(правую часть этого равенства обозначают также $\Psi \Big|_a^b$).

Доказательство. По теореме Барроу $\int_a^x f$ – одна из первообразных для f . Следовательно, $\Psi(x) = \int_a^x f + const$. Положив $x = a$, получим $const = \Psi(a)$, и $\int_a^x f = \Psi(x) - \Psi(a)$. Положив здесь $x = b$, получим формулу (14.8.1), которая называется *формулой Ньютона–Лейбница*.

Интеграл от кусочно непрерывной функции по сегменту равен приращению любой первообразной этой функции на этом сегменте.

Формула Ньютона–Лейбница сводит задачу о вычислении интеграла к отысканию первообразной для подынтегральной функции

и вычислению приращения этой первообразной на сегменте интегрирования. Этот путь вычисления интеграла мы будем называть "формальным интегрированием". Дело в том, что при отсутствии нужной первообразной в справочнике нет иного алгоритма ее отыскания, кроме представления ее в виде интеграла с переменным пределом! (Мы, естественно, не считаем алгоритмом фразу из известного учебника элементарной математики "... а теперь надо догадаться..."). Тем не менее опишем два приема преобразования интеграла, иногда помогающие выполнить формальное интегрирование.

Прием 1. Интегрирование по частям. Так именуют прием, основанный на известном правиле дифференцирования произведения

$$(f_1 \cdot f_2)' = f'_1 \cdot f_2 + f_1 \cdot f'_2.$$

Отсюда следует равенство интегралов

$$\int_a^b (f_1 \cdot f_2)' = \int_a^b f'_1 \cdot f_2 + \int_a^b f_1 \cdot f'_2. \quad (14.8.2)$$

Учитывая, что первообразной для функции $(f_1 \cdot f_2)'$ является $f_1 \cdot f_2$, и применяя к левой части равенства (14.8.2) формулу Ньютона–Лейбница, получим

$$(f_1 \cdot f_2)(b) - (f_1 \cdot f_2)(a) = \int_a^b f'_1 \cdot f_2 + \int_a^b f_1 \cdot f'_2,$$

откуда и вытекает правило интегрирования "по частям":

$$\int_a^b (f_1 \cdot f'_2) = (f_1 \cdot f_2) \Big|_a^b - \int_a^b (f'_1 \cdot f_2). \quad (14.8.3)$$

Отметим, что интегрирование "по частям" позволяет не вычислить интеграл, а лишь заменить вычисление одного интеграла на вычисление другого. Если пользователь умеет вычислять этот другой, то применение интегрирования "по частям" оправдано.

Рассмотрим технологию интегрирования "по частям" на примерах.

Примеры. 1. $\int_a^b x \cdot \exp(x) dx$. Пусть $f_1(x) = x$, $f'_2(x) = \exp(x)$. Тогда $f'_1(x) \equiv 1$, $f_2(x) = \exp(x)$.

По формуле (14.8.3)

$$\int_a^b x \cdot \exp(x) dx = x \cdot \exp(x) \Big|_a^b - \int_a^b 1 \cdot \exp(x) dx.$$

Мы свели вычисление одного интеграла – $\int_a^b x \cdot \exp(x) dx$ к вычислению другого – $\int_a^b \exp(x) dx$, у которого известна первообразная для подынтегральной функции. Применяя формулу Ньютона–Лейбница, получаем

$$\int_a^b \exp(x) dx = \exp(x) \Big|_a^b.$$

Итак,

$$\int_a^b x \cdot \exp(x) dx = x \cdot \exp(x) \Big|_a^b - \exp(x) \Big|_a^b = (b-1) \cdot \exp(b) - (a-1) \cdot \exp(a).$$

2. $\int_1^2 \ln(x) dx$. Положим $f_1(x) = \ln(x)$, $f'_1(x) \equiv 1$. Тогда $f'_1(x) = \frac{1}{x}$, $f_2(x) = x$, и по формуле (14.8.3)

$$\int_1^2 \ln(x) dx = \ln(x) \cdot x \Big|_1^2 - \int_1^2 \frac{1}{x} \cdot x dx.$$

Мы опять не вычислили интеграл, а заменили его другим. Но этот другой (интеграл от функции-константы) легко вычисляется: $\int_1^2 1 = 1$. Итак,

$$\int_1^2 \ln(x) dx = (\ln(2) \cdot 2 - \ln(1) \cdot 1) - 1 = 2 \ln(2) - 1.$$

Прием 2. Подстановка (иногда говорят "замена переменной но мы не пользуемся этим историческим названием.)

Пусть непрерывно дифференцируемая функция φ взаимно однозначно отображает сегмент $[\alpha, \beta]$ на сегмент $[a, b]$. Найдем такую непрерывную функцию $\psi : [\alpha, \beta] \rightarrow \mathbb{R}$, чтобы для всех $t \in [\alpha, \beta]$ выполнялось равенство

$$\int_{\alpha}^t \psi = \int_a^{\varphi(t)} f.$$

Дифференцируя это тождество, получим

$$\psi(t) = f(\varphi(t)) \cdot \varphi'(t), \quad \text{или} \quad \psi = (f \circ \varphi) \cdot \varphi'.$$

Итак,

$$\int_{\alpha}^t (f \circ \varphi) \cdot \varphi' = \int_a^{\varphi(t)} f.$$

Полагая $t = \beta$ ($\varphi(\beta) = b$), получаем правило подстановки:

$$\int_a^b f = \int_{\alpha}^{\beta} (f \circ \varphi) \cdot \varphi'.$$

Дадим физическую интерпретацию этого правила. Будем трактовать интеграл $\int_a^b f$ (число!) как величину электрического заряда, распределенного по стержню $[a, b]$ с непрерывной линейной плотностью f . Подвергнем этот стержень деформации, т.е. зададим непрерывно дифференцируемую функцию φ , взаимно однозначно отображающую сегмент $[\alpha, \beta]$ на сегмент $[a, b]$, причем $\varphi(\alpha) = a$ и $\varphi(\beta) = b$ (рис.14.6).

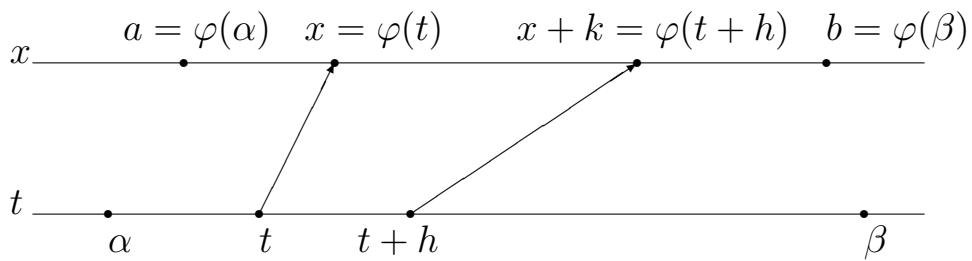


Рис.14.6

Заряд стержня при такой деформации сохраняется, а плотность его распределения вдоль стержня изменится. Обозначим новую плотность ψ . Отрезок стержня $[t, t + h]$ преобразуется при деформации в отрезок $[x, x + k]$. Пусть q – заряд отрезка $[t, t + h]$ (он же заряд отрезка $[x, x + k]$). Запишем выражения для средних плотностей заряда $\psi_{cp} = \frac{q}{h}$ (на отрезке $[t, t + h]$) и $f_{cp} = \frac{q}{k}$ (на отрезке $[x, x + k]$). Отсюда

$$\psi_{cp} = f_{cp} \cdot \frac{k}{h} = f_{cp} \cdot \frac{\varphi(t+h) - \varphi(t)}{h}.$$

По теореме о среднем $\psi_{cp} = \psi(\tau)$, $f_{cp} = f(\xi)$, где τ и ξ – некоторые точки на сегментах $[t, t+h]$ и $[x, x+k]$ соответственно. Переходя к пределу ($h = 0$), получим (вследствие непрерывности функций f и ψ)

$$\psi(t) = f(x) \cdot \varphi'(t) = (f \circ \varphi)(t) \cdot \varphi'(t).$$

Замечания. 1. Обратите внимание, что рис.14.6 соответствует *возрастающей* функции φ , когда $\varphi' > 0$. Если же $\varphi' < 0$, то отрезок $[a, b]$ при отображении "переворачивается" (если $a < b$, то $\alpha > \beta$).

2. Подстановка (как и интегрирование "по частям") лишь преобразует один интеграл в другой. Искусство пользователя состоит в выборе такой подстановки, в результате которой получится "табличный" (известный пользователю) интеграл.

Примеры. 1. $\int_0^a \sqrt{a^2 - x^2} dx, \quad a > 0.$

Функция $x = a \cdot \sin(t)$ взаимно однозначно отображает сегмент $[0, \frac{\pi}{2}]$ на сегмент $[0, a]$ и непрерывно дифференцируема. Поэтому

$$\int_0^a \sqrt{a^2 - x^2} dx = \int_0^{\pi/2} \sqrt{a^2 - (a \cdot \sin(t))^2} \cdot (a \cdot \sin(t))' dt = a^2 \int_0^{\pi/2} \cos^2(t) dt.$$

Пока что мы только заменили один интеграл другим. Но известная из школы формула $\cos^2(t) = \frac{1 + \cos(2t)}{2}$ дает

$$\int_0^{\pi/2} \cos^2(t) dt = \frac{1}{2} \left(\int_0^{\pi/2} 1 dt + \int_0^{\pi/2} \cos(2t) dt \right) = \frac{\pi}{4} + \frac{\sin(2t)}{4} \Big|_0^{\pi/2} = \frac{\pi}{4}.$$

Итак, $\int_0^a \sqrt{a^2 - x^2} dx = \frac{\pi a^2}{4}$ (заметим, что мы вычислили площадь четверти круга с радиусом a).

2. $\int_0^2 \frac{\sin(x)}{\sqrt{x}} dx$. Отметим, что хотя подынтегральная функция не определена в нуле, это не мешает интегралу существовать в силу наличия конечного правого предела: $\lim_{x=0+} \frac{\sin(x)}{\sqrt{x}} = 0$.

Функция $x = \frac{\pi}{2} t^2$ взаимно однозначно отображает сегмент $[0, \frac{2}{\sqrt{\pi}}]$ на сегмент $[0, 2]$ и непрерывно дифференцируема. Поэтому

$$\int_0^2 \frac{\sin(x)}{\sqrt{x}} dx = \int_0^{2/\sqrt{\pi}} \frac{\sin(\frac{\pi}{2}t^2)}{\sqrt{\frac{\pi}{2}t^2}} \left(\frac{\pi}{2}t^2\right)' dt = \sqrt{2\pi} \int_0^{2/\sqrt{\pi}} \sin(\frac{\pi}{2}t^2) dt.$$

Опять мы лишь преобразовали один интеграл в другой. Но если пользователь знает о существовании функции $S(x) = \int_0^x \sin(\frac{\pi}{2}t^2) dt$, называемой интегралом Френеля (мы о ней упоминали в п.14.7), то ситуация меняется:

$$\int_0^2 \frac{\sin(x)}{\sqrt{x}} dx = \sqrt{2\pi} \int_0^{2/\sqrt{\pi}} \sin(\frac{\pi}{2}t^2) dt = \sqrt{2\pi} S\left(\frac{2}{\sqrt{\pi}}\right).$$

Серьезное предупреждение. Этот пример показывает, что успех формального интегрирования зависит от тезауруса пользователя (который, конечно, следует расширять) и от его искусства, которое приобретается лишь долгой тренировкой и дается далеко не всем.

Существуют достаточно богатые таблицы первообразных, называемых по старинке "неопределенными интегралами и даже интегралов с типичными пределами интегрирования. Наиболее полные из них:

- И.С. Градштейн и И.М. Рыжик. Таблицы интегралов, сумм, рядов и произведений. Наука. М.: 1971;
- А.П. Прудников и др. Интегралы и ряды. Наука, М.: 1981;
- А.П. Прудников и др. Интегралы и ряды (специальные функции). Наука, М.: 1983;
- А.П. Прудников и др. Интегралы и ряды (дополнительные главы). Наука, М.: 1986.

Следует учесть, что роль этих превосходных таблиц в наше время уменьшается, так как основное их содержание реализовано в таких средах конечного пользователя, как MATHEMATICA и MAPLE, которые "умеют" и выполнять формальное интегрирование, и находить значение интеграла – число – с заданной пользователем точностью.

Мы считаем, что "цивилизованный пользователь" должен помнить несколько простейших первообразных, уметь грамотно провести интегрирование "по частям" и подстановку, но отнюдь не должен владеть искусством подбора подходящих путей преобразования интеграла.

Следует научиться работать с поименованными выше таблицами и средами конечного пользователя, а в сложных случаях – консультироваться со специалистами.

14.9. Численное интегрирование

В реальных задачах обычно требуется найти не сам интеграл (число), а некоторую его оценку, т.е. *интервал достаточно малой длины*, гарантированно накрывающий это число. Один из способов получения оценки интеграла мы опишем.

Для оценки интеграла $\int_a^b f$ подынтегральная функция заменяется некоторой функцией ϕ , которая должна удовлетворять двум требованиям:

- 1) должна быть известна ее первообразная, чтобы можно было применить формулу Ньютона–Лейбница;
- 2) абсолютная погрешность от замены интеграла от f на интеграл от ϕ не должна превышать заданное положительное число ε .

$$R = \left| \int_a^b f - \int_a^b \phi \right| \leq \varepsilon.$$

В качестве аппроксимирующей функции часто употребляются полиномиальные сплайны. При этом $\int_a^b f$ оказывается линейной комбинацией значений подынтегральной функции в узлах стандартной сетки и со стандартными коэффициентами. Эту линейную комбинацию называют *квадратурной формулой*.

Проиллюстрируем изложенное на простейшем примере. Возьмем на $[a, b]$ равномерную сетку

$$x_k = a + k \cdot h \quad (k = 0, \dots, n), \quad h = \frac{b - a}{n}.$$

Вычислим значения подынтегральной функции в середине каждого сегмента $J_k = [x_{k-1}, x_k]$ и построим кусочно постоянный сплайн

$$Spl(x) = f\left(\frac{x_{k-1} + x_k}{2}\right) \quad \text{при } x \in J_k.$$

Заметим, что согласно свойству 3 из п.14.3 значения сплайна в узлах сетки не влияют на значения интеграла.

Вычислим интеграл от построенного сплайна

$$\int_a^b Spl = \sum_{k=1}^n \int_{x_{k-1}}^{x_k} Spl = \sum_{k=1}^n \int_{x_{k-1}}^{x_k} f\left(\frac{x_{k-1} + x_k}{2}\right) dx = h \cdot \sum_{k=1}^n f\left(\frac{x_{k-1} + x_k}{2}\right).$$

Итак, мы получили квадратурную формулу

$$\int_a^b f \sim \int_a^b Spl = h \cdot \sum_{k=1}^n f\left(\frac{x_{k-1} + x_k}{2}\right), \quad (14.9.1)$$

которая из-за ее очевидной геометрической интерпретации называется *формулой средних прямоугольников*.

Напоминаем, что значок \sim читается "заменяется на". Мы обращаем внимание читателя на бессмысленность часто употребляющегося выражения "приближенное равенство".

Осталось оценить погрешность квадратурной формулы. И, хотя интеграл Римана определен для всех кусочно непрерывных функций, эффективную оценку погрешности можно получить только при существенном ужесточении требований к подынтегральной функции. Мы потребуем наличия у нее *непрерывной второй производной*! Обозначим

$$M_2 = \max_{x \in [a,b]} \{|f''(x)|\}, \quad M_2^{(k)} = \max_{x \in J_k} \{|f''(x)|\}.$$

Очевидно, что $M_2^{(k)} \leq M_2$ для всех k .

Пусть $x \in]x_{k-1}, x_k[$. По формуле Тейлора

$$\begin{aligned} f(x) &= f\left(\frac{x_{k-1} + x_k}{2}\right) + f'\left(\frac{x_{k-1} + x_k}{2}\right) \cdot \left(x - \frac{x_{k-1} + x_k}{2}\right) + \\ &\quad + \frac{f''(\xi(x))}{2!} \cdot \left(x - \frac{x_{k-1} + x_k}{2}\right)^2, \end{aligned}$$

где $\xi(x)$ – некоторая (неизвестная) точка на интервале $]x_{k-1}, x_k[$. Отсюда

$$\begin{aligned} \int_{x_{k-1}}^{x_k} (f - Spl) &= f'\left(\frac{x_{k-1} + x_k}{2}\right) \cdot \int_{x_{k-1}}^{x_k} \left(x - \frac{x_{k-1} + x_k}{2}\right) dx + \\ &\quad + \int_{x_{k-1}}^{x_k} \frac{f''(\xi(x))}{2!} \cdot \left(x - \frac{x_{k-1} + x_k}{2}\right)^2 dx = \int_{x_{k-1}}^{x_k} \frac{f''(\xi(x))}{2!} \cdot \left(x - \frac{x_{k-1} + x_k}{2}\right)^2 dx \end{aligned}$$

(убедитесь в том, что $\int_{x_{k-1}}^{x_k} \left(x - \frac{x_{k-1} + x_k}{2} \right) dx = 0$). Далее,

$$\begin{aligned} & \left| \int_{x_{k-1}}^{x_k} (f - Spl) \right| = \frac{1}{2} \left| \int_{x_{k-1}}^{x_k} f''(\xi(x)) \cdot \left(x - \frac{x_{k-1} + x_k}{2} \right)^2 dx \right| \leq \\ & \leq \frac{1}{2} \int_{x_{k-1}}^{x_k} |f''(\xi(x))| \cdot \left(x - \frac{x_{k-1} + x_k}{2} \right)^2 dx \leq \frac{M_2^{(k)}}{2} \cdot \int_{x_{k-1}}^{x_k} \left(x - \frac{x_{k-1} + x_k}{2} \right)^2 dx = \\ & = \frac{M_2^{(k)}}{6} \cdot \left(x - \frac{x_{k-1} + x_k}{2} \right)^3 \Big|_{x_{k-1}}^{x_k} = \frac{M_2^{(k)} h^3}{24} = \frac{M_2^{(k)} (b-a)^3}{24n^3}. \end{aligned}$$

И, наконец,

$$\begin{aligned} R = & \left| \int_a^b (f - Spl) \right| = \left| \sum_{k=1}^n \int_{x_{k-1}}^{x_k} (f - Spl) \right| \leq \sum_{k=1}^n \left| \int_{x_{k-1}}^{x_k} (f - Spl) \right| \leq \\ & \leq \frac{(b-a)^3}{24n^3} \sum_{k=1}^n M_2^{(k)} \leq M_2 \cdot \frac{(b-a)^3}{24n^2}. \end{aligned}$$

Нами доказана

Теорема. Если функция f имеет непрерывную вторую производную, то $\int_a^b f \in [S_n - \Delta_n, S_n + \Delta_n]$, где

$$S_n = \frac{b-a}{n} \cdot \sum_{k=1}^n f\left(\frac{x_{k-1} + x_k}{2}\right), \quad \Delta_n = M_2 \cdot \frac{(b-a)^3}{24n^2}.$$

Видно, что качество оценки можно улучшать, увеличивая количество точек сетки. При этом, увеличив *вдвое* объем вычислительной работы, мы уменьшаем радиус оценки в *четыре* раза.

Замечания. 1. Полученная оценка работает при наличии у подынтегральной функции *непрерывной второй производной*. Можно показать, что при наличии у подынтегральной функции только *непрерывной первой производной* оценка ухудшается:

$$\Delta_n = M_1 \cdot \frac{(b-a)^3}{24n}, \quad \text{где} \quad M_1 = \max_{x \in [a,b]} \{|f'(x)|\}.$$

Теперь в знаменателе первая степень числа узлов сетки вместо второй, т.е. увеличив *вдвое* объем вычислительной работы, мы уменьшаем радиус оценки *только в два раза*!

Поскольку обычно приходится интегрировать кусочно аналитические функции, имеющие на сегменте лишь несколько особых точек, целесообразно применять квадратурную формулу не ко всему сегменту сразу, а оценивать интегралы по каждому из отрезков аналитичности подынтегральной функции в отдельности.

2. Получение оценки погрешности квадратурной формулы связано с необходимостью находить наибольшее значение (или какую-нибудь верхнюю границу) модуля второй производной (для метода средних прямоугольников) или производной более высокого порядка (для некоторых других методов). Эта задача не может быть алгоритмизирована. Поэтому стандартные программы численного интегрирования используют так называемые апостериорные (получаемые в процессе вычислений) оценки погрешности. Простейший способ состоит в последовательном удвоении количества узлов сетки. При этом сравнивают "соседние" результаты, полученные по квадратурной формуле (S_2 с S_4 ; S_4 с S_8 и т.д.) Процесс заканчивается при выполнении неравенства $|S_n - S_{2n}| < \varepsilon$ (абсолютная погрешность) или неравенства $|S_n - S_{2n}| < \varepsilon \cdot |S_{2n}|$ (относительная погрешность).

Фортран-библиотеки (NAG, IMSL) содержат большое количество процедур, дающих, *как правило* (если подынтегральная функция не очень плохая), достоверные оценки интегралов. Среды конечного пользователя (MATHEMATICA, MAPLE, MATLAB) позволяют получать оценки интегралов, не прибегая к помощи алгоритмического языка.

Серьезное предупреждение. Необходимо помнить, что при отсутствии *априорных* оценок (основанных на знании верхних границ модуля производных) гарантировать достоверность результатов нельзя. Справедливо утверждение: для любого алгоритма численного интегрирования, основанного на *апостериорных* оценках, можно построить пример, на котором этот алгоритм "сломается".

Построим такой пример для метода средних прямоугольников. Требуется оценить интеграл

$$\int_0^{2\pi} \sin^2(32x) dx = \frac{1}{2} \int_0^{2\pi} (1 - \cos(64x)) dx = \frac{1}{2} \left(2\pi - \frac{1}{64} \sin(64x) \Big|_0^{2\pi} \right) = \pi.$$

Начинаем вычисления по квадратурной формуле, удваивая количество узлов сетки.

$$S_2 = \frac{2\pi}{2} \cdot \left(\sin\left(32 \frac{\pi}{2}\right) + \sin\left(32 \frac{3\pi}{2}\right) \right) = 0.$$

$$S_4 = \frac{2\pi}{4} \cdot \left(\sin\left(32 \frac{\pi}{4}\right) + \sin\left(32 \frac{3\pi}{4}\right) + \sin\left(32 \frac{5\pi}{4}\right) + \sin\left(32 \frac{7\pi}{4}\right) \right) = 0.$$

Два "соседних" результата совпали, и программа выдаст ответ: "ноль!". Даже если мы (на всякий случай) еще раз удвоим количество узлов, ответ не изменится: мы опять попадем в точки, где подынтегральная функция равна нулю.

Из этого примера не следует, конечно, что стандартные программы численного интегрирования нельзя использовать. Однако если пользоваться ими *без предварительного исследования поведения подынтегральной функции*, полученный результат может не иметь никакого отношения к правильному ответу.

Глава 15. КРАТНЫЕ ИНТЕГРАЛЫ РИМАНА

15.1. Определение двойного интеграла

Пусть $\Delta = [a, b] \times [c, d]$ – прямоугольник, и $f : \Delta \rightarrow \mathbb{R}$ – кусочно непрерывная функция. Будем временно обозначать координаты точки на плоскости не x_1 и x_2 , как обычно, а x и y .

Построим на сегментах $[a, b]$ и $[c, d]$ сетки

$$\{a = x_0, \dots, x_k = b\} \quad \text{и} \quad \{c = y_0, \dots, y_n = b\}.$$

Эти сетки порождают разбиение P прямоугольника Δ , состоящее из $k \cdot n$ элементарных прямоугольников

$$\Delta_{ij} = [x_{i-1}, x_i] \times [y_{j-1}, y_j], \quad (i = 1, \dots, k; j = 1, \dots, n).$$

Обозначим $S_{ij} = (x_i - x_{i-1}) \cdot (y_j - y_{j-1})$ площадь элементарного прямоугольника Δ_{ij} .

По аналогии с одномерным интегралом Римана введем суммы Дарбу, соответствующие разбиению P :

$$L(f, P) = \sum_{i=1}^k \sum_{j=1}^n m_{ij} S_{ij}, \quad U(f, P) = \sum_{i=1}^k \sum_{j=1}^n M_{ij} S_{ij}.$$

Здесь $m_{ij} = \inf_{\Delta_{ij}} \{f(x, y)\}$ и $M_{ij} = \sup_{\Delta_{ij}} \{f(x, y)\}$ – соответственно нижняя и верхняя грани множества значений функции f на элементарном прямоугольнике Δ_{ij} .

Свойства сумм Дарбу для двойного интеграла совпадают с уже известными свойствами этих сумм для одномерного интеграла. В частности, если сетки на обеих осях состоят из двух точек каждая, то "разбиение" содержит всего один прямоугольник, и этому разбиению соответствуют: $m(b-a)(d-c)$ – наименьшая из нижних и $M(b-a)(d-c)$ – наибольшая из верхних сумм Дарбу. Кроме того, любая нижняя сумма Дарбу не больше любой верхней.

Теорема. Если $f : \Delta \rightarrow \mathbb{R}$ – кусочно непрерывная функция, то существует ровно одно число, которое не меньше любой нижней суммы Дарбу и не больше любой верхней. Это число, разделяющее множество нижних и множество верхних сумм Дарбу, называют *двойным интегралом Римана* от функции f по прямоугольнику Δ .

Обозначают двойной интеграл Римана так:

$$\iint_{\Delta} f \quad \text{или} \quad \iint_{\Delta} f(x, y) dx dy.$$

Приводя эту теорему без доказательства, еще раз подчеркнем, что двойной интеграл Римана – *единственное* число, удовлетворяющее неравенству

$$L(f, P) \leq \iint_{\Delta} f \leq U(f, P) \quad (15.1.1)$$

при любом разбиении P прямоугольника Δ , и эта единственность обеспечивается кусочной непрерывностью функции f . При произвольной f может быть много чисел, удовлетворяющих неравенству (15.1.1).

Пусть теперь Ω – часть плоскости, ограниченная замкнутой кусочно гладкой кривой, и $f: \Omega \rightarrow \mathbb{R}$ – кусочно непрерывная функция. Построим *какой-нибудь* прямоугольник Δ , содержащий Ω , и определим на нем функцию

$$F(x, y) = \begin{cases} f(x, y), & \text{если } (x, y) \in \Omega; \\ 0, & \text{если } (x, y) \in \Delta \setminus \Omega. \end{cases} \quad (15.1.2)$$

Двойной интеграл Римана от f по области Ω определим так:

$$\iint_{\Omega} f = \iint_{\Delta} F.$$

"Физическое" обоснование этого определения понятно; несложно увидеть также, что это определение не зависит от выбора Δ , ибо те элементарные прямоугольники, на которых $F = 0$, дают нулевой вклад в суммы Дарбу.

Замечание. Если $f(x, y) \equiv 1$, а $\Delta = [a, b] \times [c, d]$, то, очевидно,

$$\iint_{\Delta} f = (b - a) \cdot (d - c) = S(\Delta) \quad (\text{площадь прямоугольника } \Delta).$$

Если Ω – часть плоскости, ограниченная кусочно гладкой кривой, а $f \equiv const = 1$, то число

$$\iint_{\Omega} f = \iint_{\Omega} 1 dx dy$$

естественно считать *по определению* площадью фигуры Ω . Мы будем обозначать эту площадь $S(\Omega)$.

Двойной интеграл Римана, так же, как и однократный, допускает различные физические и геометрические интерпретации. Приведем два примера.

1. Электрический заряд распределен по пластине с поверхностной плотностью ρ (кулонов на квадратный метр). Тогда полный заряд пластины $Q = \iint_{\Omega} \rho$.

2. Рассмотрим цилиндр с поперечным сечением Ω и образующей, параллельной оси Oz . Пусть тело M – часть этого цилиндра, ограниченная снизу плоскостью $z = 0$, а сверху – графиком неотрицательной непрерывной функции $f: \Omega \rightarrow \mathbb{R}$ (рис.15.1). Тогда $V(M) = \iint_{\Omega} f$ – объем тела M .

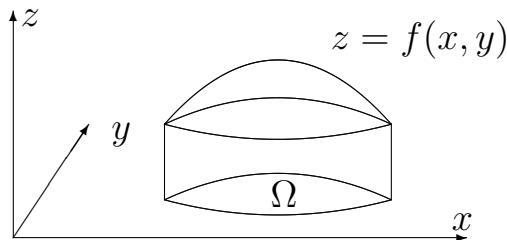


Рис.15.1

В следующем пункте мы рассмотрим еще одну интерпретацию двойного интеграла.

15.2. Площадь поверхности

Пусть $G = r(\Omega)$ – гладкая поверхность. Мы хотим придать смысл понятию *площадь поверхности*.

Для начала будем считать, что Ω – прямоугольник. Возьмем какое-нибудь его разбиение P и рассмотрим элементарный прямоугольник этого разбиения $\Delta_{ij} = [u_i - u_{i-1}] \times [v_j - v_{j-1}]$. Если бы на Δ_{ij} матрица Якоби r' была постоянной, то сужение r на Δ_{ij} имело бы вид

$$r(u, v) = r(u_{i-1}, v_{j-1}) + A \cdot \begin{bmatrix} u - u_{i-1} \\ v - v_{j-1} \end{bmatrix},$$

где $A = r' = [a^{(1)}, a^{(2)}]$ – (3×2) -матрица.

Образом прямоугольника Δ_{ij} при таком отображении был бы параллелограмм (рис.15.2), построенный на направленных отрезках

$$\overrightarrow{p} = \overrightarrow{a^{(1)}} \cdot (u_i - u_{i-1}) \quad \text{и} \quad \overrightarrow{q} = \overrightarrow{a^{(2)}} \cdot (v_j - v_{j-1}).$$

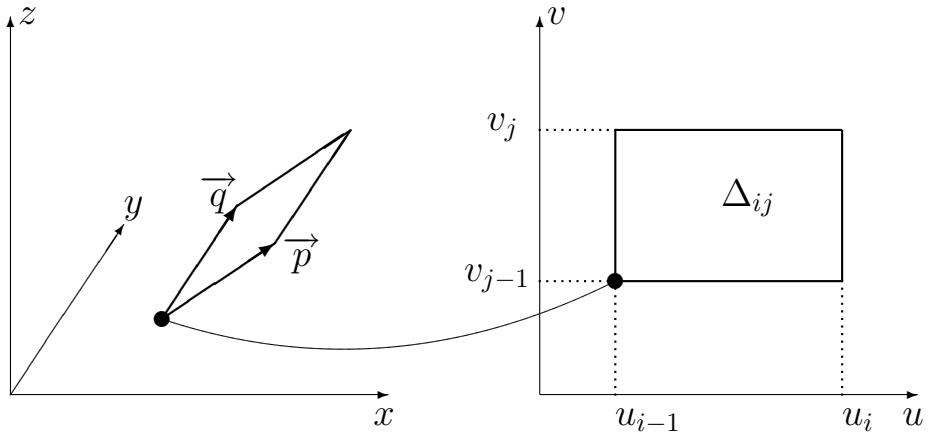


Рис.15.2

Из курса линейной алгебры известно, что площадь этого параллелограмма равна

$$\left| \det([a^{(1)}, a^{(2)}, w]) \right| \cdot S_{ij},$$

где S_{ij} – площадь Δ_{ij} , а w – нормированный вектор, ортогональный $a^{(1)}$ и $a^{(2)}$.

Поскольку определитель матрицы не меняется при ее транспонировании,

$$\begin{aligned} \left| \det([a^{(1)}, a^{(2)}, w]) \right| &= \left(\det([a^{(1)}, a^{(2)}, w]^T \cdot [a^{(1)}, a^{(2)}, w]) \right)^{1/2} = \\ &= \left(\det \begin{bmatrix} b_{11} & b_{12} & 0 \\ b_{21} & b_{22} & 0 \\ 0 & 0 & 1 \end{bmatrix} \right)^{1/2} = (\det(B))^{1/2}, \end{aligned}$$

$$\text{где } B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = A^T \cdot A = (r')^T \cdot r'.$$

Итак, если бы r' была постоянной на Δ_{ij} , то площадь куска поверхности $r(\Delta_{ij})$ равнялась бы

$$(\det((r')^T \cdot r'))^{1/2} \cdot S_{ij}.$$

Естественно предположить, что истинная площадь σ_{ij} куска поверхности $r(\Delta_{ij})$ должна удовлетворять неравенству

$$\inf_{\Delta_{ij}} \left\{ (\det((r')^T \cdot r'))^{1/2} \right\} \cdot S_{ij} \leq \sigma_{ij} \leq \sup_{\Delta_{ij}} \left\{ (\det((r')^T \cdot r'))^{1/2} \right\} \cdot S_{ij}.$$

Суммируя такие неравенства по всем i и j , получим

$$L \left((\det((r')^T \cdot r'))^{1/2}, P \right) \leq \sigma \leq U \left((\det((r')^T \cdot r'))^{1/2}, P \right).$$

Поскольку лишь одно число – двойной интеграл – удовлетворяет этому неравенству при любом разбиении P , мы заключаем, что

$$\sigma = \iint_{\Omega} (\det((r')^T \cdot r'))^{1/2}. \quad (15.2.2)$$

Замечание. Формула (15.2.2) определяет площадь гладкой поверхности. Естественно распространить это определение на случай кусочно гладкой поверхности, а также на случай, когда Ω – произвольная область в \mathbb{R}^2 , ограниченная кусочно гладкой замкнутой кривой.

Примеры. 1. Как известно, график непрерывно дифференцируемой функции $f: \Omega \rightarrow \mathbb{R}$ – гладкая поверхность (см. пример в п.11.3). В этом случае

$$r' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ D_u f & D_v f \end{bmatrix}, \quad (r')^T \cdot r' = \begin{bmatrix} 1 + (D_u f)^2 & D_u f \cdot D_v f \\ D_u f \cdot D_v f & 1 + (D_v f)^2 \end{bmatrix},$$

и формула (15.2.2) принимает вид

$$\sigma = \iint_{\Omega} (1 + (D_u f)^2 + (D_v f)^2)^{1/2} \, dudv.$$

2. Пусть непрерывно дифференцируемая функция $\varphi: \Omega \rightarrow G$ взаимно однозначно отображает область $\Omega \subset \mathbb{R}^2$ на область $G \subset \mathbb{R}^2$. Тогда область G можно рассматривать как гладкую поверхность, задаваемую отображением (рис.15.3)

$$r: \Omega \rightarrow \mathbb{R}^3; \quad r(u, v) = \begin{bmatrix} \varphi_1(u, v) \\ \varphi_2(u, v) \\ 0 \end{bmatrix}; \quad (r')^T \cdot r' = (\varphi')^T \cdot \varphi'.$$

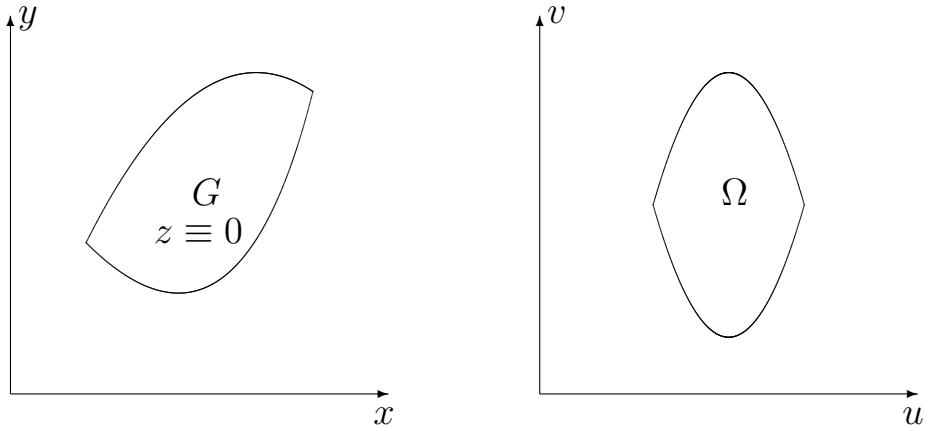


Рис.15.3

Поскольку φ' – квадратная матрица, $(\det((\varphi')^T \cdot \varphi'))^{1/2} = |\det(\varphi')|$, и формула (15.2.2) в этом случае перепишется так:

$$S(G) = \iint_{\Omega} |\det(\varphi')|. \quad (15.2.3)$$

В частном случае, когда $G = \Omega$, мы имеем $\varphi' = I_2$ и $S(G) = \iint_{\Omega} 1$, т.е. мы вновь пришли к формуле (15.1.3).

15.3. Сведение двойного интеграла к повторному

В этом пункте мы проведем *правдоподобное рассуждение* (не следует принимать его за доказательство), основанное на интерпретации двойного интеграла как электрического заряда, распределенного по прямоугольнику $\Delta = [a, b] \times [c, d]$ с поверхностной плотностью f .

Возьмем на сегменте $[a, b]$ произвольную точку с абсциссой x и "соберем" в ней заряд из всех точек прямоугольника, имеющих эту абсциссу (рис.15.4).

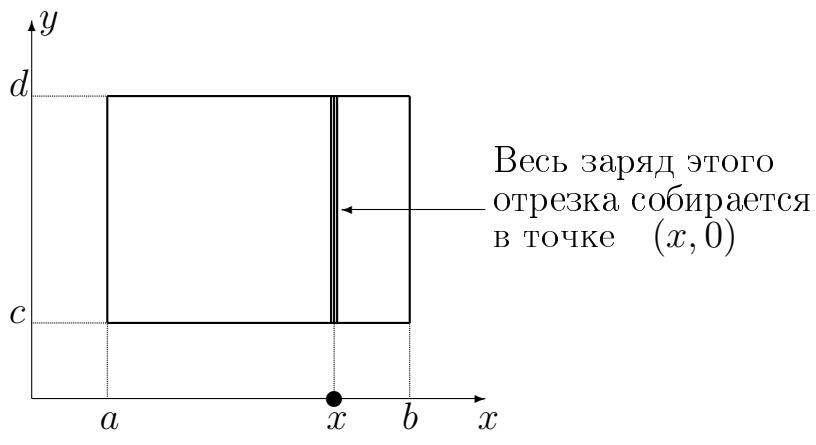


Рис.15.4

Проделав такую операцию для всех точек сегмента $[a, b]$, мы соберем заряд прямоугольника на этом сегменте. "Физически очевидно что линейная плотность заряда будет равна

$$q(x) = \int_c^d f(x, y) dy$$

(подынтегральная функция есть функция только переменной y , так как x фиксирован. Эта функция – сужение f на отрезок, изображенный жирной линией на рис.15.4).

Полный заряд сегмента (бывший ранее полным зарядом прямоугольника) находится по уже известному правилу:

$$Q = \int_a^b q(x) dx.$$

Наше рассуждение привело к формуле

$$\iint_{\Delta} f = \int_a^b \left(\int_c^d f(x, y) dy \right) dx. \quad (15.3.1)$$

Можно показать, что эта формула справедлива для всякой кусочно непрерывной функции f . Равноправие координат позволяет записать формулу, аналогичную (15.3.1):

$$\iint_{\Delta} f = \int_c^d \left(\int_a^b f(x, y) dx \right) dy. \quad (15.3.2)$$

Интегралы, стоящие в правых частях формул (15.3.1) и (15.3.2), называют *повторными*.

Пример. Пусть $\Delta = [0, 2] \times [0, 1]$, $f(x, y) = \sin(x + 2y)$. Вычислим двойной интеграл двумя способами – по формулам (15.3.1) и (15.3.2).

$$1) \quad \iint_{\Delta} f = \int_0^2 \left(\int_0^1 \sin(x + 2y) dy \right) dx.$$

Вычисляем "внутренний" интеграл (x фиксирован):

$$\int_0^1 \sin(x + 2y) dy = -\frac{1}{2} \cdot \cos(x + 2y) \Big|_{y=0}^{y=1} = \frac{1}{2} \cdot (\cos(x) - \cos(x + 2)).$$

Вычисляем "внешний" интеграл:

$$\begin{aligned} \frac{1}{2} \cdot \int_0^2 (\cos(x) - \cos(x + 2)) dx &= \frac{1}{2} \cdot (\sin(x) - \sin(x + 2)) \Big|_0^2 = \\ &= \frac{1}{2} \cdot (\sin(2) - \sin(4) + \sin(2)) = \sin(2) - \frac{\sin(4)}{2}. \end{aligned}$$

$$2) \quad \iint_{\Delta} f = \int_0^1 \left(\int_0^2 \sin(x + 2y) dx \right) dy.$$

Вычисляем "внутренний" интеграл (y фиксирован):

$$\int_0^2 \sin(x + 2y) dx = -\cos(x + 2y) \Big|_{x=0}^{x=2} = \cos(2y) - \cos(2 + 2y).$$

Вычисляем "внешний" интеграл:

$$\begin{aligned} \int_0^1 (\cos(2y) - \cos(2 + 2y)) dy &= \frac{1}{2} \cdot (\sin(2y) - \sin(2 + 2y)) \Big|_0^1 = \\ &= \frac{1}{2} \cdot (\sin(2) - \sin(4) + \sin(2)) = \sin(2) - \frac{\sin(4)}{2}. \end{aligned}$$

Замечание. При сведении двойного интеграла к повторному техническая сложность вычисления получающихся одномерных интегралов может существенно зависеть от выбранного порядка интегрирования.

15.4. Вычисление двойного интеграла от функции, заданной на "криволинейной трапеции"

"Криволинейной трапецией" мы называем область $\Omega \subset \mathbb{R}^2$, ограниченную прямыми $x = a$, $x = b$ ($b > a$) и графиками кусочно гладких функций φ , ψ , заданных на $[a, b]$ и удовлетворяющих неравенству $\varphi \leq \psi$ (рис.15.5).

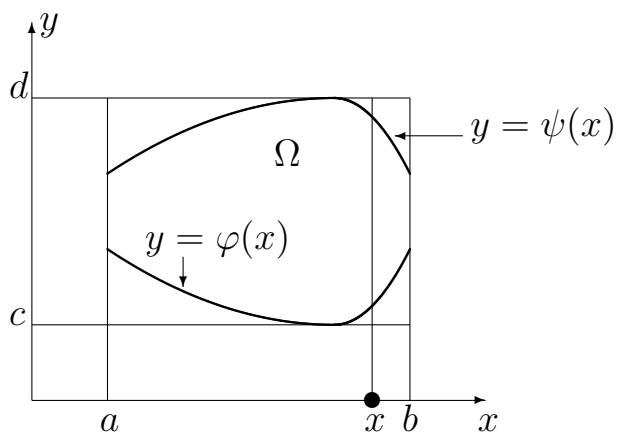


Рис.15.5

Пусть $f : \Omega \rightarrow \mathbb{R}$ – кусочно непрерывная функция. Требуется вычислить интеграл $\iint_{\Omega} f$.

Обозначим

$$c = \min_{[a,b]} \{\varphi(x)\}, \quad d = \max_{[a,b]} \{\psi(x)\}.$$

Тогда прямоугольник $\Delta = [a, b] \times [c, d]$ будет содержать Ω . Определим на Δ функцию F по формуле (15.1.2) и сведем двойной интеграл к повторному

$$\iint_{\Omega} f = \iint_{\Delta} F = \int_a^b \left(\int_c^d F(x, y) dy \right) dx.$$

Вычислим "внутренний" интеграл (x фиксирован – см. рис.15.5)

$$\int_c^d F(x, y) dy = \int_0^{\varphi(x)} F(x, y) dy + \int_{\varphi(x)}^{\psi(x)} F(x, y) dy + \int_{\psi(x)}^d F(x, y) dy. \quad (15.4.1)$$

Если $c \leq y < \varphi(x)$ или $\psi(x) < y \leq d$, то точка (x, y) не принадлежит Ω , и по построению $F(x, y) = 0$. Если $\varphi(x) \leq y \leq \psi(x)$, то точка (x, y) принадлежит Ω , и $F(x, y) = f(x, y)$.

Таким образом, подынтегральная функция в первом и третьем интегралах из (15.4.1) равна нулю тождественно, и

$$\int_c^d F(x, y) dy = \int_{\varphi(x)}^{\psi(x)} F(x, y) dy = \int_{\varphi(x)}^{\psi(x)} f(x, y) dy.$$

Мы получили правило сведения двойного интеграла к повторному для случая функции, заданной на "криволинейной трапеции":

$$\iint_{\Omega} f = \int_a^b \left(\int_{\varphi(x)}^{\psi(x)} f(x, y) dy \right) dx.$$

Аналогично, если "криволинейная трапеция" $\Omega \subset \mathbb{R}^2$ ограничена прямыми $y = a$, $y = b$ ($b > a$) и графиками кусочно гладких функций φ, ψ , заданных на $[a, b]$ и удовлетворяющих неравенству $\varphi \leq \psi$ (рис.15.6), то

$$\iint_{\Omega} f = \int_a^b \left(\int_{\varphi(y)}^{\psi(y)} f(x, y) dx \right) dy.$$

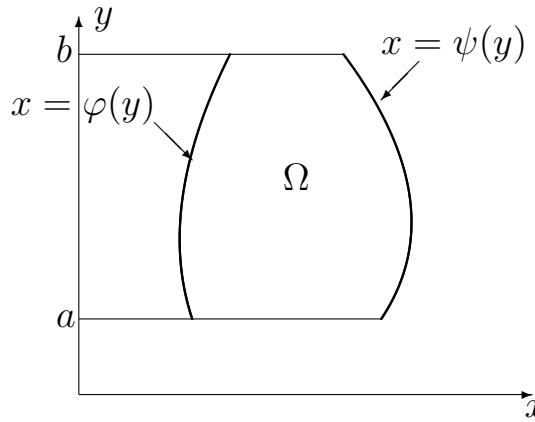


Рис.15.6

Пример. Пусть Ω – часть плоскости, ограниченная линиями $y = 0$, $y = x$, $x = 1$ (рис.15.7), $f(x, y) = \exp(-(x^2 + y^2)) \cdot y$. Тогда

$$\iint_{\Omega} f = \int_0^1 \left(\int_0^x \exp(-(x^2 + y^2)) \cdot y \, dy \right) dx.$$

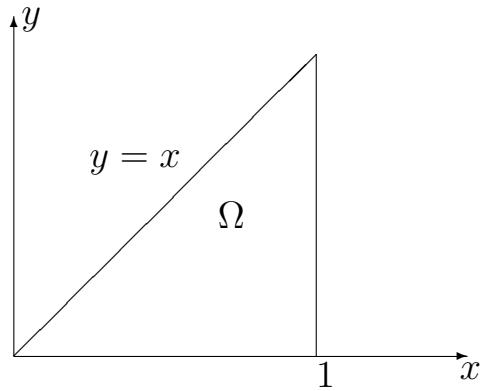


Рис.15.7

Вычисляем "внутренний" интеграл:

$$\begin{aligned} \int_0^x \exp(-(x^2 + y^2)) \cdot y \, dy &= \exp(-x^2) \cdot \int_0^x y \cdot \exp(-y^2) \, dy = \\ &= -\frac{1}{2} \cdot \exp(-x^2) \cdot \exp(-y^2) \Big|_{y=0}^{y=x} = \frac{1}{2} \cdot (\exp(-x^2) - \exp(-2x^2)). \end{aligned}$$

Вычисляем "внешний" интеграл:

$$\begin{aligned} \iint_{\Omega} f &= \frac{1}{2} \int_0^1 (\exp(-x^2) - \exp(-2x^2)) \, dx = \\ &= \frac{\sqrt{\pi}}{4} \left(\operatorname{erf}(1) - \frac{1}{\sqrt{2}} \cdot \operatorname{erf}\left(\frac{1}{\sqrt{2}}\right) \right). \end{aligned}$$

Здесь использована формула (см. пример 1 п.13.7)

$$\int_0^x \exp(-a^2 t^2) dt = \frac{1}{|a|} \cdot \int_0^{x/|a|} \exp(-u^2) du = \frac{\sqrt{\pi}}{2|a|} \cdot \operatorname{erf}\left(\frac{x}{|a|}\right).$$

Попробуйте вычислить этот двойной интеграл, используя вместо формулы (15.3.1) формулу (15.3.2). Вы убедитесь, что разные способы сведения двойного интеграла к повторному могут оказаться существенно разными по технической сложности их реализации.

15.5. Простейшие свойства двойного интеграла

Как и в случае одномерного интеграла, ограничимся перечислением: доказательства сложны, а то, что выдается обычно за доказательства, – в лучшем случае правдоподобные рассуждения.

Начиная с этого пункта мы возвращаемся к обычным обозначениям: x и y – точки плоскости (векторы в \mathbb{R}^2), x_1, x_2 и y_1, y_2 – их координаты.

1. Если f – функция-константа $f \equiv c \in \mathbb{R}$, то $\iint_{\Omega} f = c \cdot S(\Omega)$.

2. Если $\Omega = \Omega_1 \cup \Omega_2$, причем Ω_1 и Ω_2 не имеют общих точек, кроме граничных ("не налегают" друг на друга), и имеют кусочно гладкие границы, а f кусочно непрерывна на Ω , то

$$\iint_{\Omega} f = \iint_{\Omega_1} f + \iint_{\Omega_2} f.$$

(Если заряженную пластину разрезать на части, то заряд всей пластины равен сумме зарядов ее частей).

3. Для любых кусочно непрерывных на Ω функций f_1, f_2 и любых чисел α_1, α_2

$$\iint_{\Omega} (\alpha_1 f_1 + \alpha_2 f_2) = \alpha_1 \iint_{\Omega} f_1 + \alpha_2 \iint_{\Omega} f_2$$

(линейность интеграла).

4. Если $f_1, f_2 : \Omega \rightarrow \mathbb{R}$ – кусочно непрерывные функции, причем $f_1 \leq f_2$, то

$$\iint_{\Omega} f_1 \leq \iint_{\Omega} f_2.$$

5. Средним значением функции f на Ω называют число

$$\frac{1}{S(\Omega)} \cdot \iint_{\Omega} f.$$

Как и в случае одномерного интеграла, среднее значение функции не всегда является одним из ее значений, однако если f непрерывна на Ω , то найдется такая точка $x_{cp} \in \Omega$, что

$$\frac{1}{S(\Omega)} \cdot \iint_{\Omega} f = f(x_{cp}).$$

Это утверждение называется *теоремой о среднем*.

15.6. Преобразование двойного интеграла подстановкой

Будем по-прежнему интерпретировать двойной интеграл как электрический заряд области $\Omega \subset \mathbb{R}^2$, распределенный на ней с поверхностной плотностью f . Подвергнем область деформации, т.е. зададим непрерывно дифференцируемую функцию ϕ , взаимно однозначно отображающую $G \subset \mathbb{R}^2$ на Ω . Заряд области при такой ее деформации, очевидно, сохранится, а его распределение, т.е. поверхностная плотность, изменится. Обозначим новую плотность распределения заряда g .

Возьмем в G прямоугольник $\Delta = [y_1^{(0)}, y_1^{(0)} + h] \times [y_2^{(0)}, y_2^{(0)} + k]$. Образом этого прямоугольника в Ω при отображении ϕ будет некоторый "кривоугольник" P (рис.15.8).

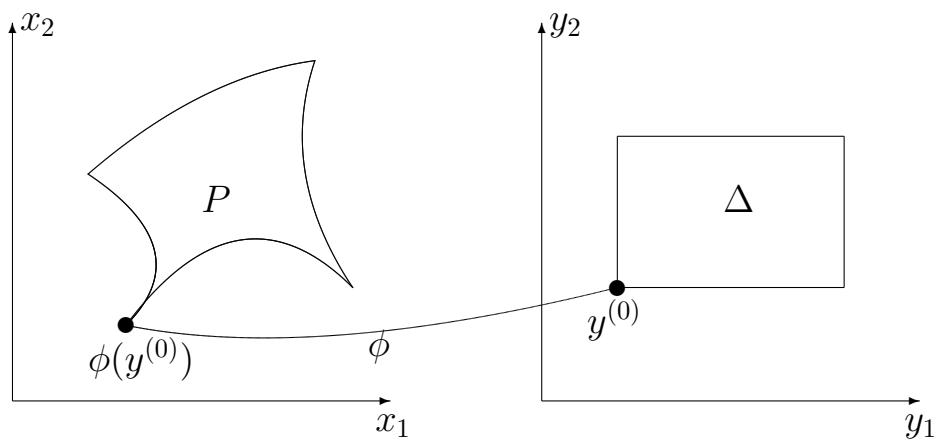


Рис.15.8

Прямоугольник Δ и "кривоугольник" P несут на себе один и тот же заряд, т.е.

$$\iint_P f = \iint_{\Delta} g.$$

Если предположить для простоты рассуждений, что обе плотности (f и g) непрерывны, то, по теореме о среднем, найдутся точки $x_{cp} \in P$ и $y_{cp} \in \Delta$, для которых

$$f(x_{cp}) \cdot S(P) = \iint_P f = \iint_{\Delta} g = g(y_{cp}) \cdot S(\Delta).$$

Отсюда

$$g(y_{cp}) = f(x_{cp}) \cdot \frac{S(P)}{S(\Delta)}.$$

Стягивая прямоугольник Δ к точке $y^{(0)}$, т.е. переходя к пределу ($h = 0, k = 0$), получим

$$g(y^{(0)}) = f(\phi(y^{(0)})) \cdot \lim_{h=k=0} \frac{S(P)}{S(\Delta)}.$$

Но по формуле (15.2.2)

$$S(P) = \iint_{\Delta} |\det(\phi')|.$$

По теореме о среднем (ϕ – непрерывно дифференцируема) найдется такая точка $c \in \Delta$, что

$$S(P) = |\det(\phi'(c))| \cdot S(\Delta).$$

При стягивании Δ к точке $y^{(0)}$ имеем

$$\lim_{h=k=0} \frac{S(P)}{S(\Delta)} = |\det(\phi'(y^{(0)}))|,$$

и, следовательно, искомая связь между плотностями распределения заряда на G и на Ω имеет вид

$$g(y) = f(\phi(y)) \cdot |\det(\phi'(y))|.$$

Действительно, можно показать, что имеет место

Теорема. Если $f: \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ – кусочно непрерывная функция, а ϕ – непрерывно дифференцируемое взаимно однозначное отображение $G \subset \mathbb{R}^2$ на Ω , то справедлива формула

$$\iint_{\Omega} f = \iint_G (f \circ \phi) |det(\phi')|.$$

Сравните эту формулу с правилом подстановки для преобразования одномерного интеграла в п.14.8.

Замечание. Условие взаимной однозначности отображения может нарушаться на границе области – теорема работает и в этом случае.

Пример. Вычислим интеграл

$$\iint_{\Omega_r} exp(-(x_1^2 + x_2^2)) dx_1 dx_2,$$

где Ω_r – круг радиуса r с центром в начале координат, являющийся образом прямоугольника G : $0 \leq \rho \leq r$, $0 \leq \varphi \leq 2\pi$ (рис.15.9) при отображении ϕ : $\begin{cases} x_1 = \rho \cdot \cos(\varphi) \\ x_2 = \rho \cdot \sin(\varphi) \end{cases}$.

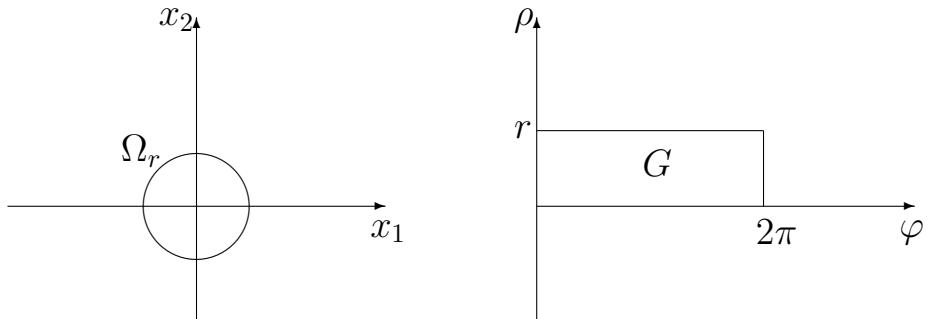


Рис.15.9

Это отображение не взаимно однозначно: правая и левая границы прямоугольника "склеиваются" а вся его нижняя граница переходит в одну точку – начало координат. Но, согласно замечанию к теореме, правило подстановки работает.

В п.11.1 было показано, что $det(\phi') = \rho$. Поэтому

$$\begin{aligned} \iint_{\Omega_r} exp(-(x_1^2 + x_2^2)) dx_1 dx_2 &= \iint_G exp(-\rho^2) \cdot \rho d\rho d\varphi = \\ &= \int_0^{2\pi} \left(\int_0^r exp(-\rho^2) \cdot \rho d\rho \right) d\varphi = \pi \cdot (1 - exp(-r^2)). \end{aligned}$$

15.7. Тройной интеграл

Конструкция тройного интеграла аналогична конструкции двойного, поэтому мы ограничимся определением и перечислением его свойств.

Пусть $\Delta = [a, b] \times [c, d] \times [q, r]$ – прямоугольный параллелепипед в \mathbb{R}^3 , и $f: \Delta \rightarrow \mathbb{R}$ – кусочно непрерывная функция. Построим какие-нибудь разбиения сегментов $[a, b]$, $[c, d]$ и $[q, r]$. Порождаемое ими разбиение параллелепипеда обозначим P .

Построим суммы Дарбу

$$L(f, P) = \sum_i \sum_j \sum_k m_{ijk} V_{ijk}, \quad U(f, P) = \sum_i \sum_j \sum_k M_{ijk} V_{ijk},$$

где

$$m_{ijk} = \inf_{\Delta_{ijk}} \{f(x, y, z)\}, \quad M_{ijk} = \sup_{\Delta_{ijk}} \{f(x, y, z)\},$$

Δ_{ijk} – элементарный параллелепипед разбиения P , V_{ijk} – его объем.

Имеет место

Теорема. Существует единственное число J , удовлетворяющее неравенству

$$L(f, P) \leq J \leq U(f, P)$$

при любом разбиении P . Его называют *тройным интегралом* от функции f по параллелепипеду Δ . Обозначают тройной интеграл так:

$$J = \iiint_{\Delta} f \quad \text{или} \quad J = \iiint_{\Delta} f(x, y, z) dx dy dz.$$

Если G – произвольное тело в \mathbb{R}^3 , ограниченное кусочно гладкой поверхностью, и f – вещественная кусочно непрерывная функция, заданная на G , то полагают *по определению*

$$\iiint_G f = \iiint_{\Delta} F,$$

где Δ – произвольный прямоугольный параллелепипед, содержащий G , а функция F совпадает с f на G и равна нулю вне G .

Объемом тела $G \subset \mathbb{R}^3$, ограниченного кусочно гладкой поверхностью, называют число

$$V(G) = \iiint_G 1.$$

Если G – часть цилиндра, поперечное сечение которого – плоская фигура Ω , а образующая параллельна оси аппликат, причем G ограничена сверху графиком кусочно гладкой функции ψ , а снизу – графиком кусочно гладкой функции φ ($\varphi \leq \psi$; см. рис.15.10), то

$$\iiint_G = \iint_{\Omega} \left(\int_{\varphi(x,y)}^{\psi(x,y)} f(x, y, z) dz \right) dx dy.$$

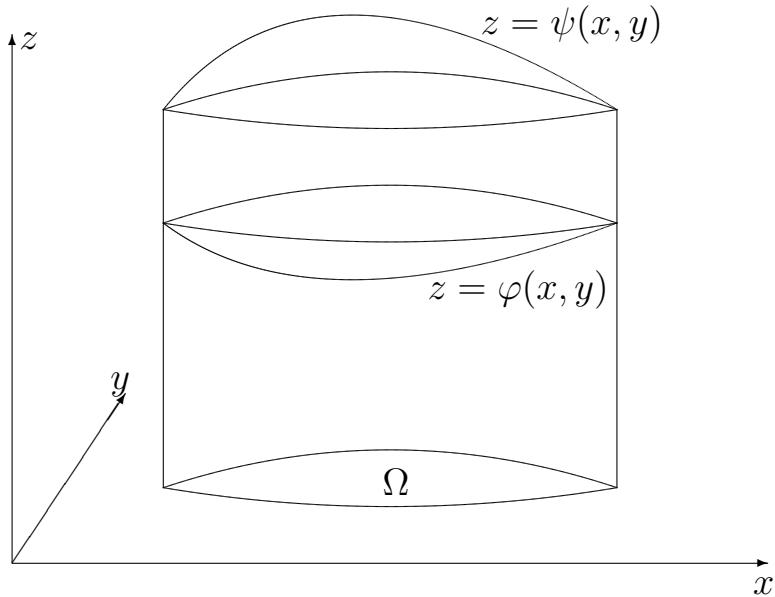


Рис.15.10

Аддитивность тройного интеграла. Если $G = G_1 \cup G_2$, причем G_1 и G_2 не имеют общих точек, кроме граничных ("не налегают" друг на друга), и имеют кусочно гладкие границы, а f кусочно непрерывна на G , то

$$\iiint_G f = \iiint_{G_1} f + \iiint_{G_2} f.$$

(Если заряженное тело разрезать на части, то заряд всего тела равен сумме зарядов его частей.)

Линейность тройного интеграла. Для любых кусочно непрерывных на G функций f_1, f_2 и любых чисел α_1, α_2

$$\iiint_G (\alpha_1 f_1 + \alpha_2 f_2) = \alpha_1 \iiint_G f_1 + \alpha_2 \iiint_G f_2.$$

Интегрирование неравенств. Если $f_1, f_2 : G \rightarrow \mathbb{R}$ – кусочно непрерывные функции, причем $f_1 \leq f_2$, то

$$\iiint_G f_1 \leq \iiint_G f_2.$$

Теорема о среднем. Если f непрерывна на G , то найдется такая точка $c \in G$, что

$$\frac{1}{V(G)} \cdot \iiint_G f = f(c).$$

Правило подстановки. Если G – тело в \mathbb{R}^3 , ограниченное кусочно гладкой поверхностью, f – кусочно непрерывная вещественная функция, заданная на G , а ϕ – непрерывно дифференцируемое взаимно однозначное отображение тела $H \subset \mathbb{R}^3$ на G , то

$$\iiint_G f = \iiint_H (f \circ \phi) \cdot |\det(\varphi')|.$$

Замечания. 1. Правило подстановки работает и в случае нарушения на границе взаимной однозначности отображения ϕ .

2. Нетрудно заметить, что аналогичным образом можно определить интегралы любой кратности – "четверной" "пятерной" и т.д. (но, конечно, уже без геометрической их интерпретации).

3. Поскольку писать большое количество "крючков" – знаков интеграла – утомительно, часто интеграл кратности n записывают в виде $\int_G f$, где G – заданная часть \mathbb{R}^n .

Глава 16. НЕСОБСТВЕННЫЕ ИНТЕГРАЛЫ

16.1. Несобственный интеграл от неограниченной функции

Начнем с примера. Рассмотрим функцию $f : [0, 1] \rightarrow \mathbb{R}$

$$f(x) = \begin{cases} \frac{1}{\sqrt{x}} & \text{при } x \neq 0; \\ A \in \mathbb{R} & \text{при } x = 0. \end{cases}$$

Она *не ограничена* на сегменте $[0, 1]$ ($\lim_{x \rightarrow 0^+} \frac{1}{\sqrt{x}} = +\infty$). Поэтому не существуют верхние суммы Дарбу, и не существует интеграл Римана от f по $[0, 1]$.

Возьмем на *интервале* $]0, 1[$ точку α и рассмотрим функцию f_α — сужение f на $[\alpha, 1]$. Она непрерывна, и существует интеграл

$$\int_{\alpha}^1 f_\alpha = \int_{\alpha}^1 \frac{1}{\sqrt{x}} dx = 2\sqrt{x} \Big|_{\alpha}^1 = 2 - 2\sqrt{\alpha},$$

предел которого при $\alpha = 0+$ конечен:

$$\lim_{\alpha \rightarrow 0^+} \int_{\alpha}^1 f_\alpha = \lim_{\alpha \rightarrow 0^+} (2 - 2\sqrt{\alpha}) = 2.$$

Этот предел называют *несобственным интегралом* от функции f по $[0, 1]$. Пишут

$$\int_0^1 f = \lim_{\alpha \rightarrow 0^+} \int_{\alpha}^1 f = 2.$$

Замечание. Так как значение функции f в нуле не играет никакой роли, можно считать, что она в нуле не задана (задана на полуинтервале $]0, 1]$) и писать

$$\int_0^1 \frac{1}{\sqrt{x}} = \lim_{\alpha \rightarrow 0^+} \int_{\alpha}^1 \frac{1}{\sqrt{x}} = 2.$$

Приведем простую физическую интерпретацию этого примера. Электрический заряд распределен вдоль стержня $[0, 1]$ с линейной плотностью $\frac{1}{\sqrt{x}}$ (значение плотности в точке $x = 0$ роли не играет). Эта

плотность не ограничена. Однако полный заряд стержня конечен и равен двум единицам заряда.

Попробуем поступить так же с функцией $g :]0, 1] \rightarrow \mathbb{R}$ $g(x) = \frac{1}{x}$.

$$\lim_{\alpha=0+} \int_{\alpha}^1 \frac{1}{x} dx = \lim_{\alpha=0+} (\ln(1) - \ln(\alpha)) = +\infty.$$

Символу $\int_0^1 \frac{1}{x} dx$ невозможно приписать никакого числового значения.

Обычно используют математические эвфемизмы и говорят "несобственный интеграл не существует" или "несобственный интеграл расходится".

Пользуясь той же физической интерпретацией, можно сказать, что такая линейная плотность распределения заряда была бы возможна только при бесконечно большом полном заряде стержня.

Определение. Если функция $f :]a, b] \rightarrow \mathbb{C}$ не ограничена в окрестности точки a , но кусочно непрерывна на сегменте $[\alpha, b]$ при любом $\alpha \in]a, b[$, и существует *конечный* предел $\lim_{\alpha=a+} \int_{\alpha}^b f$, то этот предел называют несобственным интегралом от f по $[a, b]$ и пишут

$$\int_a^b f = \lim_{\alpha=a+} \int_{\alpha}^b f.$$

Аналогично определяют несобственный интеграл от функции f , заданной на $[a, b[$, не ограниченной в окрестности точки b и кусочно непрерывной на $[a, \beta]$ при любом $\beta \in]a, b[$:

$$\int_a^b f = \lim_{\beta=b-} \int_a^{\beta} f.$$

Понятие "несобственный интеграл от неограниченной функции" можно *описать* так: если один из концов сегмента, по которому хотят проинтегрировать функцию, является для этой функции *особой точкой* (в ее окрестности функция не ограничена), то малую окрестность особой точки удаляют и вычисляют интеграл Римана по оставшейся части сегмента (где функция ведет себя прилично). Затем длину удаленной части сегмента устремляют к нулю. Если при этом интеграл Римана

имеет *конечный* предел, то этот предел называют *несобственным интегралом* (это уже НЕ интеграл Римана!). Если *конечного* предела нет, то говорят, что "несобственный интеграл не существует" (расходится).

Остается рассмотреть случай, когда особая точка лежит внутри сегмента, по которому ведется интегрирование.

Пусть функция f задана во всех точках сегмента, кроме его *внутренней* точки c , не ограничена в окрестности точки c , но сужение этой функции на $[a, \alpha] \cup [\beta, b]$, ($a < \alpha < c < \beta < b$) кусочно непрерывно. Тогда несобственный интеграл от функции f по сегменту $[a, b]$ определяют как сумму несобственных интегралов по $[a, c]$ и $[c, b]$, т.е.

$$\int_a^b f = \int_a^c f + \int_c^b f = \lim_{\alpha=c-} \int_a^\alpha f + \lim_{\beta=c+} \int_\beta^b f$$

(если оба эти предела конечны). Если хотя бы один из них не существует, то несобственный интеграл $\int_a^b f$ не существует.

Пример. Несобственный интеграл $\int_{-2}^1 \frac{1}{x} dx$ не существует, так как уже было показано, что не существует несобственный интеграл $\int_0^1 \frac{1}{x} dx$.

Если особая точка c лежит внутри сегмента $[a, b]$, и несобственный интеграл $\int_a^b f$ не существует, то можно попытаться придать смысл симметрии $\int_a^b f$, используя еще одно понятие.

Определение. *Главным значением* (по Коши) несобственного интеграла $\int_a^b f$ называется число

$$V.P.^{44} \int_a^b f = \lim_{\varepsilon=0+} \left(\int_a^{c-\varepsilon} f + \int_{c+\varepsilon}^b f \right).$$

Замечание. Обратите внимание на то, что

1) в определении главного значения удаляемый промежуток (с плохим поведением функции) симметричен относительно особой точки;

⁴⁴valeur principal (фр.) – главное значение.

2) вычисляется предел суммы вместо суммы пределов (известно, что предел суммы существует при существовании пределов слагаемых, но может существовать и при отсутствии последних).

Примеры.

$$1. \quad V.P. \int_{-2}^1 \frac{1}{x} dx = \lim_{\varepsilon \rightarrow 0+} \left(\int_{-2}^{0-\varepsilon} \frac{1}{x} dx + \int_{0+\varepsilon}^1 \frac{1}{x} dx \right) = \\ = \lim_{\varepsilon \rightarrow 0+} (\ln(\varepsilon) - \ln(2) + \ln(1) - \ln(\varepsilon)) = -\ln(2).$$

$$2. \quad V.P. \int_{-1}^1 \frac{1}{x^2} dx = \lim_{\varepsilon \rightarrow 0+} \left(\int_{-1}^{0-\varepsilon} \frac{1}{x^2} dx + \int_{0+\varepsilon}^1 \frac{1}{x^2} dx \right) = \lim_{\varepsilon \rightarrow 0+} \left(\frac{2}{\varepsilon} - 2 \right) = +\infty.$$

Показав, что не существует главное значение несобственного интеграла, мы показали также, что не существует и несобственный интеграл (в обычном смысле).

16.2. Несобственный интеграл по бесконечному промежутку

Определение. Пусть $a \in \mathbb{R}$ и $f : [a, +\infty[\rightarrow \mathbb{C}$ – функция, кусочно непрерывная на любом сегменте $[a, A]$ ($a < A < +\infty$).

Если существует конечный предел $\lim_{A \rightarrow +\infty} \int_a^A f$, то его называют несобственным интегралом от функции f по $[a, +\infty[$ и пишут

$$\int_a^{+\infty} f = \lim_{A \rightarrow +\infty} \int_a^A f.$$

Если конечный предел не существует, то говорят, что несобственный интеграл не существует (расходится).

Примеры.

$$1. \quad \int_0^{+\infty} \frac{dx}{1+x^2} = \lim_{A \rightarrow +\infty} \int_0^A \frac{dx}{1+x^2} = \lim_{A \rightarrow +\infty} (\arctg(A) - \arctg(0)) = \frac{\pi}{2}.$$

$$2. \quad \int_0^{+\infty} \frac{x dx}{1+x^2} = \lim_{A \rightarrow +\infty} \int_0^A \frac{x \cdot dx}{1+x^2} = \frac{1}{2} \cdot \lim_{A \rightarrow +\infty} (\ln(1+A^2) - \ln(1)) = +\infty$$

(несобственный интеграл расходится).

$$3. \int_0^{+\infty} \cos(x) dx = \lim_{A \rightarrow +\infty} \int_0^A \cos(x) dx = \lim_{A \rightarrow +\infty} (\sin(A) - \sin(0)).$$

Предел не существует (несобственный интеграл расходится).

Аналогично определяется несобственный интеграл по $]-\infty, b]$

$$\int_{-\infty}^b f = \lim_{B \rightarrow -\infty} \int_B^b f.$$

Несобственный интеграл по $]-\infty, +\infty[$ определяется как сумма пределов (при условии, что оба предела существуют и конечны)

$$\int_{-\infty}^{+\infty} f = \lim_{B \rightarrow -\infty} \int_B^c f + \lim_{A \rightarrow +\infty} \int_c^A f \quad (c - \text{любое вещественное число}).$$

Если хотя бы один из пределов не существует, то говорят, что несобственный интеграл расходится.

Например, несобственный интеграл $\int_{-\infty}^{+\infty} \frac{x dx}{1+x^2}$ расходится, так как уже было показано, что расходится несобственный интеграл $\int_0^{+\infty} \frac{x dx}{1+x^2}$.

В случае несобственного интеграла по $]-\infty, +\infty[$ можно ввести понятие главного значения (по Коши)

$$V.P. \int_{-\infty}^{+\infty} f = \lim_{A \rightarrow +\infty} \int_{-A}^A f.$$

Так же, как и в случае несобственного интеграла от неограниченной функции, главное значение отличается от несобственного интеграла в обычном смысле тем, что, во-первых, интегрирование ведется по симметричному относительно начала координат промежутку, и, во-вторых, вычисляется не *сумма пределов*, а *предел суммы*, который может существовать и при отсутствии пределов слагаемых.

Пример.

$$V.P. \int_{-\infty}^{+\infty} \frac{x dx}{1+x^2} = \lim_{A \rightarrow +\infty} \int_{-A}^A \frac{x dx}{1+x^2} = \frac{1}{2} \cdot \lim_{A \rightarrow +\infty} (\ln(1+A^2) - \ln(1+A^2)) = 0.$$

Рекомендуем читателю убедиться в том, что символу $\int_{-\infty}^{+\infty} x^2 dx$ нельзя приписать числовое значение ни одним из рассмотренных способов.

Замечание. Мы ввели два различных типа несобственных интегралов: интегралы от неограниченных функций и интегралы по бесконечному промежутку. В дальнейшем мы будем употреблять термин "интеграл, несобственный на правом конце промежутка $[a, b[$ " как в случае, когда $b = +\infty$, так и в случае, когда $b \in \mathbb{R}$, но подынтегральная функция не ограничена в окрестности точки b . Аналогично определим и термин "интеграл, несобственный на левом конце промежутка $]a, b]$ ".

16.3. Признаки сходимости несобственных интегралов

Можно показать, что для несобственных интегралов справедливы теоремы, аналогичные теоремам о сходимости числовых рядов.

Теорема 1 (признак сравнения для интегралов от положительных функций). Пусть $f \geq g \geq 0$ на $[a, b[,$ и интегралы $\int_a^b f,$ $\int_a^b g$ – несобственные на правом конце промежутка. Если сходится $\int_a^b f,$ то сходится и $\int_a^b g,$ причем $\int_a^b g \leq \int_a^b f,$ а если расходится $\int_a^b g,$ то расходится и $\int_a^b f.$

Теорема 2 (предельная форма признака сравнения). Пусть $f \geq 0,$ $g \geq 0$ на $[a, b[,$ и интегралы $\int_a^b f,$ $\int_a^b g$ – несобственные на правом конце промежутка. Если существует $\lim_{x=b-} \frac{g(x)}{f(x)} = L \in \mathbb{R},$ то:

- 1) при $L \neq 0$ либо оба интеграла сходятся, либо оба расходятся;
- 2) при $L = 0$ если сходится $\int_a^b f,$ то сходится и $\int_a^b g;$ если расходится $\int_a^b g,$ то расходится и $\int_a^b f.$

Теорема 3. Пусть $f : [a, b[\rightarrow \mathbb{C},$ и интеграл $\int_a^b f$ – несобственный на правом конце промежутка. Если сходится $\int_a^b |f|,$ то сходится и $\int_a^b f,$ причем $\left| \int_a^b f \right| \leq \int_a^b |f|$ (в этом случае говорят, что интеграл $\int_a^b f$ абсолютно сходится).

Эти теоремы естественным образом переносятся на случай интегралов, несобственных на *левом* конце промежутка.

Пример. Рассмотрим семейство несобственных интегралов с параметром x :

$$\int_0^{+\infty} t^{x-1} \exp(-t) dt.$$

Докажем, что при $x > 0$ эти несобственные интегралы существуют.

Отметим, что при $x < 1$ эти интегралы – несобственные на обоих концах промежутка: на правом промежуток бесконечный, а на левом – подынтегральная функция не ограничена в окрестности нуля. Поэтому рассмотрим отдельно два интеграла:

$$H_1 = \int_0^1 t^{x-1} \exp(-t) dt, \quad H_2 = \int_1^{+\infty} t^{x-1} \exp(-t) dt.$$

Ввиду очевидного неравенства $0 \leq t^{x-1} \exp(-t) \leq t^{x-1}$ имеем

$$\int_0^1 t^{x-1} dt = \lim_{\alpha \rightarrow 0+} \int_\alpha^1 t^{x-1} dt = \lim_{\alpha \rightarrow 0+} \frac{t^x}{x} \Big|_\alpha^1 = \lim_{\alpha \rightarrow 0+} \frac{1 - \alpha^x}{x} = \frac{1}{x}.$$

По признаку сравнения интеграл H_1 существует.

Для исследования H_2 используем свойство экспоненты (см. п.9.1)

$$\lim_{t \rightarrow +\infty} (t^{x+1} \exp(-t)) = 0, \quad \text{т.е.} \quad \lim_{t \rightarrow +\infty} \frac{t^{x-1} \exp(-t)}{t^{-2}} = 0.$$

Так как

$$\int_1^{+\infty} t^{-2} dt = \lim_{A \rightarrow +\infty} \int_1^A t^{-2} dt = \lim_{A \rightarrow +\infty} \left(-\frac{1}{t} \right) \Big|_1^A = 1,$$

по признаку сравнения в предельной форме интеграл H_2 существует.

Мы показали, что на $]0, +\infty[$ определена функция

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} \exp(-t) dt,$$

именуемая *Гамма-функцией*.

Имеет место рекуррентное соотношение

$$\Gamma(x+1) = x \cdot \Gamma(x), \quad x > 0.$$

Доказательство. Интегрируя "по частям получаем

$$\begin{aligned} \Gamma(x+1) &= \int_0^{+\infty} t^x \exp(-t) dt = -t^x \exp(-t) \Big|_0^{+\infty} + x \cdot \int_0^{+\infty} t^{x-1} \exp(-t) dt = \\ &= x \cdot \int_0^{+\infty} t^{x-1} \exp(-t) dt = x \cdot \Gamma(x) \end{aligned}$$

(здесь вновь использовано соотношение $\lim_{t \rightarrow +\infty} (t^x \exp(-t)) = 0$). ■

В частности,

$$\Gamma(1) = \int_0^{+\infty} \exp(-t) dt = -\exp(-t) \Big|_0^{+\infty} = 1.$$

Отсюда $\Gamma(2) = \Gamma(1+1) = 1 \cdot \Gamma(1) = 1$, $\Gamma(3) = \Gamma(2+1) = 2 \cdot \Gamma(2) = 2$, $\Gamma(4) = \Gamma(3+1) = 3 \cdot \Gamma(3) = 6$ и вообще

$$\Gamma(n) = (n-1)!, \quad n \in \mathbb{N}.$$

Вычислим еще $\Gamma(1/2)$. Подстановка $t = x^2$ дает

$$\Gamma(1/2) = \int_0^{+\infty} t^{-1/2} \exp(-t) dt = \int_0^{+\infty} x^{-1} \exp(-x^2) \cdot 2x dx = 2 \int_0^{+\infty} \exp(-x^2) dx.$$

Очевидно, что $\int_0^{+\infty} \exp(-x^2) dx = \int_{-\infty}^0 \exp(-x^2) dx$ (один из этих интегралов сводится к другому подстановкой $x = -t$).

Поэтому

$$\Gamma(1/2) = \int_{-\infty}^{+\infty} \exp(-x^2) dx.$$

Этот интеграл называется интегралом Пуассона⁴⁵. Вычислить его можно, например, так: рассмотрим интеграл

⁴⁵Симеон Дени ПУАССОН (S.D. Poisson, 1781-1840) – французский механик, физик и математик, член Парижской АН. Его работы сыграли важную роль в становлении теории вероятностей и математической физики.

$$H(A) = \int_{-A}^A \exp(-x^2) dx.$$

Имеем

$$H^2(A) = \left(\int_{-A}^A \exp(-x^2) dx \right) \cdot \left(\int_{-A}^A \exp(-y^2) dy \right)$$

(буква, обозначающая "переменную интегрирования несущественна!)

В силу линейности интеграла

$$\begin{aligned} H^2(A) &= \int_{-A}^A \left(\exp(-y^2) \cdot \int_{-A}^A \exp(-x^2) dx \right) dy = \\ &= \int_{-A}^A \left(\int_{-A}^A \exp(-x^2) \cdot \exp(-y^2) dx \right) dy. \end{aligned}$$

Преобразуем этот повторный интеграл в двойной

$$H^2(A) = \iint_{\Delta_A} \exp(-x^2) \cdot \exp(-y^2) dx dy = \iint_{\Delta_A} \exp(-(x^2 + y^2)) dx dy,$$

где Δ_A – квадрат $[-A, A] \times [-A, A]$.

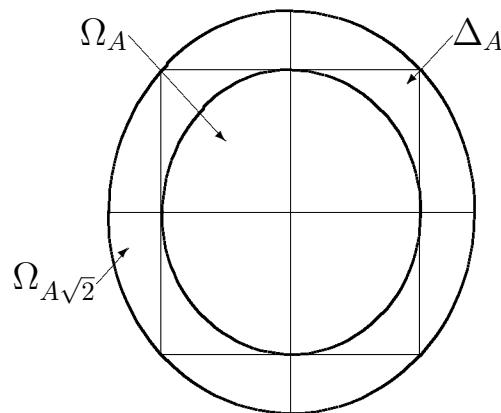


Рис.16.1

Очевидно, что (см. рис.16.1) $\Omega_A \subset \Delta_A \subset \Omega_{A\sqrt{2}}$, где Ω_A – круг радиуса A , $\Omega_{A\sqrt{2}}$ – круг радиуса $A\sqrt{2}$ (оба с центрами в начале координат). В силу положительности подынтегральной функции имеет место неравенство

$$\iint_{\Omega_A} \exp(-(x^2 + y^2)) \, dx dy \leq H^2(A) \leq \iint_{\Omega_{A\sqrt{2}}} \exp(-(x^2 + y^2)) \, dx dy.$$

Воспользовавшись результатом примера из п.15.6, получаем

$$\pi \cdot (1 - \exp(-A^2)) \leq H^2(A) \leq \pi \cdot (1 - \exp(-2A^2)).$$

Переходя к пределу, получим, наконец, что $\lim_{A \rightarrow +\infty} H^2(A) = \pi$, и

$$\Gamma(1/2) = \sqrt{\pi}.$$

Замечание. Отсюда следует, между прочим, что

$$\lim_{x \rightarrow +\infty} \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^{+\infty} \exp(-t^2) \, dt = 1.$$

16.4. Преобразование Лапласа

Определение. Будем называть *оригиналом* функцию $f : \mathbb{R} \rightarrow \mathbb{C}$, если

- 1) $f(t) = 0$ при $t < 0$;
- 2) f кусочно непрерывна;
- 3) f экспоненциально ограничена, т.е. существуют такие положительные числа a и M , что при $t \geq 0$ $|f(t)| \leq M \cdot \exp(at)$.

Примеры. 1. $\delta_1(t) = \begin{cases} 0 & \text{при } t < 0 \\ 1 & \text{при } t \geq 0 \end{cases}$. Этую функцию называют *функцией Хевисайда*⁴⁶ (*единичным скачком, единичной ступенькой*).

Функция Хевисайда – оригинал, так как она кусочно непрерывна (имеет в нуле единственную точку разрыва первого рода), ограничена и, следовательно, экспоненциально ограничена (можно положить, например, $M = 1$, $a = 0$).

$$2. f(t) = \exp(zt) \cdot \delta_1(t), \quad z \in \mathbb{C}.$$

f кусочно непрерывна. Докажем ее экспоненциальную ограниченность. Пусть $x = \operatorname{Re}(z)$. Тогда $|f(t)| = \exp(xt)$, т.е. можно положить $M = 1$, $a = x$. Итак, f – оригинал.

⁴⁶Оливер ХЕВИСАЙД (O. Heaviside, 1850-1925) – английский физик и инженер, член Лондонского Королевского общества. Предсказал открытие так называемого слоя Хевисайда в атмосфере. Один из создателей операционного исчисления.

3. $g(t) = \exp(t^2) \cdot \delta_1(t)$.

Пусть M и a – произвольные положительные числа. Решим (относительно переменной t) неравенство $\exp(t^2) \leq M \cdot \exp(at)$. Сокращая на $\exp(at) > 0$, получим

$$\exp(t^2 - at) \leq M, \quad \text{или} \quad t^2 - at \leq \ln(M).$$

Последнее неравенство не может выполняться при *всех положительных* t , так как $\lim_{t \rightarrow +\infty} (t^2 - at) = +\infty$. Следовательно, g – не оригинал.

Установим некоторые свойства оригиналов.

1. Очевидно, что линейная комбинация оригиналов есть оригинал.

Пример. Оригиналами являются функции

$$\cos(\omega t) \cdot \delta_1(t) = \frac{1}{2} \cdot \exp(i\omega t) \cdot \delta_1(t) + \frac{1}{2} \cdot \exp(-i\omega t) \cdot \delta_1(t);$$

$$\sin(\omega t) \cdot \delta_1(t) = \frac{1}{2i} \cdot \exp(i\omega t) \cdot \delta_1(t) - \frac{1}{2i} \cdot \exp(-i\omega t) \cdot \delta_1(t).$$

2. Если f – оригинал, то ее первообразная $F(t) = \int_0^t f$ – тоже оригинал. Действительно, при $t < 0$ $f(t) \equiv 0$ и $\int_0^t f = 0$, а при $t > 0$ из неравенства $|f(t)| \leq M \cdot \exp(at)$ следует

$$|F(t)| = \left| \int_0^t f \right| \leq \int_0^t |f| \leq \sup_{[0,t]} (|f|) \cdot t \leq M \cdot \exp(at) \cdot t \leq M \cdot \exp((a+1)t).$$

Таким образом, первообразная экспоненциально ограничена и потому является оригиналом.

Пример. Оригиналами являются функции $t^n \cdot \delta_1(t)$ ($n \in \mathbb{N}$), так как

$$t \cdot \delta_1(t) = \int_0^t \delta_1(t) dt; \quad t^2 \cdot \delta_1(t) = 2 \int_0^t t \delta_1(t) dt; \quad \dots, \quad t^n \cdot \delta_1(t) = n \int_0^t t^{n-1} \cdot \delta_1(t) dt.$$

Замечание. Даже если у оригинала есть производная, она не обязана быть оригиналом.

Определение. Пусть $f_1, f_2 : \mathbb{R} \rightarrow \mathbb{C}$. Тогда несобственный интеграл с вещественным параметром x

$$\int_{-\infty}^{+\infty} f_1(t) \cdot f_2(x-t) dt$$

определяет на той части \mathbb{R} , где он сходится, функцию, называемую *сверткой* функций f_1 и f_2 , и обозначаемую $f_1 \otimes f_2$.

Отметим следующие свойства свертки:

1. $f_1 \otimes f_2 = f_2 \otimes f_1$.
2. $f_1 \otimes (f_2 + f_3) = f_1 \otimes f_2 + f_1 \otimes f_3$.
3. Свертка двух оригиналов – оригинал.

Доказательство. 1. Подстановка $\tau = x - t$ дает

$$\int_{-\infty}^{+\infty} f(t) \cdot g(x-t) dt = \int_{-\infty}^{+\infty} f(x-\tau) \cdot g(\tau) d\tau.$$

2. Следует из линейности интеграла.
3. Пусть f и g – оригиналы. Тогда

$$\begin{aligned} (f \otimes g)(x) &= \int_{-\infty}^{+\infty} f(t) \cdot g(x-t) dt = \\ &= \int_{-\infty}^0 f(t) \cdot g(x-t) dt + \int_0^x f(t) \cdot g(x-t) dt + \int_x^{+\infty} f(t) \cdot g(x-t) dt. \end{aligned}$$

В первом интеграле $t < 0$, следовательно, *оригинал* f – тождественный нуль, в третьем интеграле $x - t < 0$, следовательно, *оригинал* g – тождественный нуль. Итак,

$$(f \otimes g)(x) = \int_0^x f(t) \cdot g(x-t) dt. \quad (16.4.1)$$

Отсюда видно, что свертка оригиналов определена при всех $x \in \mathbb{R}$. Далее, если $x < 0$, то $(f \otimes g)(x) = \int_0^x f(t) \cdot g(x-t) dt = 0$, так как $f(t) \equiv 0$.

На положительной полуоси имеют место оценки

$$|f(t)| \leq M \cdot \exp(at) \leq M \cdot \exp(ct), \quad |g(t)| \leq N \cdot \exp(bt) \leq N \cdot \exp(ct)$$

(здесь $c = \max\{a, b\}$). Поэтому

$$\begin{aligned}
\left| \int_0^x f(t) \cdot g(x-t) dt \right| &\leq \int_0^x |f(t)| \cdot |g(x-t)| dt \leq \\
&\leq MN \cdot \int_0^x \exp(ct) \cdot \exp(c(x-t)) dt = MN \cdot \int_0^x \exp(cx) dt = \\
&= MN \cdot x \cdot \exp(cx) < MN \cdot \exp((c+1)x).
\end{aligned}$$

Таким образом, свертка экспоненциально ограничена и, следовательно, является оригиналом. ■

Рассмотрев свойства функций-оригиналов, перейдем к определению преобразования Лапласа⁴⁷.

Пусть f – оригинал. Рассмотрим семейство несобственных интегралов с *комплексным* параметром $s = \sigma + i\omega$ ($\sigma, \omega \in \mathbb{R}$)

$$\int_0^{+\infty} f(t) \cdot \exp(-st) dt.$$

Оценим подынтегральную функцию. Поскольку f экспоненциально ограничена, существуют такие числа M и a , что $|f(t)| \leq M \cdot \exp(at)$. С учетом равенства $|\exp(-st)| = \exp(-\sigma t)$ получаем

$$|f(t) \cdot \exp(st)| \leq M \cdot \exp(-(\sigma - a)t). \quad (16.4.2)$$

При $\sigma > a$

$$\begin{aligned}
\int_0^{+\infty} \exp(-(\sigma - a)t) dt &= \lim_{A \rightarrow +\infty} \int_0^A \exp(-(\sigma - a)t) dt = \\
&= \lim_{A \rightarrow +\infty} \frac{\exp(-(\sigma - a)t)}{\sigma - a} \Big|_0^A = \lim_{A \rightarrow +\infty} \frac{1 - \exp(-(\sigma - a)A)}{\sigma - a} = \frac{1}{\sigma - a}.
\end{aligned}$$

Признак сравнения и неравенство (16.4.2) доказывают сходимость несобственного интеграла $\int_0^{+\infty} |f(t) \cdot \exp(st)| dt$, т.е. абсолютную сходимость несобственного интеграла $\int_0^{+\infty} f(t) \cdot \exp(st) dt$.

⁴⁷Пьер Симон ЛАПЛАС (P.S. Laplace, 1749-1827) – французский математик, физик и астроном, член многих академий и научных обществ, автор фундаментальных работ по теории вероятностей, математической физике, небесной механике.

Мы показали, что на части комплексной плоскости, лежащей правее прямой $Re(s) > a$, определена функция

$$\tilde{f}(s) = \int_0^{+\infty} f(t) \cdot \exp(-st) dt.$$

Эту функцию называют *изображением* оригинала f . Мы будем записывать соответствие между оригиналом и его изображением так:

$$\tilde{f} = \mathcal{L}(f).$$

Оператор \mathcal{L} , который каждому оригиналу (функции) сопоставляет его изображение (функцию), называют *преобразованием Лапласа*.

Терминологическое замечание. Мы условились считать слова *отображение, функция, оператор* синонимами. Теперь к этому списку добавляется еще один синоним – *преобразование*.

Подумайте, как будет звучать фраза, если из всех синонимов оставить только один (например, "функция"): "функция сопоставляет каждой функции-оригиналу функцию-изображение". Наличие синонимов делает речь более понятной.

Пример. Найдем изображение функции Хевисайда.

$$\begin{aligned} \tilde{\delta}_1(s) &= \int_0^{+\infty} 1 \cdot \exp(-st) dt = \lim_{A \rightarrow +\infty} \int_0^{+\infty} \exp(-st) dt = \\ &= \lim_{A \rightarrow +\infty} \frac{\exp(-(s+i\omega)t)}{-s} \Big|_0^A = \lim_{A \rightarrow +\infty} \frac{1 - \exp(-sA) \cdot \exp(-i\omega A)}{s} = \frac{1}{s}. \end{aligned}$$

Итак, $\mathcal{L}(\delta_1) = \frac{1}{s}$, $(Re(s) > 0)$.

Докажем некоторые свойства преобразования Лапласа.

1. Линейность преобразования Лапласа: для любых оригиналов f_1 и f_2 и любых комплексных чисел α_1 и α_2

$$\mathcal{L}(\alpha_1 f_1 + \alpha_2 f_2) = \alpha_1 \cdot \mathcal{L}(f_1) + \alpha_2 \cdot \mathcal{L}(f_2).$$

Доказательство. Следует из линейности интеграла.

2. Теорема смещения. Если $\mathcal{L}(f) = \tilde{f}$ и $\alpha \in \mathbb{C}$, то

$$\mathcal{L}(f(t) \cdot \exp(\alpha t)) = \tilde{f}(s - \alpha).$$

Доказательство.

$$\int_0^{+\infty} (f(t) \exp(\alpha t)) \cdot \exp(-st) dt = \int_0^{+\infty} f(t) \cdot \exp(-(s-\alpha)t) dt = \tilde{f}(s-\alpha). \blacksquare$$

3. Теорема запаздывания. Если $\mathcal{L}(f) = \tilde{f}$ и $\tau > 0$, то

$$\mathcal{L}(f(t - \tau)) = \exp(-s\tau) \cdot \tilde{f}(s).$$

Доказательство. Подстановка $x = t - \tau$ дает

$$\int_0^{+\infty} f(t - \tau) \cdot \exp(-st) dt = \int_{-\tau}^{+\infty} f(x) \cdot \exp(-s(\tau + x)) dx.$$

Поскольку f – оригинал, $f(x) = 0$ при $x < 0$, и интеграл равен

$$\begin{aligned} & \int_0^{+\infty} f(x) \cdot \exp(-s\tau) \cdot \exp(-sx) dx = \\ &= \exp(-s\tau) \cdot \int_0^{+\infty} f(x) \cdot \exp(-sx) dx = \exp(-s\tau) \cdot \tilde{f}(s). \blacksquare \end{aligned}$$

Название "теорема запаздывания" объясняется тем, что график функции $f_\tau(t) = f(t - \tau)$ получается сдвигом графика функции $f(t)$ вправо на величину τ ("новая функция начинает работать на τ единиц времени позже, чем старая").

4. Изображение первообразной. Если f – оригинал, то

$$\mathcal{L}\left(\int_0^t f\right) = \frac{1}{s} \cdot \mathcal{L}(f).$$

Доказательство. Пусть $g(t) = \int_0^t f$. Тогда (по свойству 2 оригиналов) g – тоже оригинал. Кроме того, $g(0) = 0$. Интегрируя "по частям" имеем

$$\begin{aligned} \tilde{f}(s) &= \int_0^{+\infty} f(t) \cdot \exp(-st) dt = g(t) \cdot \exp(-st) \Big|_0^{+\infty} + \\ &+ s \cdot \int_0^{+\infty} g(t) \cdot \exp(-st) dt = \lim_{t \rightarrow +\infty} (g(t) \cdot \exp(-st)) + s \cdot \tilde{g}(s). \end{aligned}$$

Из оценки $|g(t)| \leq M \cdot \exp(at)$ следует (см. (16.4.2)), что при $\sigma = \operatorname{Re}(s) > a$ предел равен нулю. Поэтому $\tilde{f}(s) = s \cdot \tilde{g}(s)$. ■

5. Изображение производной. Пусть функция f имеет на \mathbb{R} непрерывную производную, причем $f \cdot \delta_1$ и $f' \cdot \delta_1$ – оригиналы. Тогда

$$\mathcal{L}(f' \cdot \delta_1) = s \cdot \mathcal{L}(f \cdot \delta_1) - f(0).$$

Доказательство. По формуле Ньютона-Лейбница

$$\int_0^t (f' \cdot \delta_1) = (f(t) - f(0)) \cdot \delta_1(t) = (f \cdot \delta_1)(t) - f(0) \cdot \delta_1(t).$$

По теореме об изображении первообразной

$$\mathcal{L}(f \cdot \delta_1 - f(0) \cdot \delta_1) = \frac{1}{s} \cdot \mathcal{L}(f' \cdot \delta_1).$$

Учитывая, что $\mathcal{L}(\delta_1) = \frac{1}{s}$, получаем

$$\mathcal{L}(f' \cdot \delta_1) = s \cdot \mathcal{L}(f \cdot \delta_1) - f(0) \cdot s \cdot \mathcal{L}(\delta_1) = s \cdot \mathcal{L}(f \cdot \delta_1) - f(0). \blacksquare$$

Следствие. Если $f^{(n)} \cdot \delta_1$ – оригинал, то

$$\begin{aligned} \mathcal{L}(f^{(n)} \delta_1) &= s \cdot \mathcal{L}(f^{(n-1)} \cdot \delta_1) - f^{(n-1)}(0) = \\ &= s \cdot \left(s \cdot \mathcal{L}(f^{(n-2)} \cdot \delta_1) - f^{(n-2)}(0) \right) - f^{(n-1)}(0) = \dots \\ &\dots = s^n \cdot \mathcal{L}(f \cdot \delta_1) - s^{n-1} f(0) - \dots - f^{(n-1)}(0). \end{aligned}$$

6. Изображение свертки. Если f_1 и f_2 – оригиналы, то

$$\mathcal{L}(f_1 \otimes f_2) = \mathcal{L}(f_1) \cdot \mathcal{L}(f_2).$$

Доказательство. Согласно формуле (16.4.1) для свертки оригиналлов,

$$g(t) = (f_1 \otimes f_2)(t) = \int_0^t f_1(x) \cdot f_2(t-x) dx.$$

Отсюда

$$\begin{aligned} \tilde{g}(s) &= \int_0^{+\infty} g(t) \cdot \exp(-st) dt = \int_0^{+\infty} \left(\int_0^t f_1(x) \cdot f_2(t-x) dx \right) \cdot \exp(-st) dt = \\ &= \lim_{A \rightarrow \infty} \int_0^A \left(\int_0^t f_1(x) \cdot f_2(t-x) \cdot \exp(-st) dx \right) dt. \end{aligned}$$

Преобразуем повторный интеграл в двойной:

$$\tilde{g}(s) = \lim_{A \rightarrow \infty} \iint_{\Delta} f_1(x) \cdot f_2(t-x) \cdot \exp(-st) dx dt,$$

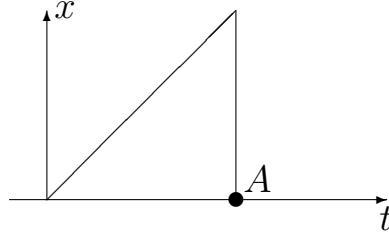


Рис.16.2

где Δ – треугольник ($0 \leq x \leq t \leq A$) (рис.16.2). Двойной интеграл снова преобразуем в повторный, но с другим порядком интегрирования:

$$\tilde{g}(s) = \lim_{A \rightarrow \infty} \int_0^A \left(\int_x^A f_1(x) \cdot f_2(t-x) \cdot \exp(-st) dx \right) dt.$$

Во внутреннем интеграле сделаем подстановку $t = x + y$ (x здесь фиксирован!)

$$\begin{aligned} \tilde{g}(s) &= \lim_{A \rightarrow \infty} \int_0^A \left(\int_0^{A-x} f_1(x) \cdot f_2(y) \cdot \exp(-s(x+y)) dy \right) dx = \\ &= \int_0^{+\infty} \left(\int_0^{+\infty} f_1(x) \cdot f_2(y) \cdot \exp(-s(x+y)) dy \right) dx. \end{aligned}$$

Наконец, в силу линейности интеграла

$$\begin{aligned} \tilde{g}(s) &= \int_0^{+\infty} \left(f_1(x) \cdot \exp(-sx) \int_0^{+\infty} f_2(y) \cdot \exp(-sy) dy \right) dx = \\ &= \left(\int_0^{+\infty} f_1(x) \cdot \exp(-sx) dx \right) \cdot \left(\int_0^{+\infty} f_2(y) \cdot \exp(-sy) dy \right) = \tilde{f}_1(s) \cdot \tilde{f}_2(s). \end{aligned}$$

Примеры. 1. Мы показали, что $\mathcal{L}(\delta_1) = \frac{1}{s}$. Отсюда по теореме смешения получаем

$$\mathcal{L}(\exp(\alpha t) \cdot \delta_1(t)) = \frac{1}{s - \alpha}.$$

2. Используя линейность преобразования Лапласа, получим далее

$$\begin{aligned}\mathcal{L}(\cos(\omega t)\delta_1(t)) &= \mathcal{L}\left(\frac{1}{2}(exp(i\omega t) \cdot \delta_1(t) + exp(-i\omega t) \cdot \delta_1(t))\right) = \\ &= \frac{1}{2}\left(\frac{1}{s-i\omega} + \frac{1}{s+i\omega}\right) = \frac{s}{s^2+\omega^2}.\end{aligned}$$

$$\begin{aligned}\mathcal{L}(\sin(\omega t)\delta_1(t)) &= \mathcal{L}\left(\frac{1}{2i}(exp(i\omega t) \cdot \delta_1(t) - exp(-i\omega t) \cdot \delta_1(t))\right) = \\ &= \frac{1}{2i}\left(\frac{1}{s-i\omega} - \frac{1}{s+i\omega}\right) = \frac{\omega}{s^2+\omega^2}.\end{aligned}$$

3. По теореме об изображении первообразной

$$\mathcal{L}(t \cdot \delta_1(t)) = \frac{1}{s} \cdot \mathcal{L}(\delta_1(t)) = \frac{1}{s^2};$$

$$\mathcal{L}\left(\frac{t^2}{2} \cdot \delta_1(t)\right) = \frac{1}{s} \cdot \mathcal{L}(t \cdot \delta_1(t)) = \frac{1}{s^3};$$

.....

$$\mathcal{L}\left(\frac{t^n}{n!} \cdot \delta_1(t)\right) = \frac{1}{s} \cdot \mathcal{L}\left(\frac{t^{n-1}}{(n-1)!} \cdot \delta_1(t)\right) = \frac{1}{s^{n+1}}.$$

4. Применяя к последнему равенству теорему смещения, получим очень важную для приложений формулу:

$$\mathcal{L}\left(\frac{t^n}{n!} \cdot exp(\alpha t) \cdot \delta_1(t)\right) = \frac{1}{(s-\alpha)^{n+1}}$$

при $n = 0, 1, 2, \dots$ и любом $\alpha \in \mathbb{C}$.

Раздел 2

ЛИНЕЙНАЯ АЛГЕБРА И ЕЕ ПРИЛОЖЕНИЯ

Глава 1. СИСТЕМЫ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

1.1. Метод полного исключения

Система m линейных алгебраических уравнений с n переменными⁴⁸ имеет вид

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2 \\ \dots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n = b_m \end{cases}. \quad (1.1.1)$$

Здесь буквой a с двумя индексами обозначены заданные числа – *коэффициенты системы* (первый индекс – номер уравнения, второй – номер переменной), буквой b с одним индексом обозначены заданные числа – *свободные члены* уравнений. Буквой x с одним индексом обозначены числовые переменные.

Решением системы (1.1.1) называется упорядоченный набор из n чисел $\tilde{x}_1, \dots, \tilde{x}_n$, подстановка которых вместо переменных (\tilde{x}_1 вместо x_1, \dots, \tilde{x}_n вместо x_n) превращает все уравнения системы в верные числовые равенства.

Систему (1.1.1) можно записать короче, если ввести удобные обозначения.

Определение. Прямоугольную числовую таблицу из m строк и n столбцов будем называть *матрицей размера $m \times n$* или *($m \times n$)-матрицей*.

Примеры.

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \text{ – матрица размера } 2 \times 3;$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \text{ – матрица размера } 3 \times 2.$$

Матрица размера $m \times m$ называется *квадратной*, а число m – ее *порядком*. Элементы квадратной матрицы, у которых совпадают

⁴⁸Напомним, что под *переменной* понимается буква, вместо которой разрешено подставлять элементы некоторого заданного множества (вместо *числовой переменной* можно подставлять любые числа).

номер строки и номер столбца, называются *диагональными*; остальные – *внедиагональными*.

Матрицу размера $n \times 1$ будем называть *матрицей-столбцом* (или просто *столбцом*), а число n – *высотой* столбца.

Матрицу размера $1 \times m$ будем называть *матрицей-строкой* (или просто *строкой*), а число m – *шириной* строки.

Квадратную матрицу 1-го порядка (одноэлементную) мы будем отождествлять с ее элементом – числом.

Примеры.

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} \text{ – столбец высоты 2; } \quad [1 \ 2 \ 3] \text{ – строка ширины 3;}$$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \text{ – квадратная матрица порядка 2.}$$

Используя введенные термины, зададим систему (1.1.1) матрицей ее коэффициентов

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{m1} & \dots & a_{mn} \end{bmatrix}$$

и столбцом свободных членов

$$b = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}.$$

Иногда матрицу коэффициентов и столбец свободных членов объединяют в *расширенную матрицу* системы

$$\left[\begin{array}{ccc|c} a_{11} & \dots & a_{1n} & b_1 \\ \dots & \dots & \dots & \dots \\ a_{m1} & \dots & a_{mn} & b_m \end{array} \right].$$

Пример. Система

$$\begin{cases} x_2 + 3x_3 - x_4 = 5 \\ x_1 + 2x_2 + x_4 = 0 \\ 3x_1 + x_2 - x_3 + 2x_4 = -1 \end{cases}.$$

имеет матрицу коэффициентов

$$A = \begin{bmatrix} 0 & 1 & 3 & -1 \\ 1 & 2 & 0 & 1 \\ 3 & 1 & -1 & 2 \end{bmatrix}$$

и столбец свободных членов

$$b = \begin{bmatrix} 5 \\ 0 \\ -1 \end{bmatrix}.$$

Можно объединить их в расширенную матрицу системы

$$\left[\begin{array}{cccc|c} 0 & 1 & 3 & -1 & 5 \\ 1 & 2 & 0 & 1 & 0 \\ 3 & 1 & -1 & 2 & -1 \end{array} \right].$$

При рассмотрении системы (1.1.1) возникают два вопроса:

- 1) существуют ли решения у этой системы?
- 2) если решения существуют, то сколько их и как их найти?

Прежде чем отвечать на эти вопросы, рассмотрим два частных случая.

1. Число уравнений системы равно числу переменных, а квадратная матрица коэффициентов имеет специальный вид (такую матрицу называют *единичной*):

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix}. \quad (1.1.2)$$

Запишем эту систему в "школьной" форме

$$\begin{cases} 1 \cdot x_1 + 0 \cdot x_2 + \dots + 0 \cdot x_n = b_1 \\ 0 \cdot x_1 + 1 \cdot x_2 + \dots + 0 \cdot x_n = b_2 \\ \dots \\ \dots \\ 0 \cdot x_1 + 0 \cdot x_2 + \dots + 1 \cdot x_n = b_n \end{cases}$$

Очевидно, что система имеет решение, оно единственное, и, более того, уже фактически найдено:

$$x_1 = b_1, x_2 = b_2, \dots, x_n = b_n.$$

2. Число уравнений системы (m) меньше числа переменных (n), и $(m \times n)$ -матрица коэффициентов имеет вид

$$\begin{bmatrix} 1 & 0 & \dots & 0 & a_{1,m+1} & \dots & a_{1n} \\ 0 & 1 & \dots & 0 & a_{2,m+1} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & a_{m,m+1} & \dots & a_{mn} \end{bmatrix}. \quad (1.1.3)$$

Запишем эту систему в "школьной" форме

$$\left\{ \begin{array}{l} 1 \cdot x_1 + a_{1,m+1}x_{m+1} + \dots + a_{1n}x_n = b_1 \\ 1 \cdot x_2 + a_{2,m+1}x_{m+1} + \dots + a_{2n}x_n = b_2 \\ \dots \\ \dots \\ 1 \cdot x_m + a_{m,m+1}x_{m+1} + \dots + a_{mn}x_n = b_m \end{array} \right.,$$

а затем перепишем ее в виде

$$\left\{ \begin{array}{l} x_1 = b_1 - a_{1,m+1}x_{m+1} - \dots - a_{1n}x_n \\ x_2 = b_2 - a_{2,m+1}x_{m+1} - \dots - a_{2n}x_n \\ \dots \\ \dots \\ x_m = b_m - a_{m,m+1}x_{m+1} - \dots - a_{mn}x_n \end{array} \right. \quad (1.1.4)$$

Задав *произвольно* значения переменных x_{m+1}, \dots, x_n и вычислив x_1, \dots, x_m по формулам (1.1.4), мы получим, очевидно, решение системы.

Пусть теперь произвольная система вида (1.1.1) задана своей расширенной матрицей. Попытаемся привести матрицу ее коэффициентов к одному из видов (1.1.2) или (1.1.3) с помощью так называемых *элементарных* преобразований, которые не изменяют множество решений системы.

Под *элементарными* преобразованиями системы линейных алгебраических уравнений мы понимаем:

- 1) изменение порядка уравнений в системе (перестановка строк ее расширенной матрицы);
- 2) изменение порядка расположения переменных в уравнениях (перестановка столбцов в матрице коэффициентов);
- 3) умножение обеих частей уравнения (строки расширенной матрицы) на отличное от нуля число;

4) прибавление к одному из уравнений системы другого, умноженного предварительно на некоторое число (прибавление к одной строке расширенной матрицы системы другой ее строки, умноженной на число).

Замечания. 1. Нетрудно убедиться в том, что перечисленные элементарные преобразования системы действительно не изменяют множество ее решений.

2. Порядок расположения переменных в уравнении важен, и его изменения должны запоминаться.

Теперь мы можем сформулировать алгоритм *полного исключения*, с помощью которого находятся все решения произвольной системы линейных алгебраических уравнений с расширенной матрицей

$$\left[\begin{array}{cccc|c} a_{11}^{(0)} & a_{12}^{(0)} & \dots & a_{1n}^{(0)} & b_1^{(0)} \\ a_{21}^{(0)} & a_{22}^{(0)} & \dots & a_{2n}^{(0)} & b_2^{(0)} \\ \dots & \dots & \dots & \dots & \dots \\ a_{m1}^{(0)} & a_{m2}^{(0)} & \dots & a_{mn}^{(0)} & b_m^{(0)} \end{array} \right].$$

Алгоритм полного исключения.

1-й шаг.

1) Находим в матрице коэффициентов системы *ведущий* (наибольший по модулю) элемент. Если наибольших по модулю элементов несколько, то в качестве ведущего может быть взят любой из них.

Замечание. Мы предполагаем, естественно, что не все элементы матрицы коэффициентов равны нулю.

2) Переставляя (если нужно) строки расширенной матрицы и столбцы матрицы коэффициентов, помещаем ведущий элемент в первую строку и в первый столбец. Перестановки столбцов при этом запоминаем!

3) Делим первую строку расширенной матрицы на ведущий элемент, При этом на месте элемента $a_{11}^{(0)}$ появляется единица.

Если матрица состоит из одной строки, *работа алгоритма заканчивается*. Иначе переходим к п.4.

4) Прибавляя ко второй, третьей и т.д. строкам расширенной матрицы ее первую строку, умноженную соответственно на $-a_{21}^{(0)}, -a_{31}^{(0)}$ и т.д., преобразуем эту матрицу к виду

$$\left[\begin{array}{cccc|c} 1 & a_{12}^{(1)} & \dots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & \mathbf{a}_{22}^{(1)} & \dots & \mathbf{a}_{2n}^{(1)} & b_2^{(1)} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \mathbf{a}_{m2}^{(1)} & \dots & \mathbf{a}_{mn}^{(1)} & b_m^{(1)} \end{array} \right]. \quad (1.1.5)$$

Если в расширенной матрице появились нулевые строки, эти строки следует отбросить, уменьшив тем самым число уравнений в системе. Действительно, нулевая строка соответствует уравнению

$$0 \cdot x_1 + \dots + 0 \cdot x_n = 0,$$

которое удовлетворяется тождественно.

Если в расширенной матрице появилась строка вида $[0 \dots 0 | b_r^{(1)}]$, где $b_r^{(1)} \neq 0$, соответствующая уравнению

$$0 \cdot x_1 + \dots + 0 \cdot x_n = b_r^{(1)} \neq 0,$$

которое не удовлетворяется ни при каких значениях переменных, то преобразованная система несовместна, и, следовательно, несовместна равносильная ей исходная система. *Работа алгоритма закончена.*

Если после первого шага работа алгоритма не закончилась (не обнаружена несовместность системы и число строк в преобразованной расширенной матрице больше одной), то переходим ко второму шагу.

2-й шаг.

1) Рассмотрим выделенную жирным шрифтом подматрицу преобразованной матрицы⁴⁹ (1.1.5):

$$\left[\begin{array}{ccccc} a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \dots & \dots & \dots \\ a_{s2}^{(1)} & \dots & a_{sn}^{(1)} \end{array} \right].$$

Найдем в этой подматрице ведущий элемент.

⁴⁹Отметим, что количество строк в этой матрице обозначено буквой s , а не m , так как после первого шага количество строк могло уменьшиться за счет отбрасывания нулевых ($s \leq m$).

2) Переставляя (если нужно) строки расширенной матрицы и столбцы матрицы коэффициентов, помещаем ведущий элемент во вторую строку и во второй столбец. Перестановки столбцов запоминаем!

3) Делим вторую строку расширенной матрицы на ведущий элемент, При этом на месте элемента $a_{22}^{(1)}$ появляется единица.

4) Прибавляя ко всем строкам расширенной матрицы, кроме второй, ее вторую строку, умноженную соответственно на $-a_{12}^{(1)}, -a_{32}^{(1)}$ и т.д., преобразуем эту матрицу к виду

$$\left[\begin{array}{cccc|c} 1 & 0 & a_{13}^{(2)} & \dots & a_{1n}^{(2)} & b_1^{(2)} \\ 0 & 1 & a_{23}^{(2)} & \dots & a_{2n}^{(2)} & b_2^{(2)} \\ 0 & 0 & a_{33}^{(2)} & \dots & a_{3n}^{(2)} & b_3^{(2)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & a_{s3}^{(2)} & \dots & a_{sn}^{(2)} & b_s^{(2)} \end{array} \right].$$

Если после второго шага обнаружена несовместность системы или (после удаления возможно появившихся нулевых строк) расширенная матрица содержит две строки, то *работа алгоритма заканчивается*. Иначе – переходим к 3-му шагу.

Пусть теперь выполнен $(k-1)$ -й шаг, не обнаружена несовместность системы и в матрице осталось $s \geq k$ строк. Переходим к k -му шагу.

k -й шаг.

1) Рассмотрим подматрицу преобразованной матрицы коэффициентов

$$\left[\begin{array}{ccc} a_{kk}^{(k-1)} & \dots & a_{kn}^{(k-1)} \\ \dots & \dots & \dots \\ a_{sk}^{(k-1)} & \dots & a_{sk}^{(k-1)} \end{array} \right].$$

Находим в этой подматрице ведущий элемент.

2) Переставляя (если нужно) строки расширенной матрицы и столбцы матрицы коэффициентов, помещаем ведущий элемент в k -ю строку и в k -й столбец. Перестановки столбцов при этом запоминаем!

3) Делим k -ю строку расширенной матрицы на ведущий элемент. При этом на месте элемента $a_{kk}^{(k-1)}$ появляется единица.

4) Прибавляя ко всем строкам расширенной матрицы, кроме k -й, ее k -ю строку, умноженную соответственно на $-a_{1k}^{(k-1)}, \dots, -a_{k-1,k}^{(k-1)}$, $-a_{k+1,k}^{(k-1)}, \dots, -a_{s,k}^{(k-1)}$, преобразуем эту матрицу к виду

$$\left[\begin{array}{cccc|ccc} 1 & 0 \dots 0 & a_{1,k+1}^{(k)} & \dots & a_{1n}^{(k)} & b_1^{(k)} \\ 0 & 1 \dots 0 & a_{2,k+1}^{(k)} & \dots & a_{2n}^{(k)} & b_2^{(k)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 \dots 1 & a_{k,k+1}^{(k)} & \dots & a_{kn}^{(k)} & b_k^{(k)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 \dots 0 & a_{s,k+1}^{(k)} & \dots & a_{sn}^{(k)} & b_s^{(k)} \end{array} \right].$$

Если после k -го шага обнаружена несовместность системы или (после удаления возможно появившихся нулевых строк) расширенная матрица содержит k строк, то *работа алгоритма заканчивается*. Иначе – переходим к $k + 1$ -му шагу.

Очевидно, что после выполнения не более, чем m шагов (m – число уравнений системы) работа алгоритма полного исключения заканчивается одним из трех исходов:

- 1) обнаруживается несовместность системы;
- 2) расширенная матрица системы имеет вид

$$\left[\begin{array}{cccc|c} 1 & 0 \dots 0 & b_1^{(r)} \\ 0 & 1 \dots 0 & b_2^{(r)} \\ \dots & \dots & \dots \\ 0 & 0 \dots 1 & b_n^{(r)} \end{array} \right] \quad (r \text{ -- число шагов}),$$

т.е. на месте столбца свободных членов получено *единственное* решение системы.

Напомним, что в процессе исключения столбцы матрицы коэффициентов могли меняться местами (перестановки столбцов запоминались!). Поэтому элементы полученного столбца свободных членов следует переупорядочить (см. пример ниже);

- 3) расширенная матрица системы имеет вид

$$\left[\begin{array}{cccc|c} 1 & 0 \dots 0 & a_{1,s+1}^{(r)} & \dots & a_{1,n}^{(r)} & b_1^{(r)} \\ 0 & 1 \dots 0 & a_{2,s+1}^{(r)} & \dots & a_{2,n}^{(r)} & b_2^{(r)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 \dots 1 & a_{s,s+1}^{(r)} & \dots & a_{s,n}^{(r)} & b_s^{(r)} \end{array} \right],$$

т.е. система имеет бесконечно много решений, которые находятся по формулам (1.1.4).

Терминологическое замечание. Алгоритм полного исключения называют также алгоритмом Гаусса–Йордана⁵⁰.

Серьезное предупреждение. Мы предполагаем, что элементы расширенной матрицы системы заданы *точно*, а все арифметические операции выполняются *без округления или усечения*. Только в этом редко встречающемся случае верны доказанные выше утверждения. Влияние погрешностей исходных данных и погрешностей, вносимых при вычислениях, будет рассмотрено в главе 13.

Рассмотрим несколько несложных примеров, задавая системы их расширенными матрицами. Над столбцами матрицы коэффициентов будем указывать из порядковые номера.

Пример 1.

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 14 \\ 1 & 2 & 3 & 32 \\ 4 & 5 & 6 & 49 \\ 7 & 9 & 8 & \end{array} \right].$$

1-й шаг.

- 1) Ведущий элемент 1-го шага $a_{32}^{(0)} = 9$.
- 2) Меняем местами первую и третью строки, первый и второй столбцы. Перестановки столбцов (изменение порядка переменных) запоминаем!

$$\left[\begin{array}{ccc|c} 2 & 1 & 3 & 49 \\ 9 & 7 & 8 & 14 \\ 5 & 4 & 6 & 32 \\ 2 & 1 & 3 & \end{array} \right].$$

⁵⁰Карл Фридрих ГАУСС (K.F. Gauss, 1777-1855) – немецкий математик, астроном, физик и геодезист, внесший существенный вклад практически во все области математики, почетный член Петербургской АН.

Вильгельм ЙОРДАН (W. Jordan, 1842-1899) – немецкий геодезист. Часто метод полного исключения ошибочно связывают с именем французского математика К.М.Э. ЖОРДАНА (C.M.E. Jordan).

3) Делим первую строку расширенной матрицы на ведущий элемент.

$$\left[\begin{array}{ccc|c} 2 & 1 & 3 & \frac{49}{9} \\ 1 & \frac{7}{9} & \frac{8}{9} & \frac{49}{9} \\ 5 & 4 & 6 & 32 \\ 2 & 1 & 3 & 14 \end{array} \right].$$

4) Прибавляя ко второй и третьей строкам расширенной матрицы ее первую строку, умноженную соответственно на -5 и -2 , получаем расширенную матрицу

$$\left[\begin{array}{ccc|c} 2 & 1 & 3 & \frac{49}{9} \\ 1 & \frac{7}{9} & \frac{8}{9} & \frac{49}{9} \\ 0 & \frac{1}{9} & \frac{14}{9} & \frac{43}{9} \\ 0 & -\frac{5}{9} & \frac{11}{9} & \frac{28}{9} \end{array} \right].$$

2-й шаг.

1) Ведущий элемент 2-го шага $a_{23}^{(1)} = \frac{14}{9}$.

2) Помещаем ведущий элемент во вторую строку и второй столбец, переставляя второй и третий столбцы. Перестановки столбцов запоминаем!

$$\left[\begin{array}{ccc|c} 2 & 3 & 1 & \frac{49}{9} \\ 1 & \frac{8}{9} & \frac{7}{9} & \frac{49}{9} \\ 0 & \frac{14}{9} & \frac{1}{9} & \frac{43}{9} \\ 0 & \frac{11}{9} & -\frac{5}{9} & \frac{28}{9} \end{array} \right].$$

3) Делим вторую строку расширенной матрицы на ведущий элемент.

$$\left[\begin{array}{ccc|c} 2 & 3 & 1 & \frac{49}{9} \\ 1 & \frac{8}{9} & \frac{7}{9} & \frac{49}{9} \\ 0 & 1 & \frac{1}{14} & \frac{43}{14} \\ 0 & \frac{11}{9} & -\frac{5}{9} & \frac{28}{9} \end{array} \right].$$

4) Прибавляя к первой и третьей строкам расширенной матрицы ее вторую строку, умноженную соответственно на $-\frac{8}{9}$ и $-\frac{11}{9}$, получаем расширенную матрицу

$$\left[\begin{array}{ccc|c} 2 & 3 & 1 & \frac{19}{7} \\ 1 & 0 & \frac{5}{7} & \frac{19}{7} \\ 0 & 1 & \frac{1}{14} & \frac{43}{14} \\ 0 & 0 & -\frac{9}{14} & -\frac{9}{14} \end{array} \right].$$

3-й шаг.

Подматрица, получающаяся из матрицы коэффициентов при удалении первой и второй строк, первого и второго столбца, состоит из единственного элемента $-\frac{9}{14}$. Он и является ведущим элементом 3-го шага и стоит на положенном ему месте. Пункты 1) и 2) уже выполнены.

3) Делим третью строку расширенной матрицы на ведущий элемент.

$$\left[\begin{array}{ccc|c} 2 & 3 & 1 & \frac{19}{7} \\ 1 & 0 & \frac{5}{7} & \frac{7}{7} \\ 0 & 1 & \frac{1}{14} & \frac{43}{14} \\ 0 & 0 & 1 & 1 \end{array} \right].$$

4) Прибавляя к первой и второй строкам расширенной матрицы ее третью строку, умноженную соответственно на $-\frac{5}{7}$ и $-\frac{1}{14}$, получаем расширенную матрицу

$$\left[\begin{array}{ccc|c} 2 & 3 & 1 & 2 \\ 1 & 0 & 0 & 3 \\ 0 & 1 & 0 & 1 \end{array} \right].$$

Работа алгоритма закончена. На месте столбца свободных членов стоит единственное решение системы. Порядок переменных указан в строке, стоящей над матрицей коэффициентов. Учитывая его, получаем

$$x_2 = 2, \quad x_3 = 3, \quad x_1 = 1.$$

Пример 2.

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 6 \\ 1 & 2 & 3 & 6 \\ 4 & 5 & 6 & 15 \\ 7 & 8 & 9 & 24 \end{array} \right].$$

1-й шаг.

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 6 \\ 1 & 2 & 3 & 6 \\ 4 & 5 & 6 & 15 \\ 7 & 8 & 9 & 24 \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|c} 3 & 2 & 1 & 24 \\ 9 & 8 & 7 & 24 \\ 6 & 5 & 4 & 15 \\ 3 & 2 & 1 & 6 \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|c} 3 & 2 & 1 & \frac{8}{3} \\ 1 & \frac{8}{9} & \frac{7}{9} & \frac{8}{3} \\ 6 & 5 & 4 & 15 \\ 3 & 2 & 1 & 6 \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|c} 3 & 2 & 1 & \frac{8}{3} \\ 1 & \frac{8}{9} & \frac{7}{9} & -1 \\ 0 & -\frac{1}{3} & -\frac{2}{3} & -2 \\ 0 & -\frac{2}{3} & -\frac{4}{3} & -2 \end{array} \right].$$

2-й шаг.

$$\left[\begin{array}{ccc|c} 3 & 2 & 1 & \frac{8}{3} \\ 1 & \frac{8}{9} & \frac{7}{9} & -1 \\ 0 & -\frac{1}{3} & -\frac{2}{3} & -2 \\ 0 & -\frac{2}{3} & -\frac{4}{3} & \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|c} 3 & 1 & 2 & \frac{8}{3} \\ 1 & \frac{7}{9} & \frac{8}{9} & -2 \\ 0 & -\frac{4}{3} & -\frac{2}{3} & -1 \\ 0 & -\frac{2}{3} & -\frac{1}{3} & \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|c} 3 & 1 & 2 & \frac{8}{3} \\ 1 & \frac{7}{9} & \frac{8}{9} & \frac{3}{2} \\ 0 & 1 & \frac{1}{2} & -1 \\ 0 & -\frac{2}{3} & -\frac{1}{3} & \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|c} 3 & 1 & 2 & \frac{3}{2} \\ 1 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & \frac{1}{2} & \frac{3}{2} \\ 0 & 0 & 0 & 0 \end{array} \right].$$

Удаляем полученную нулевую строку.

Работа алгоритма закончена. С помощью равносильных преобразований система уравнений приведена к виду (1.1.3).

$$\left[\begin{array}{ccc|c} 3 & 1 & 2 & \frac{3}{2} \\ 1 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & \frac{1}{2} & \frac{3}{2} \end{array} \right].$$

Перенося слагаемые, содержащие x_2 , в правую часть, получим

$$x_3 = \frac{3}{2} - \frac{1}{2} \cdot x_2; \quad x_1 = \frac{3}{2} - \frac{1}{2} \cdot x_2. \quad (1.1.6)$$

Эта система уравнений (а, следовательно, и равносильная ей исходная) имеет бесконечно много решений, которые получаются из (1.1.6) при произвольно заданном значении x_2 .

Пример 3.

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 6 \\ 1 & 2 & 3 & 12 \\ 4 & 5 & 6 & 12 \\ 7 & 8 & 9 & 15 \end{array} \right].$$

1-й шаг.

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 6 \\ 1 & 2 & 3 & 12 \\ 4 & 5 & 6 & 12 \\ 7 & 8 & 9 & 15 \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|c} 3 & 2 & 1 & 15 \\ 9 & 8 & 7 & 12 \\ 6 & 5 & 4 & 6 \\ 3 & 2 & 1 & 6 \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|c} 3 & 2 & 1 & \frac{5}{3} \\ 1 & \frac{8}{9} & \frac{7}{9} & 12 \\ 6 & 5 & 4 & 6 \\ 3 & 2 & 1 & 6 \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|c} 3 & 2 & 1 & \frac{5}{3} \\ 1 & \frac{8}{9} & \frac{7}{9} & 2 \\ 0 & -\frac{1}{3} & -\frac{2}{3} & 0 \\ 0 & -\frac{2}{3} & -\frac{4}{3} & 1 \end{array} \right].$$

2-й шаг.

$$\left[\begin{array}{ccc|c} 3 & 2 & 1 & \frac{5}{3} \\ \hline 1 & \frac{8}{9} & \frac{7}{9} & 2 \\ 0 & -\frac{1}{3} & -\frac{2}{3} & 1 \\ 0 & -\frac{2}{3} & -\frac{4}{3} & 1 \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|c} 3 & 1 & 2 & \frac{5}{3} \\ \hline 1 & \frac{7}{9} & \frac{8}{9} & 2 \\ 0 & -\frac{4}{3} & -\frac{2}{3} & 1 \\ 0 & -\frac{2}{3} & -\frac{1}{3} & 2 \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|c} 3 & 1 & 2 & \frac{5}{3} \\ \hline 1 & \frac{7}{9} & \frac{8}{9} & 2 \\ 0 & 1 & \frac{1}{2} & \frac{3}{4} \\ 0 & -\frac{2}{3} & -\frac{1}{3} & 2 \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|c} 3 & 1 & 2 & \frac{9}{4} \\ \hline 1 & 0 & \frac{1}{2} & -\frac{3}{4} \\ 0 & 1 & \frac{1}{2} & \frac{3}{2} \\ 0 & 0 & 0 & \frac{3}{2} \end{array} \right].$$

Работа алгоритма закончена. Третье уравнение полученной системы имеет вид

$$0 \cdot x_3 + 0 \cdot x_1 + 0 \cdot x_2 = \frac{3}{2}.$$

Оно не удовлетворяется ни при каких значениях переменных. Полученная система (а, следовательно, и равносильная ей исходная) *несовместна*.

1.2. Однородные системы линейных уравнений

Определение. Система линейных уравнений называется *однородной*, если все ее свободные члены равны нулю.

Теорема. Если число уравнений в однородной системе меньше, чем число переменных, то система имеет бесконечно много решений.

Доказательство. 1. Всякая однородная система имеет нулевое решение. Это утверждение проверяется подстановкой. Следовательно, однородная система не может быть несовместной.

2. В исходной системе количество строк меньше количества столбцов. При работе алгоритма Гаусса–Йордана количество переменных не меняется, а количество уравнений не растет (уменьшиться оно может за счет отбрасывания нулевых строк расширенной матрицы). Следовательно, результирующая матрица коэффициентов квадратной быть не может, т.е. решение не может быть единственным. Теорема доказана. ■

Глава 2. АЛГЕБРА МАТРИЦ

Матрицы, введенные нами для компактной записи систем линейных уравнений, имеют право и на самостоятельное существование. В этой главе мы рассмотрим операции над матрицами – матричную алгебру. Отметим сразу, что все операции матричной алгебры реализованы в средах конечного пользователя и в виде Фортран-программ.

2.1. Транспонирование и эрмитово сопряжение

Определение. Если A – строка ширины n , то матрицу-столбец высоты n , элементы которой равны соответствующим элементам A , называют *транспонированной по отношению к A* и обозначают A^T .

Если B – столбец высоты n , то матрицу-строку ширины n , элементы которой равны соответствующим элементам B , называют *транспонированной по отношению к B* и обозначают B^T .

Примеры.

$$\text{Если } B = \begin{bmatrix} 4 & 5 \end{bmatrix}, \text{ то } B^T = \begin{bmatrix} 4 \\ 5 \end{bmatrix}; \quad \text{если } A = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \text{ то } A^T = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}.$$

Определение. Если A – $(m \times n)$ -матрица, то $(n \times m)$ -матрицу, столбцы которой равны соответствующим транспонированным строкам A (строки равны соответствующим транспонированным столбцам A), называют *транспонированной по отношению к A* и обозначают A^T .

$$a_{ij}^T = a_{ji} \quad (i = 1, \dots, m; j = 1, \dots, n).$$

Пример.

$$\text{Если } A = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}, \text{ то } A^T = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}.$$

Определение. Квадратная матрица A , удовлетворяющая условию $A^T = A$, называется *симметричной*. Для симметричной матрицы порядка n выполнены равенства

$$a_{ij}^T = a_{ji} = a_{ij} \quad (i = 1, \dots, n; j = 1, \dots, n).$$

Определение. Если транспонировать матрицу A , а затем заменить каждый ее элемент комплексно сопряженным ему числом, то получится

матрица, которую называют *эрмитовой⁵¹ сопряженной* (или просто сопряженной) по отношению к A и обозначают A^* .

$$a_{ij}^* = \overline{a_{ji}} \quad (i = 1, \dots, m; j = 1, \dots, n).$$

Пример.

Если $A = \begin{bmatrix} 1+2i & 4 \\ 2-i & 5+3i \\ 3 & 6-2i \end{bmatrix}$, то $A^* = \begin{bmatrix} 1-2i & 2+i & 3 \\ 4 & 5-3i & 6+2i \end{bmatrix}$.

Имеют место очевидные соотношения

$$(A^T)^T = A; \quad (A^*)^* = A.$$

Определение. Квадратную матрицу, удовлетворяющую условию $A^* = A$, называют *самосопряженной* или *эрмитовой*. Для эрмитовой матрицы

$$a_{ij}^* = \overline{a_{ji}} = a_{ij} \quad (i = 1, \dots, n; j = 1, \dots, n)$$

(элементы, симметричные относительно диагонали, комплексно сопряжены). Отсюда, в частности, следует, что диагональные элементы эрмитовой матрицы вещественны.

2.2. Линейные операции над матрицами

Линейными операциями называют умножение матрицы на число и сложение матриц. Линейные операции выполняются *поэлементно*.

Определение. Если $A - (m \times n)$ -матрица, а α – число, то символом αA обозначают матрицу, получающуюся из A умножением каждого ее элемента на α .

$$B = \alpha A \iff b_{ij} = \alpha \cdot a_{ij} \quad (i = 1, \dots, m; j = 1, \dots, n).$$

Пример.

$$(-1.5) \cdot \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} -1.5 & -3.0 & -4.5 \\ -6.0 & -7.5 & -9.0 \end{bmatrix}.$$

Определение. Если A и $B - (m \times n)$ -матрицы *одного размера*, то их суммой называют матрицу, получающуюся сложением соответствующих элементов A и B .

⁵¹Шарль ЭРМИТ (C. Hermite, 1822-1901) – французский математик, член Парижской АН, почетный член Петербургской АН.

$$C = A + B \iff c_{ij} = a_{ij} + b_{ij} \quad (i = 1, \dots, m; j = 1, \dots, n).$$

Пример. $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix} = \begin{bmatrix} 8 & 10 & 12 \\ 14 & 16 & 18 \end{bmatrix}.$

Матрицы разных размеров складывать нельзя!

Из определений очевидны следующие свойства линейных операций над матрицами (матрицы A, B, C, Θ – одного размера!):

- 1) $B + A = A + B;$
- 2) $A + (B + C) = (A + B) + C;$
- 3) $A + \Theta = \Theta + A = A$

(здесь Θ – нулевая матрица, т.е. матрица, все элементы которой – нули);

- 4) $\alpha(A + B) = \alpha A + \alpha B; \quad (\alpha + \beta)A = \alpha A + \beta A;$
- 5) $(A + B)^T = A^T + B^T; \quad (A + B)^* = A^* + B^*;$
- 6) $(\alpha A)^T = \alpha A^T; \quad (\alpha A)^* = \bar{\alpha} A^*$

(здесь α и β – числа).

2.3. Умножение матриц

Умножение матрицы на матрицу определим сначала для случая, когда *левый сомножитель – строка* ширины n , а *правый сомножитель – столбец* высоты n .

$$[a_1, \dots, a_n] \cdot \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = [a_1 \cdot b_1 + \dots + a_n \cdot b_n] = a_1 \cdot b_1 + \dots + a_n \cdot b_n.$$

Произведение строки (слева) и столбца (справа) есть (1×1) -матрица, которую мы ранее договорились отождествлять с числом.

Пример. $\begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = 1 \cdot 4 + 2 \cdot 5 + 3 \cdot 6 = 32.$

Рассмотрим теперь общий случай. Пусть левый сомножитель – A – имеет строки ширины n , а правый – B – столбцы высоты n . Тогда матрица-произведение $C = A \cdot B$ определяется так: элемент c_{ik} есть произведение i -й строки левого сомножителя на k -й столбец правого.

$$C = A \cdot B \iff c_{ik} = a_{i1}b_{1k} + \dots + a_{in}b_{nk} = \sum_{j=1}^n a_{ij}b_{jk}.$$

Пример.

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \cdot \begin{bmatrix} 7 & 10 \\ 8 & 11 \\ 9 & 12 \end{bmatrix} = \begin{bmatrix} 50 & 68 \\ 122 & 167 \end{bmatrix}.$$

$$1 \cdot 10 + 2 \cdot 11 + 3 \cdot 12 = 68.$$

Замечание. Из определения следует, что произведение $(m \times p)$ -матрицы на $(p \times n)$ -матрицу есть $(m \times n)$ -матрица.

Матрицы, размеры которых не согласованы, т.е. ширина левой не равна высоте правой, перемножать нельзя!

Известно, что умножение чисел

- а) коммутативно: $a \cdot b = b \cdot a$,
- б) ассоциативно: $a \cdot (b \cdot c) = (a \cdot b) \cdot c$,
- в) дистрибутивно относительно сложения: $a \cdot (b + c) = a \cdot b + a \cdot c$.

Покажем, что умножение матриц, вообще говоря, *не коммутативно*.

Во-первых, при изменении порядка сомножителей может нарушиться согласованность размеров. Например, если A – (3×4) -матрица, а B – (4×2) -матрица, то AB – (3×2) -матрица, а произведение BA не определено.

Произведение не определено.

Во-вторых, даже если произведения двух матриц определены при любом порядке сомножителей, то размеры этих произведений могут быть различными. Так, например, если A – (2×4) -матрица, а B – (4×2) -матрица, то AB – (2×2) -матрица, а BA – (4×4) -матрица.

$$\begin{array}{|c|c|} \hline \text{ } & \text{ } \\ \hline \text{ } & \text{ } \\ \hline \end{array} \cdot \begin{array}{|c|c|} \hline \text{ } & \text{ } \\ \hline \text{ } & \text{ } \\ \hline \end{array} = \begin{array}{|c|c|} \hline \text{ } & \text{ } \\ \hline \text{ } & \text{ } \\ \hline \end{array}; \quad \begin{array}{|c|c|} \hline \text{ } & \text{ } \\ \hline \text{ } & \text{ } \\ \hline \end{array} \cdot \begin{array}{|c|c|c|} \hline \text{ } & \text{ } & \text{ } \\ \hline \text{ } & \text{ } & \text{ } \\ \hline \end{array} = \begin{array}{|c|c|c|} \hline \text{ } & \text{ } & \text{ } \\ \hline \text{ } & \text{ } & \text{ } \\ \hline \end{array}.$$

Наконец, произведения квадратных матриц одного порядка заведомо определены при любом расположении сомножителей и представляют собой квадратные матрицы того же порядка. Однако и в этом случае произведения могут зависеть от расположения сомножителей. Например:

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}; \quad \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

В то же время существуют такие пары *квадратных* матриц, что $AB = BA$. В таком случае говорят, что матрицы A и B *коммутируют*. Например:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} 0 & 2 \\ 3 & 3 \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 3 & 3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 6 & 8 \\ 12 & 18 \end{bmatrix}.$$

Определение. Квадратная матрица, у которой все *внедиагональные* элементы равны нулю, называется *диагональной*. Мы будем писать

$$diag[d_1, d_2, \dots, d_n] = \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_n \end{bmatrix}.$$

Отметим, что умножение матрицы A на диагональную матрицу D слева приводит к умножению строк A на соответствующие (по номерам) элементы диагонали D , а умножение на D справа – к умножению на те же элементы столбцов A .

Примеры.

$$diag[2, 3, 4] \cdot \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 3 & 3 \\ 4 & 4 \end{bmatrix};$$

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \cdot diag[2, 3] = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 2 & 3 \\ 2 & 3 \end{bmatrix}.$$

Определение. Матрица $I = diag[1, 1, \dots, 1]$ называется *единичной*. Это название объясняется тем, что она коммутирует с любой квадратной

матрицей того же порядка, причем $A \cdot I = I \cdot A = A$. Если необходимо указать порядок единичной матрицы, пишут I_n . Если A – $(m \times n)$ -матрица, то $I_m \cdot A = A \cdot I_n = A$.

Остальные два свойства умножения чисел – ассоциативность и дистрибутивность относительно сложения – верны и для матриц (конечно, при условии согласованности размеров). Докажем это.

Ассоциативность. Пусть матрица A имеет размер $(m \times n)$, матрица B – $(n \times k)$ и матрица C – $(k \times s)$. Тогда

$$(A \cdot B) \cdot C = A \cdot (B \cdot C).$$

Доказательство. Определим матрицы

$$P = A \cdot B, \quad Q = B \cdot C, \quad U = P \cdot C = (A \cdot B) \cdot C, \quad V = A \cdot Q = A \cdot (B \cdot C)$$

(проверьте согласованность размеров!).

Докажем равенство матриц U и V , имеющих общий размер $(m \times s)$, рассмотрев их соответствующие элементы:

$$\begin{aligned} u_{ij} &= \sum_{r=1}^k p_{ir} c_{rj} = \sum_{r=1}^k \left(\sum_{t=1}^n a_{it} b_{tr} \right) c_{rj} = \sum_{r=1}^k \sum_{t=1}^n a_{it} b_{tr} c_{rj}; \\ v_{ij} &= \sum_{t=1}^n a_{it} q_{tj} = \sum_{t=1}^n a_{it} \left(\sum_{r=1}^k b_{tr} c_{rj} \right) = \sum_{t=1}^n \sum_{r=1}^k a_{it} b_{tr} c_{rj}; \end{aligned}$$

Видно, что u_{ij} и v_{ij} представляют собой суммы одних и тех же слагаемых и поэтому совпадают. ■

Дистрибутивность относительно сложения. Пусть A – матрица размера $(m \times n)$, а B и C – матрицы размера $(n \times k)$. Тогда

$$A \cdot (B + C) = A \cdot B + A \cdot C.$$

Доказательство. Определим матрицы

$$R = A \cdot (B + C), \quad P = A \cdot B, \quad Q = A \cdot C.$$

(проверьте согласованность размеров!). Докажем, что $R = P + Q$. Действительно,

$$r_{ij} = \sum_{t=1}^n a_{it} (b_{tj} + c_{tj}) = \sum_{t=1}^n a_{it} b_{tj} + \sum_{t=1}^n a_{it} c_{tj} = p_{ij} + q_{ij}. \quad ■$$

Точно так же доказывается равенство

$$(L + M) \cdot N = L \cdot N + M \cdot N,$$

где L, M, N – матрицы с согласованными размерами.

Взаимодействие умножения матриц с операциями транспонирования и эрмитова сопряжения задается соотношениями

$$(A \cdot B)^T = B^T \cdot A^T; \quad (A \cdot B)^* = B^* \cdot A^*.$$

Проверьте эти равенства, обратив внимание на порядок сомножителей слева и справа.

В терминах умножения матриц может быть записана система линейных алгебраических уравнений. Пусть A – заданная числовая матрица размера $(m \times n)$, b – заданный числовой столбец высоты m , x – переменный столбец высоты n , т.е. $x = [x_1, \dots, x_n]^T$, где x_1, \dots, x_n – числовые переменные. Тогда

$$Ax = b \quad \text{или} \quad \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \dots \\ b_m \end{bmatrix}$$

есть краткая запись системы линейных уравнений

$$\begin{cases} a_{11}x_1 + \dots + a_{1n}x_n = b_1 \\ \dots \\ a_{m1}x_1 + \dots + a_{mn}x_n = b_m \end{cases}.$$

2.4. Матричные уравнения

Рассмотрим уравнение

$$AX = B, \tag{2.4.1}$$

где A – заданная числовая $(m \times n)$ -матрица, B – заданная числовая $(m \times p)$ -матрица, X – переменная $(n \times p)$ -матрица:

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \cdot \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} b_{11} & \dots & b_{1p} \\ \dots & \dots & \dots \\ b_{m1} & \dots & b_{mp} \end{bmatrix}.$$

Решением этого уравнения называется такая числовая $(n \times p)$ -матрица \tilde{X} , подстановка которой вместо переменной матрицы X превращает уравнение в равенство матриц $A\tilde{X} = B$.

Вспоминая правило умножения матриц, запишем уравнение (2.4.1) "по столбцам":

$$A \cdot \begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix} = \begin{bmatrix} b_{11} \\ \vdots \\ b_{m1} \end{bmatrix}; \dots; A \cdot \begin{bmatrix} x_{1p} \\ \vdots \\ x_{np} \end{bmatrix} = \begin{bmatrix} b_{1p} \\ \vdots \\ b_{mp} \end{bmatrix}. \quad (2.4.2)$$

Отсюда видно, что система линейных уравнений (1.1.1) есть частный случай матричного уравнения (2.4.1) при $p = 1$, и что матричное уравнение (2.4.1) равносильно p системам линейных уравнений.

Все системы в (2.4.2) имеют общую матрицу коэффициентов. Применяя к ним алгоритм Гаусса–Йордана, мы будем многократно повторять одни и те же вычисления – отличие будет только при работе со свободными членами. Поэтому следует решать эти системы одновременно. Технология видна из приведенного ниже примера.

Пример.

$$\begin{bmatrix} 1 & 2 & 4 \\ 1 & 2 & 3 \\ 2 & 3 & 4 \end{bmatrix} \cdot X = \begin{bmatrix} 17 & 8 \\ 14 & 6 \\ 20 & 9 \end{bmatrix}.$$

Записываем расширенную матрицу (над чертой – номера столбцов).

$$\left[\begin{array}{ccc|cc} 1 & 2 & 3 & 17 & 8 \\ 1 & 2 & 4 & 14 & 6 \\ 1 & 2 & 3 & 20 & 9 \end{array} \right]$$

1-й шаг.

$$\left[\begin{array}{ccc|cc} 1 & 2 & 3 & 17 & 8 \\ 1 & 2 & 4 & 14 & 6 \\ 1 & 2 & 3 & 20 & 9 \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|cc} 3 & 2 & 1 & 17 & 8 \\ 4 & 2 & 1 & 14 & 6 \\ 3 & 2 & 1 & 20 & 9 \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|cc} 3 & 2 & 1 & \frac{17}{4} & 2 \\ 1 & \frac{1}{2} & \frac{1}{4} & 14 & 6 \\ 3 & 2 & 1 & 20 & 9 \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|cc} 3 & 2 & 1 & \frac{17}{4} & 2 \\ 1 & \frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{4} & 0 & 1 \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|cc} 3 & 2 & 1 & \frac{17}{4} & 2 \\ 1 & \frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{2} \\ 0 & 1 & 1 & 0 & 1 \end{array} \right]$$

2-й шаг.

$$\left[\begin{array}{ccc|cc} 3 & 2 & 1 & \frac{17}{4} & 2 \\ 1 & \frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{4} & 0 & 1 \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|cc} 3 & 2 & 1 & \frac{17}{4} & 2 \\ 1 & \frac{1}{2} & \frac{1}{4} & 0 & 1 \\ 0 & \frac{1}{2} & \frac{1}{4} & 0 & 0 \end{array} \right] \Leftrightarrow \left[\begin{array}{ccc|cc} 3 & 2 & 1 & \frac{11}{4} & \frac{3}{2} \\ 1 & 0 & -\frac{1}{4} & 0 & 1 \\ 0 & 1 & 1 & -\frac{1}{4} & -\frac{1}{2} \end{array} \right]$$

3-й шаг.

$$\left| \begin{array}{ccc|cc} 3 & 2 & 1 & \frac{11}{4} & \frac{3}{2} \\ 1 & 0 & -\frac{1}{4} & 3 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{2} \end{array} \right| \Leftrightarrow \left| \begin{array}{ccc|cc} 3 & 2 & 1 & \frac{11}{4} & \frac{3}{2} \\ 1 & 0 & -\frac{1}{4} & 3 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 2 \end{array} \right| \Leftrightarrow \left| \begin{array}{ccc|cc} 3 & 2 & 1 & 3 & 2 \\ 1 & 0 & 0 & 2 & -1 \\ 0 & 1 & 0 & 1 & 2 \\ 0 & 0 & 1 & 1 & 2 \end{array} \right|$$

$$X = \begin{bmatrix} 1 & 2 \\ 2 & -1 \\ 3 & 2 \end{bmatrix}. \quad \text{Обратите внимание на порядок строк!}$$

Матричное уравнение

$$XA = B \quad (2.4.3)$$

приводят к виду (2.4.1), транспонируя обе его части:

$$A^T X^T = B^T.$$

Найдя описанным выше способом матрицу X^T , еще раз применяют операцию транспонирования:

$$X = (X^T)^T.$$

2.5. Обратная матрица

Определение. Если $A - (m \times n)$ -матрица, и существует решение уравнения $AX = I_m$, то это решение – $(n \times m)$ -матрицу X – называют *правой обратной матрицей* для A . Аналогично, если существует решение уравнения $YA = I_n$, то $(n \times m)$ -матрицу Y называют *левой обратной матрицей* для A .

Покажем, что из существования у матрицы и левой и правой обратных следует их совпадение.

$$\begin{aligned} (AX = I_m) \wedge (YA = I_n) &\implies \\ \implies Y = YI_m = Y(AX) &= (YA)X = I_nX = X. \end{aligned} \quad (2.5.1)$$

В этом случае матрицу $X = Y$ называют *обратной* к A и обозначают A^{-1} , а саму матрицу A называют *обратимой*.

Пример. Решив матричное уравнение

$$\begin{bmatrix} 1 & 2 & 4 \\ 1 & 2 & 3 \\ 2 & 3 & 4 \end{bmatrix} \cdot X = I, \quad \text{получим} \quad X = \begin{bmatrix} 1 & -4 & 2 \\ -2 & 4 & -1 \\ 1 & -1 & 0 \end{bmatrix}.$$

Убедитесь в том, что $XA = I$ и, следовательно, $X = A^{-1}$.

Если матрица A обратима, то решение матричного уравнения (2.4.1) существует, единственno и задается формулой

$$X = A^{-1}B. \quad (2.5.2)$$

Действительно, умножив (2.4.1) на A^{-1} слева, получим (2.5.2), умножив же (2.5.2) на A слева, получим (2.4.1).

Аналогично, если A обратима, то решение матричного уравнения (2.4.3) существует, единственno и задается формулой

$$X = BA^{-1}. \quad (2.5.3)$$

Если матрица A обратима, то обратимы и матрицы A^{-1} , A^T , A^* . Убедитесь в справедливости равенств

$$(A^{-1})^{-1} = A; \quad (A^T)^{-1} = (A^{-1})^T; \quad (A^*)^{-1} = (A^{-1})^*.$$

Теорема. Неквадратная матрица не может иметь обратной.

Доказательство. Пусть $A - (m \times n)$ -матрица, и $m < n$. Применим метод Гаусса–Йордана к уравнению $AX = I_m$. Поскольку число переменных в процессе исключения не меняется, а число уравнений разве что уменьшается, решение единственным быть не может – A необратима.

Если $m > n$, то такое же рассуждение показывает, что необратима $(n \times m)$ -матрица A^T , а, следовательно, необратима и A . ■

Итак, обратимыми могут быть только квадратные матрицы. Условия обратимости квадратных матриц будут получены в п.3.4.

Серьезное предупреждение. Может создаться впечатление, что для решения матричного уравнения $AX = B$ нужно найти матрицу A^{-1} и воспользоваться формулами (2.5.2) или (2.5.3). На самом деле этот путь бессмыслен, так как для нахождения обратной матрицы все равно нужно

решить матричное уравнение $AX = I$. Более того, матрица может быть необратимой, а уравнение $AX = B$ – все-таки разрешимым. Поэтому формулы (2.5.2) и (2.5.3) применяются для теоретических построений и никогда не используются в вычислениях.

2.6. Сведение комплексного матричного уравнения к вещественному

Рассмотрим матричное уравнение

$$CZ = W. \quad (2.6.1)$$

Здесь $C = A + iB$ – заданная комплексная $(m \times n)$ -матрица, $W = U + iV$ – заданная комплексная $(m \times p)$ -матрица и $Z = X + iY$ – переменная комплексная $(n \times p)$ -матрица, а A, B, U, V, X, Y – вещественные матрицы соответствующих размеров.

Перепишем (2.6.1) в виде

$$(A + iB) \cdot (X + iY) = U + iV. \quad (2.6.2)$$

Приравнивая по отдельности вещественные и мнимые части уравнений (2.6.2), получим систему *вещественных* матричных уравнений

$$\begin{cases} AX - BY = U \\ BX + AY = V \end{cases}.$$

Нетрудно видеть, что эта система равносильна одному матричному уравнению

$$\begin{bmatrix} A & | & -B \\ -- & - & -- \\ B & | & A \end{bmatrix} \cdot \begin{bmatrix} X \\ - \\ Y \end{bmatrix} = \begin{bmatrix} U \\ - \\ V \end{bmatrix}, \quad (2.6.3)$$

которое называют овеществлением уравнения (2.6.1).

Заметим, что работать с уравнением (2.6.1) выгоднее, чем с (2.6.3), так как оно требует в два раза меньше оперативной памяти компьютера. Однако программы, работающие с комплексными уравнениями, к сожалению, до сих пор встречаются у нас реже, чем потребность в них.

Глава 3. ОПРЕДЕЛИТЕЛЬ КВАДРАТНОЙ МАТРИЦЫ

3.1. Определение. Примеры

Зададим на множестве всех *квадратных числовых матриц* функцию, которая ставит в соответствие каждой такой матрице число, называемое ее *определителем (детерминантом)*. Определитель квадратной матрицы A принято обозначать символом $\det(A)$.

Определение этой функции построим индуктивно.

Определение. *Определителем матрицы первого порядка* называется число – ее единственный элемент:

$$\det [a_{11}] = a_{11}.$$

Определение. Пусть теперь $n > 1$ – порядок матрицы, и мы умеем вычислять определитель матрицы $(n - 1)$ -го порядка.

Удалив из матрицы одну строку (i -ю) и один столбец (k -й), получим матрицу $(n - 1)$ -го порядка. Ее определитель (который мы по предположению умеем вычислять) назовем *дополнительным минором* элемента a_{ik} (стоящего на пересечении удаленных строки и столбца). Обозначается дополнительный минор M_{ik} . Число $A_{ik} = (-1)^{i+k} \cdot M_{ik}$ именуется *алгебраическим дополнением* элемента матрицы a_{ik} .

Пример.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}; \quad M_{11} = \det \begin{bmatrix} 5 & 6 \\ 8 & 9 \end{bmatrix}; \quad M_{23} = \det \begin{bmatrix} 1 & 2 \\ 7 & 8 \end{bmatrix};$$

$$A_{11} = (-1)^{1+1} \cdot M_{11} = M_{11}; \quad A_{23} = (-1)^{2+3} \cdot M_{23} = (-1) \cdot M_{23}.$$

Определение. *Определителем квадратной числовой матрицы A порядка $n > 1$* называется число

$$\det(A) = a_{1k}A_{1k} + \dots + a_{nk}A_{nk} = \sum_{i=1}^n a_{ik}A_{ik}; \quad (k = 1, \dots, n). \quad (3, 1, 1)$$

Определитель квадратной матрицы равен сумме произведений элементов *любого* ее столбца на их алгебраические дополнения.

Замечания. 1. Сформулированное определение будет корректно, если доказать, что получаемое число *не зависит от выбора столбца*. Мы не приводим доказательство из-за его технической сложности.

2. Выражение (3.1.1) обычно называют *разложением определителя по k-му столбцу матрицы*.

Примеры. 1. $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$.

Разложим определитель матрицы по ее первому столбцу

$$\begin{aligned} \det(A) &= a_{11}A_{11} + a_{21}A_{21} = a_{11}(-1)^{1+1}M_{11} + a_{21}(-1)^{2+1}M_{21} = \\ &= a_{11}\det[a_{22}] - a_{21}\det[a_{12}] = a_{11}a_{22} - a_{21}a_{12}. \end{aligned}$$

Убедитесь в том, что результат не изменится, если разложение выполнить по второму столбцу.

Определитель матрицы *второго порядка* равен разности между произведением ее диагональных элементов и произведением ее внедиагональных элементов.

2. $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$.

Разложим определитель этой матрицы по ее третьему столбцу

$$\begin{aligned} \det(A) &= a_{13}A_{13} + a_{23}A_{23} + a_{33}A_{33} = \\ &= a_{13}(-1)^{1+3}M_{13} + a_{23}(-1)^{2+3}M_{23} + a_{33}(-1)^{3+3}M_{33} = \\ &= a_{13}\det \begin{bmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} - a_{23}\det \begin{bmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{bmatrix} + a_{33}\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \\ &= a_{13}(a_{21}a_{32} - a_{22}a_{31}) - a_{23}(a_{11}a_{32} - a_{12}a_{31}) + a_{33}(a_{11}a_{22} - a_{12}a_{21}). \end{aligned}$$

Убедитесь, что результат не изменится, если разлагать определитель по первому или по второму столбцу.

Замечание. Вычисление определителя матрицы второго порядка требует выполнения двух умножений и двух сложений. Вычисление определителя матрицы третьего порядка сводится (в соответствии с определением) к вычислению трех определителей матриц второго порядка, выполнению трех умножений и трех сложений. Соответственно,

вычисление определителя матрицы порядка n сводится к вычислению n определителей матриц $(n - 1)$ -го порядка и выполнению n умножений и n сложений.

Обозначим $F(n)$ количество арифметических операций, затрачиваемых на вычисление определителя матрицы порядка n . Из предыдущих рассуждений следует, что

$$F(n) > n \cdot F(n) > F(n - 1), \quad \text{т.е.} \quad F(n) > n!$$

Эта простая оценка показывает, что "по определению" (3.1.1) определители матриц вычислять нельзя. Действительно, если принять, что одна арифметическая операция выполняется за 10^{-9} сек., то для вычисления определителя матрицы 20-го порядка потребуется более $20! \cdot 10^{-9}$ сек. ≈ 77 лет, а для вычисления определителя матрицы 30-го порядка – около 10^{16} лет!

Отметим один класс квадратных матриц, определители которых (в отличие от общего случая) легко вычислять "по определению". Это так называемые *треугольные* матрицы, т.е. квадратные матрицы, у которых равны нулю либо все поддиагональные элементы (*верхние треугольные* матрицы), либо все наддиагональные элементы (*нижние треугольные* матрицы).

Разложим определитель верхней треугольной матрицы по элементам первого столбца:

$$\det \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22} & a_{23} & \dots & a_{2n} \\ 0 & 0 & a_{33} & \dots & a_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & a_{nn} \end{bmatrix} = a_{11} \cdot \det \begin{bmatrix} a_{22} & a_{23} & \dots & a_{2n} \\ 0 & a_{33} & \dots & a_{3n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix}.$$

Полученный определитель опять разлагаем по элементам первого столбца

$$a_{11} \cdot \det \begin{bmatrix} a_{22} & a_{23} & \dots & a_{2n} \\ 0 & a_{33} & \dots & a_{3n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix} = a_{11} \cdot a_{22} \cdot \det \begin{bmatrix} a_{33} & \dots & a_{3n} \\ \dots & \dots & \dots \\ 0 & \dots & a_{nn} \end{bmatrix}.$$

Повторяя эту операцию, получим

$$\det(A) = a_{11}a_{22}\dots a_{nn}.$$

Тот же результат, очевидно получится для нижней треугольной матрицы, если разлагать ее определитель по элементам последнего столбца. Итак,

Определитель треугольной матрицы равен произведению ее диагональных элементов.

В частности, $\det(I) = \det(\text{diag}[1, 1, \dots, 1]) = 1$.

Эффективный метод вычисления определителя произвольной квадратной матрицы будет изложен в п.4.2.

3.2. Свойства определителя

Для начала условимся о компактной записи матрицы, при которой указываются не все ее элементы, а только имена ее столбцов:

$$A = [a^{(1)}, \dots, a^{(n)}],$$

где

$$a^{(1)} = [a_{11}, \dots, a_{n1}]^T, \quad \dots, \quad a^{(n)} = [a_{1n}, \dots, a_{nn}]^T.$$

1. Перестановка двух столбцов матрицы приводит к умножению ее определителя на (-1) .

Доказательство. Переставим в матрице A два *соседних* столбца, стоящих на k -м и $(k+1)$ -м местах:

$$A = [\dots \quad p \quad q \quad \dots]; \quad A' = [\dots \quad q \quad p \quad \dots].$$

Разложим $\det(A)$ по элементам k -го столбца, а $\det(A')$ – по элементам $(k+1)$ -го столбца:

$$\det(A) = p_1 A_{1k} + \dots + p_n A_{nk};$$

$$\det(A') = p_1 A'_{1,k+1} + \dots + p_n A'_{n,k+1}.$$

Поскольку при удалении из матриц A и A' столбца p получается одна и та же матрица, имеем $M_{ik} = M'_{i,k+1}$ ($i = 1, \dots, n$). Поэтому

$$A'_{i,k+1} = (-1)^{i+k+1} M'_{i,k+1} = (-1)(-1)^{i+k} M_{ik} = (-1) \cdot A_{ik},$$

откуда $\det(A') = (-1) \cdot \det(A)$, т.е. при перестановке *соседних* столбцов определитель матрицы умножается на (-1) .

Рассмотрим теперь произвольную пару столбцов

$$A = [\dots \overset{(i)}{p} \dots \overset{(k)}{q} \dots]; \quad A' = [\dots \overset{(i)}{q} \dots \overset{(k)}{p} \dots].$$

Поменяем их местами, последовательно переставляя соседние столбцы:

$$i \leftrightarrow i+1, \quad i+1 \leftrightarrow i+2, \quad \dots, \quad k-1 \leftrightarrow k, \quad k-1 \leftrightarrow k-2, \quad \dots, \quad i+1 \leftrightarrow i.$$

Таких перестановок $2(k - i) - 1$. При каждой из них определитель умножается на (-1) , следовательно, в результате он умножится на $(-1)^{2(k-i)-1} = -1$. ■

2. Определитель матрицы с двумя одинаковыми столбцами равен нулю.

Доказательство. Пусть в матрице есть два одинаковых столбца. Поменяв их местами, мы матрицу не изменим, т.е. $A' = A$, и потому $\det(A') = \det(A)$. Но, согласно свойству 1, $\det(A') = (-1) \cdot \det(A)$. Отсюда $\det(A) = 0$. ■

Зафиксируем в матрице все столбцы, кроме k -го, а k -й сделаем переменным. Тогда каждому значению этого столбца будет соответствовать число – значение определителя матрицы, т.е. на множестве столбцов будет задана функция

$$f(x) = \det([a^{(1)}, \dots, a^{(k-1)}, x, a^{(k+1)}, \dots, a^{(n)}]).$$

Покажем, что f – линейная функция, т.е. что

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$$

для любых числовых столбцов x, y и любых чисел α, β .

3. Определитель матрицы – линейная функция каждого ее столбца.

В частности, умножение столбца матрицы на число приводит к умножению на это число ее определителя.

Доказательство. Разлагая определитель

$$f(\alpha x + \beta y) = \det([\dots, \alpha x + \beta y, \dots])$$

по элементам переменного k -го столбца, получим

$$\begin{aligned} f(\alpha x + \beta y) &= \sum_{i=1}^n (\alpha x_i + \beta y_i) A_{ik} = \alpha \sum_{i=1}^n x_i A_{ik} + \beta \sum_{i=1}^n y_i A_{ik} = \\ &= \alpha \cdot \det([\dots, x, \dots]) + \beta \cdot \det([\dots, y, \dots]) = \alpha f(x) + \beta f(y). \end{aligned}$$

4. Определитель матрицы не изменится, если к некоторому ее столбцу прибавить другой столбец, умножив его предварительно на любое число.

Доказательство. Выделим в матрице два столбца: $A = [\dots p \dots q \dots]$. Тогда в силу свойства 3 для любого числа α

$$\det([\dots p + \alpha q \dots q \dots]) = \det([\dots p \dots q \dots]) + \alpha \cdot \det([\dots q \dots q \dots]).$$

Но, согласно свойству 2, $\det([\dots q \dots q \dots]) = 0$, т.е.

$$\det([\dots p + \alpha q \dots q \dots]) = \det([\dots p \dots q \dots]).$$

5. Сумма произведений элементов столбца матрицы на алгебраические дополнения *соответствующих* элементов *другого* ее столбца равна нулю.

Доказательство. Сделаем k -й столбец матрицы A переменным. Тогда получим тождество относительно x :

$$\det[\dots x \dots] = x_1 A_{1k} + \dots + x_n A_{nk}.$$

Подставив вместо x m -й столбец ($m \neq k$), получим определитель матрицы с одинаковыми столбцами, который равен нулю:

$$0 = \det([\dots \overset{(k)}{a^{(m)}} \dots \overset{(m)}{a^{(m)}} \dots]) = a_{1m} A_{1k} + \dots + a_{nm} A_{nk}.$$

Доказательство следующего утверждения технически сложно, и мы его опускаем.

6. Транспонирование матрицы не меняет ее определителя: $\det(A^T) = \det(A)$.

Следствия. 1. Утверждения 1 – 5, доказанные для столбцов матрицы, верны и для ее строк.

2. Имеет место разложение определителя матрицы по элементам любой ее строки:

$$\det(A) = a_{k1}A_{k1} + \dots + a_{kn}A_{kn} = \sum_{i=1}^n a_{ki}A_{ki} \quad (k = 1, \dots, n).$$

3. Из свойства 6 и определения эрмитова сопряжения следует, что

7. Определители эрмитово сопряженных матриц – сопряженные комплексные числа: $\det(A^*) = \overline{\det(A)}$.

Доказательство следующего свойства также опускается из-за его технической сложности.

8. Определитель произведения двух *квадратных* матриц равен произведению определителей сомножителей: $\det(AB) = \det(A) \cdot \det(B)$.

Для матриц второго порядка проверим это свойство прямым вычислением.

Пусть $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, $B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$. Тогда

$$\det(AB) = \det \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}.$$

Рассматривая первый столбец матрицы-произведения как сумму двух столбцов, получим по свойству 3:

$$\det(AB) = \det \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix} + \det \begin{bmatrix} a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}.$$

Теперь вторые столбцы матриц представлены в виде сумм. Поэтому

$$\begin{aligned} \det(AB) &= \det \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} \\ a_{21}b_{11} & a_{21}b_{12} \end{bmatrix} + \det \begin{bmatrix} a_{11}b_{11} & a_{12}b_{22} \\ a_{21}b_{11} & a_{22}b_{22} \end{bmatrix} + \\ &+ \det \begin{bmatrix} a_{12}b_{21} & a_{11}b_{12} \\ a_{22}b_{21} & a_{21}b_{12} \end{bmatrix} + \det \begin{bmatrix} a_{12}b_{21} & a_{12}b_{22} \\ a_{22}b_{21} & a_{22}b_{22} \end{bmatrix}. \end{aligned}$$

Элементы матрицы B – общие множители в столбцах. По свойству 3

$$\begin{aligned} \det(AB) &= b_{11}b_{12} \cdot \det \begin{bmatrix} a_{11} & a_{11} \\ a_{21} & a_{21} \end{bmatrix} + b_{11}b_{22} \cdot \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \\ &+ b_{21}b_{12} \cdot \det \begin{bmatrix} a_{12} & a_{11} \\ a_{22} & a_{21} \end{bmatrix} + b_{21}b_{22} \cdot \det \begin{bmatrix} a_{12} & a_{12} \\ a_{22} & a_{22} \end{bmatrix}. \end{aligned}$$

Теперь во втором и в третьем слагаемом определители матриц равны соответственно $\det(A)$ и $-\det(A)$, а в первом и в четвертом – нулю (свойство 2). Поэтому

$$\det(AB) = \det(A) \cdot (b_{11}b_{22} - b_{21}b_{12}) = \det(A) \cdot \det(B).$$

Рассмотренные свойства определителя позволяют вычислять его путем преобразования матрицы в треугольную (определитель которой равен произведению диагональных элементов). Технологию рассмотрим на примере так называемой матрицы Вандермонда⁵² (z_1, \dots, z_n – комплексные числа)

$$E = \begin{bmatrix} 1 & z_1 & z_1^2 & \dots & z_1^{n-1} \\ 1 & z_2 & z_2^2 & \dots & z_2^{n-1} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & z_n & z_n^2 & \dots & z_n^{n-1} \end{bmatrix}.$$

Обозначим определитель матрицы Вандермонда $V(z_1, z_2, \dots, z_n)$. Вычтем из n -го столбца матрицы ее $(n-1)$ -й столбец, умноженный на z_1 ; вычтем из $(n-1)$ -го столбца матрицы ее $(n-2)$ -й столбец, умноженный на $z_1; \dots$; вычтем из 2-го столбца матрицы ее 1-й столбец, умноженный на z_1 . В силу свойства 4 определитель не изменится:

$$V(z_1, z_2, \dots, z_n) = \det \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & z_2 - z_1 & z_2(z_2 - z_1) & \dots & z_2^{n-2}(z_2 - z_1) \\ \dots & \dots & \dots & \dots & \dots \\ 1 & z_n - z_1 & z_n(z_n - z_1) & \dots & z_n^{n-2}(z_n - z_1) \end{bmatrix}.$$

Разложим определитель по 1-й строке:

⁵²Александр Теофил ВАНДЕРМОНД (A.T. Vandermonde, 1735-1796) – французский математик, член Парижской АН.

$$V(z_1, z_2, \dots, z_n) = 1 \cdot \det \begin{bmatrix} z_2 - z_1 & z_2(z_2 - z_1) & \dots & z_2^{n-2}(z_2 - z_1) \\ \dots & \dots & \dots & \dots \\ z_n - z_1 & z_n(z_n - z_1) & \dots & z_n^{n-2}(z_n - z_1) \end{bmatrix}.$$

В силу свойства 3 можно вынести за знак определителя общие множители строк $(z_2 - z_1), \dots, (z_n - z_1)$:

$$V(z_1, z_2, \dots, z_n) = (z_2 - z_1) \cdot \dots \cdot (z_n - z_1) \cdot \det \begin{bmatrix} 1 & z_2 & \dots & z_2^{n-2} \\ \dots & \dots & \dots & \dots \\ 1 & z_n & \dots & z_n^{n-2} \end{bmatrix}.$$

Мы видим, что оставшийся определитель равен $V(z_2, \dots, z_n)$. Повторяя эту операцию, понижающую порядок матрицы, в конце концов получим

$$V(z_1, z_2, \dots, z_n) = \prod_{k=2}^n (z_k - z_1) \cdot \prod_{k=3}^n (z_k - z_2) \cdot \dots \cdot (z_n - z_{n-1}) = \prod_{m>k} (z_m - z_k).$$

Заметим, что если z_1, z_2, \dots, z_n – попарно различные числа, то определитель матрицы Вандермонда отличен от нуля.

3.3. Матричные уравнения с квадратной матрицей коэффициентов

Определение. Квадратная матрица, определитель которой равен нулю, называется *вырожденной*.

Применим к уравнению с квадратной матрицей коэффициентов алгоритм Гаусса–Йордана. Используемые в нем элементарные преобразования либо не меняют определителя матрицы коэффициентов, либо умножают определитель на *отличное от нуля* число (проверьте это!).

1. Если определитель матрицы коэффициентов отличен от нуля, то в результате работы алгоритма он не может стать нулем. Поэтому в матрице коэффициентов не может появиться нулевая строка, и по завершении работы алгоритма эта матрица превратится в единичную.

Матричное уравнение с *квадратной невырожденной* матрицей коэффициентов имеет решение, и это решение единственное.

Это утверждение известно как теорема Крамера⁵³.

⁵³Габриэль КРАМЕР (G. Cramer, 1704-1752) – швейцарский математик.

2. Если определитель матрицы коэффициентов равен нулю, то работа алгоритма Гаусса–Йордана не может завершиться превращением этой матрицы в единичную. Следовательно,

Матричное уравнение с *квадратной вырожденной* матрицей коэффициентов либо не имеет решения, либо имеет бесконечно много решений.

Однородное уравнение не может быть несовместным. Поэтому

Однородное матричное уравнение с *квадратной вырожденной* матрицей коэффициентов имеет бесконечно много решений.

3.4. Структура обратной матрицы. Формулы Крамера

Рассмотрим теперь вопрос об обратности квадратной матрицы A , т.е. о существовании решения матричных уравнений $AX = XA = I$.

Для начала заметим, что если матрица A обратима, то

$$\det(A) \cdot \det(A^{-1}) = \det(A \cdot A^{-1}) = \det(I) = 1. \quad (3.4.1)$$

Поэтому вырожденная матрица не может иметь обратной.

Для дальнейшего введем понятие матрицы, *присоединенной* к матрице A .

Пусть A – квадратная матрица. Заменим каждый ее элемент его алгебраическим дополнением и транспонируем результат. Полученную матрицу называют присоединенной к A и обозначают \tilde{A} :

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}; \quad \tilde{A} = \begin{bmatrix} A_{11} & \dots & A_{n1} \\ \dots & \dots & \dots \\ A_{1n} & \dots & A_{nn} \end{bmatrix}.$$

Вычислим произведения $P = A\tilde{A}$ и $Q = \tilde{A}A$:

$$P_{ik} = \sum_{r=1}^n a_{ir}A_{kr} = \delta_{ik} \cdot \det(A)$$

и, аналогично,

$$Q_{ik} = \sum_{r=1}^n A_{ri}a_{rk} = \delta_{ik} \cdot \det(A).$$

Здесь $\delta_{ik} = \begin{cases} 1 & \text{при } i = k \\ 0 & \text{при } i \neq k \end{cases}$ – так называемый *символ Кронекера*⁵⁴.

Итак, $A\tilde{A} = \tilde{A}A = I \cdot \det(A)$.

Если $\det(A) \neq 0$, то

$$A \cdot \left(\frac{1}{\det(A)} \cdot \tilde{A} \right) = \left(\frac{1}{\det(A)} \cdot \tilde{A} \right) \cdot A = I,$$

т.е.

$$A^{-1} = \frac{1}{\det(A)} \cdot \tilde{A}. \quad (3.4.2)$$

Мы не только доказали обратимость невырожденной матрицы, но и получили "явное выражение" для обратной матрицы.

Серьезное предупреждение. Формула (3.4.2) используется только в теоретических построениях. Вычислять обратную матрицу следует, решая уравнение $AX = I$. Убедиться в этом можно, хотя бы сравнив количества операций, необходимых для реализации этих двух методов.

Пусть теперь $Ax = b$ – система уравнений с квадратной невырожденной матрицей.

Найдем решение, умножив ее слева на $A^{-1} = \frac{1}{\det(A)} \cdot \tilde{A}$:

$$x = \frac{1}{\det(A)} \cdot \tilde{A} \cdot b = \frac{1}{\det(A)} \cdot \begin{bmatrix} A_{11} & \dots & A_{n1} \\ \dots & \dots & \dots \\ A_{1n} & \dots & A_{nn} \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ \dots \\ b_n \end{bmatrix}.$$

Отсюда

$$x_k = \frac{b_1 A_{1k} + \dots + b_n A_{nk}}{\det(A)} = \frac{\det(B_k)}{\det(A)}, \quad (k = 1, \dots, n), \quad (3.4.3)$$

где B_k - матрица, получающаяся из A заменой ее k -го столбца столбцом свободных членов.

Формулы (3.4.3) называют формулами Крамера. Они дают представление о структуре решения системы линейных уравнений с квадратной матрицей и *не должны использоваться для численного решения систем вследствие их очевидной неэффективности*.

⁵⁴Леопольд КРОНЕКЕР (L. Kronecker, 1823-1891) – немецкий математик, член Берлинской АН и Петербургской АН.

Глава 4. ТРЕУГОЛЬНОЕ РАЗЛОЖЕНИЕ КВАДРАТНОЙ МАТРИЦЫ

4.1. LU-разложение

При решении различных задач матричной алгебры оказывается полезным представить заданную матрицу в виде произведения нескольких матриц специальной структуры. Один из способов *факторизации* (разложения на множители) матрицы мы сейчас рассмотрим.

Итак, пусть задана квадратная матрица A .

1-й шаг. Найдем ведущий (наибольший по модулю) элемент A . Переставляя (если нужно) строки и столбцы, поместим ведущий элемент в 1-ю строку и в 1-й столбец. Далее обнулим *поддиагональные* элементы 1-го столбца, прибавляя к m -й строке ($m = 2, \dots, n$) первую строку, умноженную на $(-\frac{a_{m1}}{a_{11}})$.

Полученную матрицу обозначим $A^{(1)}$.

2-й шаг. Найдем ведущий элемент в подматрице, получающейся из $A^{(1)}$ вычеркиванием 1-го столбца и 1-й строки. Переставляя (если нужно) строки и столбцы, поместим ведущий элемент во 2-ю строку и во 2-й столбец. Далее обнулим *поддиагональные* элементы 2-го столбца, прибавляя к m -й строке ($m = 3, \dots, n$) вторую строку, умноженную на $(-\frac{a_{m2}}{a_{22}})$.

Полученную матрицу обозначим $A^{(2)}$.

k -й шаг. Рассмотрим в матрице $A^{(k-1)}$, полученной на предыдущем шаге, подматрицу

$$\begin{bmatrix} a_{kk}^{(k-1)} & \dots & a_{kn}^{(k-1)} \\ \dots & \dots & \dots \\ a_{nk}^{(k-1)} & \dots & a_{nn}^{(k-1)} \end{bmatrix}.$$

Найдем ее ведущий элемент. Переставляя (если нужно) строки и столбцы, поместим ведущий элемент в k -ю строку и в k -й столбец. Далее обнулим *поддиагональные* элементы k -го столбца, прибавляя к m -й строке ($m = k + 1, \dots, n$) k -ю строку, умноженную на $(-\frac{a_{mk}}{a_{kk}})$.

Если на каком-то шаге ведущий элемент окажется нулем, *работа алгоритма заканчивается*. Таким образом, после не более, чем $(n - 1)$ шагов мы преобразуем матрицу A в верхнюю треугольную матрицу, которую принято обозначать U (от английского "Upper").

В рассмотренном выше алгоритме использовались два элементарных преобразования матрицы – прибавление к ее строке другой строки, умноженной на число, и перестановки строк (столбцов). Покажем, что эти элементарные преобразования матрицы равносильны умножению ее на матрицы специального вида.

Обозначим $E_{km}(\alpha)$ ($k \neq m$) квадратную матрицу, получаемую из единичной заменой нуля в k -й строке и m -м столбце числом $\alpha \neq 0$.

Пример. Матрица 4-го порядка

$$E_{31}(-2) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Замечание. Если $k > m$, то $E_{km}(\alpha)$ – нижняя треугольная, если $k < m$ – верхняя треугольная.

Нетрудно убедиться, что умножив матрицу A на $E_{km}(\alpha)$ слева, мы прибавим к ее k -й строке m -ю строку, умноженную на α .

Пример.

$$\begin{aligned} E_{31}(-2) \cdot A &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} a_{11} & \dots & a_{14} \\ a_{21} & \dots & a_{24} \\ a_{31} & \dots & a_{34} \\ a_{41} & \dots & a_{44} \end{bmatrix} = \\ &= \begin{bmatrix} a_{11} & \dots & a_{14} \\ a_{21} & \dots & a_{24} \\ a_{31} - 2a_{11} & \dots & a_{34} - 2a_{14} \\ a_{11} & \dots & a_{14} \end{bmatrix}. \end{aligned}$$

Несложно также видеть (проверьте это!), что

$$E_{km}(\alpha) \cdot E_{km}(-\alpha) = I. \quad (4.1.1)$$

Если бы можно было провести описанный выше процесс преобразования матрицы в верхнюю треугольную без выбора ведущих элементов (т.е. без перестановок строк и столбцов), то с учетом установленных свойств матриц $E_{km}(\alpha)$ можно было бы написать

$$E_{n,n-1} \left(-\frac{a_{n,n-1}^{(n-2)}}{a_{n-1,n-1}^{(n-2)}} \right) \cdot \dots \cdot E_{31} \left(-\frac{a_{31}}{a_{11}} \right) \cdot E_{21} \left(-\frac{a_{21}}{a_{11}} \right) \cdot A = U. \quad (4.1.2)$$

Здесь U – верхняя треугольная матрица, а количество сомножителей слева от A равно количеству обнуляемых поддиагональных элементов матрицы n -го порядка, т.е. $\frac{n(n-1)}{2}$.

Выразим матрицу A из (4.1.2), используя тождество (4.1.1):

$$A = E_{21} \left(\frac{a_{21}}{a_{11}} \right) \cdot E_{31} \left(\frac{a_{31}}{a_{11}} \right) \cdots \cdot E_{n,n-1} \left(\frac{a_{n,n-1}^{(n-2)}}{a_{n-1,n-1}^{(n-2)}} \right) \cdot U. \quad (4.1.3)$$

Нетрудно убедиться (проверьте это!), что произведение нижних треугольных матриц есть нижняя треугольная матрица. Если же на диагоналях сомножителей стоят единицы, то и произведение имеет единичную диагональ.

Таким образом, слева от матрицы U в (4.1.3) стоит нижняя треугольная матрица с единичной диагональю. Обозначив ее L (от английского "Lower"), перепишем (4.1.3) в виде

$$A = LU.$$

Однако в реальном алгоритме присутствуют еще перестановки строк и столбцов матрицы. Введем *матрицу элементарных перестановок* Π_{km} , получающуюся из единичной перемещением единицы, стоящей в k -й строке в m -й столбец, а единицы, стоящей в m -й строке, – в k -й столбец.

Пример.

$$\begin{aligned} \Pi_{13} \cdot A &= \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} a & b & c & d \\ * & * & * & * \\ p & q & r & s \\ * & * & * & * \end{bmatrix} = \begin{bmatrix} p & q & r & s \\ * & * & * & * \\ a & b & c & d \\ * & * & * & * \end{bmatrix} \\ A \cdot \Pi_{13} &= \begin{bmatrix} a & * & p & * \\ b & * & q & * \\ c & * & r & * \\ d & * & s & * \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} p & * & a & * \\ q & * & b & * \\ r & * & c & * \\ s & * & d & * \end{bmatrix}. \end{aligned}$$

Здесь звездочками обозначены элементы матриц, не меняющие своего положения. Видно, что умножая матрицу A на Π_{km} слева, мы переставляем в $A k$ -ю и m -ю строки, а умножая ее на Π_{km} справа, переставляем k -й и m -й столбцы.

Отметим свойства матрицы элементарных перестановок.

1. В каждой ее строке и каждом столбце ровно одна единица; остальные элементы – нули.

2. $\Pi_{km} \cdot \Pi_{km} = I$ (дважды поменяв местами одну и ту же пару столбцов или строк, получаем исходную матрицу).

3. $\det(\Pi_{km}) = -1$ (перестановка двух строк или двух столбцов матрицы приводит к умножению ее определителя на (-1)).

Произведение нескольких матриц элементарных перестановок есть матрица, меняющая местами уже несколько пар столбцов (строк). Такая матрица называется *матрицей перестановок* и обладает свойством 1. Ее определитель равен либо $(+1)$, либо (-1) . Матрица, обратная матрице перестановок, также есть матрица перестановок.

Теперь можно сформулировать следующую теорему.

Теорема. Для каждой квадратной матрицы A существуют такие матрицы перестановок Π_1 и Π_2 , что

$$\Pi_1 \cdot A \cdot \Pi_2 = L \cdot U, \quad (4.1.4)$$

где L – нижняя треугольная матрица с единичной диагональю, а U – верхняя треугольная матрица.

Формула (4.1.4) называется *LU-разложением* или *треугольным разложением* матрицы A .

Доказательство мы опускаем из-за его технической сложности. Отметим, что Π_1 и Π_2 (а, следовательно, L и U) определены не единственным образом. Это следует, например, из того, что при выборе ведущего элемента может встретиться несколько равных по модулю наибольших элементов и, отдав предпочтение одному из них, мы получим одно из возможных *LU*-разложений матрицы.

Замечание. Иногда вместо *LU*-разложения используют так называемое *LDU*-разложение матрицы. Если A – невырожденная матрица, то в формуле (4.1.3) U – также невырожденная матрица. Положим $D = \text{diag}[u_{11}, \dots, u_{nn}]$ и $\tilde{U} = D^{-1}U$. Тогда \tilde{U} – верхняя треугольная матрица с единичной диагональю, и справедливо соотношение

$$\Pi_1 \cdot A \cdot \Pi_2 = L D \tilde{U},$$

которое и называется *LDU*-разложением матрицы A .

4.2. Некоторые применения *LU*-разложения

1. Вычисление определителей. Используя формулу (4.1.4) и свойства определителей, мы можем написать:

$$\begin{aligned}
\det(A) &= \det(\Pi_1^{-1} \cdot L \cdot U \cdot \Pi_2^{-1}) = \\
&= \det(\Pi_1^{-1}) \cdot \det(L) \cdot \det(U) \cdot \det(\Pi_2^{-1}) = \\
&= (\pm 1) \cdot u_{11} \cdot \dots \cdot u_{nn} \quad (4.2.1)
\end{aligned}$$

(согласно п.3.1., $\det(L) = 1$, $\det(U) = u_{11} \cdot \dots \cdot u_{nn}$, а знак в правой части зависит от количества элементарных перестановок при выполнении LU -разложения матрицы).

В п.3.1 было показано, что вычисление определителя матрицы "по определению" невозможно, поскольку требуемое для его реализации количество операций превышает $n!$ (n – порядок матрицы). *Можно показать, что LU -разложение требует порядка $\frac{n^3}{3}$ операций.* При $n = 30$ это составит около $3 \cdot 10^4$ (сравните с полученной ранее оценкой $3 \cdot 10^{23}$ для вычисления "по определению"). После получения LU -разложения определитель вычисляется по формуле (4.2.1) элементарно.

2. Решение систем линейных уравнений. Если получено LU -разложение матрицы A , то решение системы $Ax = b$ сводится к последовательному решению двух систем с треугольными матрицами:

$$Ly = b, \quad Ux = y. \quad (4.2.2)$$

Решение системы с треугольной матрицей коэффициентов требует выполнения порядка n^2 операций (против n^3 для системы с заполненной матрицей коэффициентов). Правда, затраты времени на LU -разложение сравнимы с временем решения системы с заполненной матрицей и, на первый взгляд, выгоды не видно. Однако в приложениях весьма часто приходится неоднократно решать системы с одной и той же матрицей и различными правыми частями. В этом случае факторизация производится один раз, а затем каждый раз решаются системы (4.2.2), что при матрицах большого порядка дает весьма значительный выигрыш во времени.

5. ЛИНЕЙНОЕ ПРОСТРАНСТВО

5.1. Основные понятия

Рассмотрим множество всех матриц-столбцов высоты n с комплексными элементами. Матрицы-столбцы будем обозначать в этом пункте латинскими буквами, а числа – греческими. Столбец с нулевыми элементами будем обозначать θ (или θ_n , если необходимо указать его высоту). Положим по определению $-x = (-1) \cdot x$ и будем называть столбец $-x$ *противоположным* столбцу x .

Известно, что матрицы-столбцы можно складывать и умножать на числа. Перечислим свойства этих операций.

- $$\begin{aligned} 1) x+y &= y+x, & 2) (x+y)+z &= (x+z)+y, & 3) x+\theta &= x, & 4) x+(-x) &= \theta; \\ 5) 1 \cdot x &= x, & 6) (\alpha + \beta) \cdot x &= \alpha x + \beta x, & 7) (\alpha\beta)x &= \alpha(\beta x); \\ 8) \alpha \cdot (x+y) &= \alpha x + \alpha y. \end{aligned}$$

Определение. Множество всех комплексных матриц-столбцов высоты n с введенными выше операциями сложения и умножения на число называется *линейным (векторным) пространством* и обозначается \mathbb{C}^n . Его элементы – $(n \times 1)$ -матрицы называются *векторами*.

Замечания. 1. Линейное пространство \mathbb{C}^n называют также *комплексным линейным пространством* в отличие от *вещественного линейного пространства* \mathbb{R}^n – множества всех *вещественных* матриц-столбцов высоты n , в котором разрешено умножение *только на вещественные* числа.

2. Линейное пространство \mathbb{R}^3 имеет очевидную геометрическую интерпретацию: каждому его вектору $x = [x_1, x_2, x_3]^T$ можно сопоставить направленный отрезок, начало которого совмещено с началом координат, а конец расположен в точке с координатами x_1, x_2, x_3 (рис.5.1).

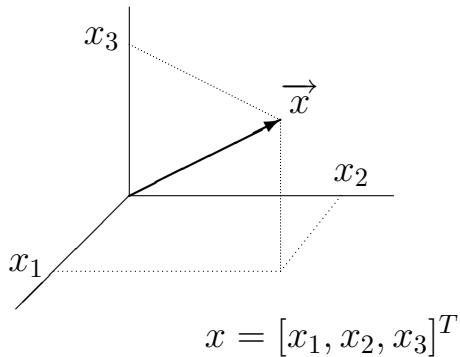


Рис.5.1

Аналогично, геометрическую интерпретацию линейного пространства \mathbb{R}^2 дают направленные отрезки на плоскости.

Будем обозначать направленный отрезок той же буквой, что и соответствующий ему вектор, но со стрелкой сверху. Нетрудно видеть, что

$$\overrightarrow{x+y} = \overrightarrow{x} + \overrightarrow{y}; \quad \overrightarrow{\alpha x} = \alpha \cdot \overrightarrow{x}.$$

5.2. Абстрактное линейное пространство

Может показаться, что уделено излишнее внимание перечислению известных свойств линейных операций над матрицами. Дело, однако, в том, что линейное пространство – одно из основных понятий современной математики.

Линейным пространством называют любое непустое множество, над элементами которого (векторами) можно производить две операции, именуемые *сложение* и *умножение на число*.

Сложение. Каждой паре элементов пространства (x, y) (векторов) ставится в соответствие вектор, называемый их суммой и обозначаемый $x + y$. Эта операция должна удовлетворять следующим правилам:

1. $x + y = y + x$.
2. $(x + y) + z = x + (y + z)$.
3. Существует такой вектор θ , именуемый *нулевым*, что $x + \theta = x$ для любого вектора x .
4. Для каждого вектора x есть такой вектор $(-x)$, именуемый *противоположным* к x , что $x + (-x) = \theta$.

Умножение на число. Каждому вектору x и каждому числу α ставится в соответствие вектор αx , называемый их произведением. При этом

5. $1 \cdot x = x$.
6. $(\alpha + \beta) \cdot x = \alpha \cdot x + \beta \cdot x$.
7. $(\alpha\beta) \cdot x = \alpha \cdot (\beta \cdot x)$.
8. $\alpha \cdot (x + y) = \alpha \cdot x + \alpha \cdot y$.

Перечисленные восемь свойств называют *аксиомами линейного пространства*. Подчеркнем еще раз, что при построении теории не существенно, что представляют собой элементы линейного пространства –

векторы (лишь бы их можно было "складывать" умножать на числа и выполнялись аксиомы). Тогда будут справедливы все выводы построенной ниже теории. В нашем курсе мы рассматриваем, в основном, простейшие частные случаи – пространства \mathbb{C}^n и \mathbb{R}^n . Один пример линейного пространства, отличного от \mathbb{C}^n и \mathbb{R}^n , будет рассмотрен в п.5.5.

5.3. Линейная зависимость векторов

Пусть заданы векторы $a^{(1)}, \dots, a^{(k)} \in \mathbb{C}^n$. Составим уравнение

$$\alpha_1 a^{(1)} + \dots + \alpha_k a^{(k)} = \theta, \quad (5.3.1)$$

в котором искомыми являются числа $\alpha_1, \dots, \alpha_k$.

Уравнение (5.3.1) может быть переписано в виде $A\alpha = \theta_n$, где $A = [a^{(1)}, \dots, a^{(k)}]$ – заданная $(n \times k)$ -матрица, $\alpha = [\alpha_1, \dots, \alpha_k]^T$ – искомый вектор-столбец, т.е. оно представляет собой *однородную* систему линейных уравнений.

Очевидно, что при любых заданных векторах эта система имеет *нулевое* решение $\alpha_1 = \alpha_2 = \dots = \alpha_k = 0$ ($\alpha = \theta_k$).

Может оказаться, что нулевое решение *единственно*. Так, например, если $a^{(1)} = [1 \ 0 \ 0]^T$, $a^{(2)} = [0 \ 1 \ 0]^T$, то уравнение

$$\alpha_1 a^{(1)} + \alpha_2 a^{(2)} = \theta_3 \iff [\alpha_1 \ \alpha_2 \ 0]^T = [0 \ 0 \ 0]^T$$

имеет *только нулевое* решение $\alpha_1 = \alpha_2 = 0$. В то же время известно, что однородная система линейных уравнений может иметь решения, отличные от нулевого (например, если число уравнений меньше, чем число переменных).

Определение. Если уравнение (5.3.1) имеет *только нулевое* решение, то множество векторов $\{a^{(1)}, \dots, a^{(k)}\}$ называется *линейно независимым*. Если же это уравнение имеет *ненулевые* решения, то упомянутое множество векторов называется *линейно зависимым*.

Докажем теперь несколько утверждений, связанных с понятием линейной зависимости множества векторов.

1. Множество, состоящее из одного *ненулевого* вектора, линейно независимо.

Доказательство. $a \neq \theta \wedge aa = \theta \implies \alpha = 0$. ■

2. Множество, содержащее нулевой вектор, линейно зависимо.

Доказательство. Уравнение $\alpha_1 a^{(1)} + \dots + \alpha_k a^{(k)} + \alpha\theta = \theta$ имеет очевидное ненулевое решение $\alpha_1 = \dots = \alpha_k = 0, \alpha = 1$.

Для дальнейшего нам необходимо ввести важное новое понятие.

Определение. Вектор $a = \alpha_1 a^{(1)} + \dots + \alpha_k a^{(k)}$ называется *линейной комбинацией векторов* $a^{(1)}, \dots, a^{(k)}$.

3. Если множество векторов линейно зависимо, то хотя бы один из них есть линейная комбинация остальных.

Доказательство. Пусть $\alpha_1, \dots, \alpha_k$ – ненулевое решение уравнения (5.3.1), существующее в силу линейной зависимости множества $\{a^{(1)}, \dots, a^{(k)}\}$. Пусть для определенности $\alpha_m \neq 0$. Перенося в (5.3.1) слагаемое $\alpha_m a^{(m)}$ в правую часть и деля обе части на $-\alpha_m$, получим

$$\begin{aligned} a^{(m)} &= \left(-\frac{\alpha_1}{\alpha_m}\right)a^{(1)} + \dots + \left(-\frac{\alpha_{m-1}}{\alpha_m}\right)a^{(m-1)} + \\ &\quad + \left(-\frac{\alpha_{m+1}}{\alpha_m}\right)a^{(m+1)} + \dots + \left(-\frac{\alpha_k}{\alpha_m}\right)a^{(k)}. \end{aligned}$$

■

4. Если хотя бы один из векторов множества есть линейная комбинация остальных, то множество линейно зависимо.

Доказательство. Пусть, например,

$$a^{(m)} = \alpha_1 a^{(1)} + \dots + \alpha_{m-1} a^{(m-1)} + \alpha_{m+1} a^{(m+1)} + \dots + \alpha_k a^{(k)}.$$

Перенося все слагаемые в правую часть, получим

$$\theta = \alpha_1 a^{(1)} + \dots + \alpha_{m-1} a^{(m-1)} + (-1)a^{(m)} + \alpha_{m+1} a^{(m+1)} + \dots + \alpha_k a^{(k)}.$$

Мы построили *ненулевое* ($\alpha_m = -1$) решение уравнения (5.3.1), т.е. множество $\{a^{(1)}, \dots, a^{(k)}\}$ линейно зависимо. ■

Замечание. Доказанные выше утверждения 3 и 4 часто формулируют так: *линейная зависимость множества векторов равносильна возможности представления одного из них в виде линейной комбинации остальных*.

5. Множество векторов, содержащее линейно зависимую часть, линейно зависимо.

Доказательство. Пусть множество $a^{(1)}, \dots, a^{(k)}$ содержит k векторов, и его часть $a^{(1)}, \dots, a^{(m)}$ ($m < k$) линейно зависима, т.е. уравнение $\beta_1 a^{(1)} + \dots + \beta_m a^{(m)} = \theta$ имеет ненулевое решение (среди чисел β_1, \dots, β_m есть отличные от нуля). Тогда и уравнение

$$\alpha_1 a^{(1)} + \dots + \alpha_m a^{(m)} + \alpha_{m+1} a^{(m+1)} + \dots + \alpha_k a^{(k)} = \theta$$

имеет очевидное ненулевое решение

$$\alpha_1 = \beta_1, \dots, \alpha_m = \beta_m; \quad \alpha_{m+1} = \dots = \alpha_k = 0. \quad \blacksquare$$

6. Всякая непустая часть линейно независимого множества линейно независима.

Доказательство. От противного.

Замечания. 1. Если два ненулевых вектора в $\mathbb{R}^2(\mathbb{R}^3)$ линейно зависимы, то соответствующие им направленные отрезки *коллинеарны* (лежат на одной прямой).

2. Если три ненулевых вектора в \mathbb{R}^3 линейно зависимы, то соответствующие им направленные отрезки *компланарны* (лежат в одной плоскости).

5.4. Размерность линейного пространства и его базис

Определение. Если в линейном пространстве P существует линейно независимое множество из n векторов, а *всякое* множество, содержащее более, чем n векторов, линейно зависимо, то говорят, что пространство имеет размерность n и пишут⁵⁵:

$$\dim(P) = n.$$

Пример. Докажем, что $\dim(\mathbb{C}^n) = n$.

Доказательство. Множество, состоящее из n векторов

$$e^{(1)} = [1, 0, \dots, 0]^T, e^{(2)} = [0, 1, \dots, 0]^T, \dots, e^{(n)} = [0, 0, \dots, 1]^T,$$

⁵⁵dimension (англ.) – размерность.

линейно независимо, так как уравнение $\alpha_1 e^{(1)} + \dots + \alpha_n e^{(n)} = \theta$, или

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

или, наконец, $I_n \alpha = \theta_n$, имеет единственное решение $\alpha_1 = \dots = \alpha_n = 0$.

В то же время любая часть \mathbb{C}^n , содержащая больше, чем n векторов, линейно зависима, так как при $m > n$ уравнение $\alpha_1 a^{(1)} + \dots + \alpha_m a^{(m)} = \theta_n$ имеет ненулевое решение как однородная система, в которой количество переменных (m) превышает количество уравнений (n). ■

Замечание. Из нашего рассуждения следует, что и $\dim(\mathbb{R}^n) = n$.

Если не ограничиваться линейными пространствами \mathbb{C}^n и \mathbb{R}^n , то нельзя исключить случай, когда существуют линейно независимые множества, содержащие как угодно много векторов. Поэтому мы введем

Определение. Если для любого натурального числа n в линейном пространстве существует линейно независимое множество, состоящее из n векторов, то линейное пространство называется *бесконечномерным*.

Для *конечномерных* линейных пространств введем понятие базиса.

Определение. *Базисом* n -мерного линейного пространства называется любой упорядоченный линейно независимый набор из n векторов этого пространства.

Пример. Векторы $e^{(1)}, \dots, e^{(n)}$, введенные выше, образуют базис в \mathbb{C}^n и в \mathbb{R}^n . Его называют *стандартным* базисом.

Роль базиса в конечномерном пространстве определяет

Теорема. Всякий вектор конечномерного пространства может быть представлен в виде линейной комбинации векторов базиса, и такое представление единственно.

Доказательство. Пусть $a^{(1)}, \dots, a^{(n)}$ – базис, и b – произвольный вектор. По определению размерности пространства множество $b, a^{(1)}, \dots, a^{(n)}$, содержащее более, чем n векторов, линейно зависимо. Поэтому уравнение $ab + \alpha_1 a^{(1)} + \dots + \alpha_n a^{(n)} = \theta$ имеет ненулевое решение. В частности, $\alpha \neq 0$, ибо иначе имело бы ненулевое решение уравнение $\alpha_1 a^{(1)} + \dots + \alpha_n a^{(n)} = \theta$, и базис оказался бы линейно зависимым множеством! А если $\alpha \neq 0$, то

$$b = \left(-\frac{\alpha_1}{\alpha}\right) a^{(1)} + \dots + \left(-\frac{\alpha_n}{\alpha}\right) a^{(n)},$$

и возможность представления произвольного вектора в виде линейной комбинации базисных векторов доказана.

Докажем единственность такого представления. Пусть имеются два представления вектора в виде линейной комбинации базисных векторов: $b = \beta_1 a^{(1)} + \dots + \beta_n a^{(n)}$ и $b = \gamma_1 a^{(1)} + \dots + \gamma_n a^{(n)}$. Вычитая второе равенство из первого, получаем

$$\theta = (\beta_1 - \gamma_1) a^{(1)} + \dots + (\beta_n - \gamma_n) a^{(n)}.$$

Отсюда, вследствие линейной независимости базиса, все коэффициенты при его векторах – нули, т.е. $\beta_1 = \gamma_1, \dots, \beta_n = \gamma_n$.

Определение. Представление вектора в виде линейной комбинации векторов базиса называют *разложением вектора по базису*, а коэффициенты этого разложения – *координатами* этого вектора в этом базисе.

Замечания. 1. Отметим существенность фиксации порядка векторов в базисе (упорядоченности базиса). Изменив порядок векторов в базисе, получим, конечно, опять базис, но другой!

2. Обратите внимание на то, что разложение вектора b по базису $a^{(1)}, \dots, a^{(n)}$, т.е. решение уравнения

$$\alpha_1 a^{(1)} + \dots + \alpha_n a^{(n)} = b \quad (5.4.1)$$

сводится в \mathbb{C}^n и в \mathbb{R}^n к решению матричного уравнения

$$A\alpha = b, \quad (5.4.2)$$

где, как всегда, $A = [a^{(1)}, \dots, a^{(n)}]$, $\alpha = [\alpha_1, \dots, \alpha_n]^T$.

Таким образом, (5.4.1) и (5.4.2) представляют собой попросту две различные формы записи одного и того же уравнения – "векторную" и "матричную".

Следствия. 1. Если столбцы квадратной матрицы линейно независимы (т.е. составляют базис \mathbb{C}^n (\mathbb{R}^n)), то система линейных уравнений (5.4.2) имеет единственное решение при любом столбце свободных членов.

2. Поскольку единственность решения линейной системы с квадратной матрицей равносильна (как уже известно) невырожденности этой

матрицы, то линейная независимость столбцов квадратной матрицы также равносильна ее невырожденности. Итак, если A – квадратная матрица порядка n , то равносильны следующие три утверждения:

- 1) $\det(A) \neq 0$;
- 2) столбцы матрицы A образуют базис \mathbb{C}^n (\mathbb{R}^n);
- 3) система линейных уравнений $Ax = b$ при любом столбце b имеет единственное решение.

5.5. Пространство полиномов порядка n

Как известно, полиномом *степени* $k \geq 0$ называется функция, действующая из \mathbb{C} в \mathbb{C} по правилу

$$p(z) = p_1 + p_2 z + \dots + p_{k+1} z^k, \quad (5.5.1)$$

где p_1, \dots, p_{k+1} – заданные комплексные числа, называемые *коэффициентами полинома*, причем $p_{k+1} \neq 0$.

Замечание. Если коэффициенты полинома – вещественные числа, то (5.5.1) определяет также функцию, действующую из \mathbb{R} в \mathbb{R} .

Определение. Назовем линейным пространством полиномов *порядка* n множество, содержащее все полиномы, *степень* которых *строго меньше* n , и функцию $\theta(z)$, тождественно равную нулю (напомним, что эта функция называется *нуль-полиномом*; степень нуль-полинома не определена). Иначе говоря, полиномом порядка n будем называть функцию $p(z) = p_1 + p_2 z + \dots + p_n z^{n-1}$, где p_1, \dots, p_n – произвольный набор комплексных чисел (в том числе он может состоять из одних нулей).

Линейное пространство полиномов порядка n будем обозначать \mathbb{P}_n . В этом пространстве естественным образом определены сумма полиномов и произведение полинома на число:

$$(p + q)(z) = (p_1 + q_1) + (p_2 + q_2)z + \dots + (p_n + q_n)z^{n-1};$$

$$(\alpha p)(z) = (\alpha p_1) + (\alpha p_2)z + \dots + (\alpha p_n)z^{n-1}.$$

Замечание. Обратите внимание на то, что *степень* полинома не обязана сохраняться при сложении и умножении на число. Например, складывая два полинома второй степени $p(z) = z^2 + 1$ и $q(z) = -z^2$, получаем полином нулевой степени $(z^2 + 1) + (-z^2) = 1$. Поскольку для полинома $p(z)$ любой степени $0 \cdot p(z) = \theta(z) \equiv 0$, степень произведения полинома на число может быть и вовсе не определена. Поэтому *порядок* полинома является для наших целей более удобной характеристикой, чем *степень*.

Предлагаем читателю убедиться самостоятельно, что множество \mathbb{P}_n с введенными в нем операциями сложения и умножения на число, "авансом" названное нами *линейным пространством*, действительно является таковым (т.е. удовлетворяет аксиомам, приведенным в п.5.2). Роль нулевого элемента играет при этом нуль-полином, а роль противоположного полиному p элемента – полином

$$(-p)(z) = (-p_1) + (-p_2)z + \dots + (-p_n)z^{n-1}.$$

Теорема. Размерность пространства \mathbb{P}_n равна n .

Доказательство. Рассмотрим множество, состоящее из n полиномов

$$e^{(1)}(z) \equiv 1, \quad e^{(2)}(z) = z, \dots, \quad e^{(n)}(z) = z^{n-1}.$$

Все эти полиномы, очевидно, принадлежат \mathbb{P}_n . Составим уравнение

$$\alpha_1 e^{(1)}(z) + \dots + \alpha_n e^{(n)}(z) = \theta(z) \equiv 0. \quad (5.5.2)$$

Зафиксируем n попарно различных чисел z_1, \dots, z_n и, полагая в (5.5.2) $z = z_k$, $k = 1, \dots, n$, получим систему линейных уравнений

$$\begin{cases} \alpha_1 + \alpha_2 z_1 + \dots + \alpha_n z_1^{n-1} = 0 \\ \dots \dots \dots \dots \dots \dots \\ \alpha_1 + \alpha_2 z_n + \dots + \alpha_n z_n^{n-1} = 0 \end{cases}$$

или, в матричной записи, $E\alpha = \theta$, где E – матрица Вандермонда.

Поскольку числа z_1, \dots, z_n попарно различны, $\det(E) \neq 0$ (см. п.3.2). Поэтому уравнение (5.5.2) имеет только нулевое решение, и, значит, рассматриваемый набор полиномов линейно независим.

С другой стороны, если множество

$$\begin{aligned} p^{(1)}(z) &= p_{11} + p_{21}z + \dots + p_{n1}z^{n-1}, \\ &\dots \dots \dots \dots \dots \dots \\ p^{(k)}(z) &= p_{1k} + p_{2k}z + \dots + p_{nk}z^{n-1} \end{aligned}$$

содержит более, чем n полиномов ($k > n$), то, приравнивая коэффициенты при одинаковых степенях z в левой и правой частях уравнения

$$\alpha_1 p^{(1)}(z) + \dots + \alpha_k p^{(k)}(z) = \theta(z),$$

получим однородную систему из n уравнений с k переменными ($k > n$), которая имеет ненулевые решения. Следовательно, всякое множество, содержащее более, чем n полиномов из \mathbb{P}^n , линейно зависимо.

Мы доказали, что $\dim(\mathbb{P}_n) = n$, и показали, что упорядоченный набор полиномов $e^{(1)}(z) \equiv 1, e^{(2)}(z) = z, \dots, e^{(n)}(z) = z^{(n-1)}$ образует базис \mathbb{P}_n . Этот базис называют стандартным.

Замечания. 1. Разложение полинома порядка n по стандартному базису очевидно совпадает с его записью по возрастающим степеням переменной.

2. Из-за специфики конкретных задач могут оказаться более удобными разложения полинома по другим базисам \mathbb{P}_n . С примерами таких задач мы познакомимся в п.п. 5.6 и 11.3. Здесь следует лишь отметить, что разложения полинома в разных базисах суть тождественные преобразования этого полинома.

5.6. Полиномиальная интерполяция

Рассмотрим задачу, часто встречающуюся в приложениях: построить полином по его значениям в заданных точках.

Такая задача называется задачей *полиномиальной интерполяции*. Ее решение базируется на следующей теореме.

Теорема. Существует полином порядка n , который в заданных n точках принимает заданные значения, и такой полином один.

Доказательство. Пусть z_1, \dots, z_n – заданные попарно различные числа. Положим

$$\begin{aligned}
l^{(1)}(z) &= \frac{(z - z_2) \cdot (z - z_3) \cdot \dots \cdot (z - z_n)}{(z_1 - z_2) \cdot (z_1 - z_3) \cdot \dots \cdot (z_1 - z_n)}; \\
l^{(2)}(z) &= \frac{(z - z_1) \cdot (z - z_3) \cdot \dots \cdot (z - z_n)}{(z_2 - z_1) \cdot (z_2 - z_3) \cdot \dots \cdot (z_2 - z_n)}; \\
&\dots \\
l^{(n)}(z) &= \frac{(z - z_1) \cdot (z - z_2) \cdot \dots \cdot (z - z_{n-1})}{(z_n - z_1) \cdot (z_n - z_2) \cdot \dots \cdot (z_n - z_{n-1})}.
\end{aligned} \tag{5.6.1}$$

Очевидно, что степень каждого из этих полиномов равна $n - 1$ и, следовательно, они принадлежат \mathbb{P}_n . Далее, легко видеть, что

$$l^{(m)}(z_k) = \delta_{mk}, \quad k, m = 1, \dots, n \quad (5.6.2)$$

(напомним, что δ_{mk} – символ Кронекера).

Составим уравнение

$$\alpha_1 l^{(1)}(z) + \dots + \alpha_n l^{(n)}(z) = \theta(z) \equiv 0 \quad (5.6.3)$$

и положим в нем поочередно $z = z_1, \dots, z = z_n$. Используя (5.6.2), находим, что уравнение (5.6.3) имеет *только нулевое* решение, т.е. полиномы (5.6.1) линейно независимы и образуют базис \mathbb{P}_n .

Построим теперь полином L порядка n , принимающий в n заданных точках заданные значения:

$$L(z_k) = y_k, \quad k = 1, \dots, n. \quad (5.6.4)$$

Искать его будем в виде разложения по построенному базису (5.6.1)

$$L(z) = \sum_{m=1}^n \alpha_m l^{(m)}(z).$$

Выписывая условия (5.6.4), получим систему уравнений для определения коэффициентов разложения

$$L(z_k) = \sum_{m=1}^n \alpha_m l^{(m)}(z_k) = y_k, \quad k = 1, \dots, n.$$

Подставив сюда (5.6.2), получим

$$\sum_{m=1}^n \alpha_m l^{(m)}(z_k) = \sum_{m=1}^n \alpha_m \delta_{mk} = \alpha_k = y_k, \quad k = 1, \dots, n. \quad \blacksquare$$

Мы не только доказали теорему, но и получили явное выражение для единственного полинома порядка n , принимающего в n заданных точках заданные значения:

$$L(z) = \sum_{m=1}^n y_m l^{(m)}(z).$$

Его называют *интерполяционным полиномом в форме Лагранжа*.

Замечание. Существенно, что в теореме фиксируется *порядок* интерполяционного полинома, а не его *степень*. Взяв, например, $y_k \equiv 1$ при $k = 1, \dots, n$, получим для *любого* n $L(z) \equiv 1$ – полином *нулевой* степени.

Пример. Построим полином третьего порядка, соответствующий таблице

z	1	2	3	
y	4	5	6	

$$L(z) = 4 \cdot \frac{(z-2)(z-3)}{(1-2)(1-3)} + 5 \cdot \frac{(z-1)(z-3)}{(2-1)(2-3)} + 6 \cdot \frac{(z-1)(z-2)}{(3-1)(3-2)} = z + 3.$$

Глава 6. СОБСТВЕННЫЕ ЧИСЛА И СОБСТВЕННЫЕ ВЕКТОРЫ КВАДРАТНОЙ МАТРИЦЫ

6.1. Основные понятия

Пусть A – матрица размера $(m \times n)$. Умножение вектора из \mathbb{C}^n на эту матрицу слева дает вектор из \mathbb{C}^m . Таким образом, можно сказать, что $(m \times n)$ -матрица A порождает отображение линейного пространства \mathbb{C}^n в линейное пространство \mathbb{C}^m :

$$x \longrightarrow Ax.$$

Отметим два важных свойства этого отображения.

1. Образ суммы двух векторов есть сумма их образов:

$$A(x^{(1)} + x^{(2)}) = Ax^{(1)} + Ax^{(2)}.$$

2. При умножении вектора на число его образ умножается на то же число:

$$A(\alpha x) = \alpha(Ax).$$

Свойства **1** и **2** можно объединить и переформулировать так: для любых векторов $x^{(1)}, x^{(2)}$ и для любых чисел α_1, α_2

$$A(\alpha_1 x^{(1)} + \alpha_2 x^{(2)}) = \alpha_1 Ax^{(1)} + \alpha_2 Ax^{(2)}.$$

Отображения, обладающие этим свойством, называются *линейными* (сравните со свойством **3** определителя матрицы, п.3.2). Итак, матрица порождает *линейное отображение* \mathbb{C}^n в \mathbb{C}^m (*линейный оператор*, действующий из \mathbb{C}^n в \mathbb{C}^m).

Заметим, что если A – $(m \times n)$ -матрица с *вещественными* элементами и $x \in \mathbb{R}^n$, то $Ax \in \mathbb{R}^m$, т.е. вещественная матрица порождает еще и линейный оператор, действующий из \mathbb{R}^n в \mathbb{R}^m .

Если A – квадратная матрица, то она порождает отображение линейного пространства в себя. Рассмотрим пример такого отображения. Пусть $n = 2$,

$$A = \begin{bmatrix} 5 & -6 \\ 3 & -4 \end{bmatrix}, \quad y = \begin{bmatrix} -6 \\ -4 \end{bmatrix}, \quad x = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

Тогда

$$Ay = \begin{bmatrix} 5 & -6 \\ 3 & -4 \end{bmatrix} \cdot \begin{bmatrix} -6 \\ -4 \end{bmatrix} = \begin{bmatrix} -6 \\ -2 \end{bmatrix}, \quad Ax = \begin{bmatrix} 5 & -6 \\ 3 & -4 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} = 2 \cdot x.$$

Видно (рис.6.1), что умножение вектора y на матрицу A не только "растягивает" соответствующий этому вектору направленный отрезок \overrightarrow{y} , но и "поворачивает" его, в то время как направленные отрезки \overrightarrow{x} и \overrightarrow{Ax} коллинеарны.

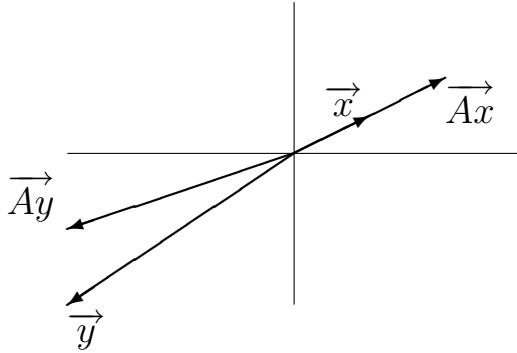


Рис.6.1

Определение. Пусть A – квадратная $(n \times n)$ -матрица. Если для некоторого комплексного числа λ и некоторого *ненулевого* вектора $x \in \mathbb{C}^n$ выполняется равенство

$$Ax = \lambda x, \quad (6.1.1)$$

то λ называется *собственным числом* (или *собственным значением*) матрицы A , а x – *собственным вектором* этой матрицы, *соответствующим собственному числу* λ .

Замечания. 1. В случае, когда речь идет о собственных числах нескольких матриц, например, A и B , целесообразно использовать обозначения $\lambda(A)$ и $\lambda(B)$.

2. Условие $x \neq \theta$ для собственного вектора существенно, так как $A\theta = \lambda\theta$ при *любом* λ , и этот случай не представляет интереса.

6.2. Полная проблема собственных значений

Полной проблемой собственных значений называют задачу о нахождении всех собственных чисел и собственных векторов квадратной матрицы. Эта задача наряду с задачей о решении системы линейных уравнений составляет основное содержание линейной алгебры.

Преобразуем уравнение (6.1.1), определяющее собственные числа и соответствующие им собственные векторы.

$$Ax = \lambda x \iff Ax - \lambda x = \theta \iff (A - \lambda I)x = \theta. \quad (6.2.1)$$

В (6.2.1) n уравнений (*нелинейных!*) с $(n + 1)$ переменными: x_1, \dots, x_n и еще λ .

Заметим, однако, что при фиксированном λ эта система становится линейной и однородной, и, следовательно, существование у нее ненулевых решений равносильно вырожденности матрицы ее коэффициентов. Итак, собственные числа матрицы A – это в точности корни уравнения

$$\det(A - \lambda I) = 0. \quad (6.2.2)$$

Исследуем это уравнение. Как известно, определитель матрицы вычисляется через ее элементы с помощью операций умножения и сложения. С другой стороны, элементы матрицы $A - \lambda I$ – это полиномы относительно λ . Следовательно, $\det(A - \lambda I)$ – тоже полином относительно λ . Он называется *характеристическим полиномом матрицы A*, и мы будем обозначать его $P_A(\lambda)$.

Примеры. 1. $n = 1$, $A = [a_{11}]$; $P_A(\lambda) = \det[a_{11} - \lambda] = a_{11} - \lambda$.

$$2. \quad n = 2, \quad A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}; \quad P_A(\lambda) = \det \begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} = \\ = \det(A) - Sp(A) \cdot \lambda + \lambda^2$$

(здесь $Sp(A)$ – *след матрицы* – сумма ее диагональных элементов).

$$3. \quad n = 3, \quad A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}; \quad P_A(\lambda) = \det \begin{bmatrix} a_{11} - \lambda & a_{12} & a_{13} \\ a_{21} & a_{22} - \lambda & a_{23} \\ a_{31} & a_{32} & a_{33} - \lambda \end{bmatrix} = \\ = \det(A) - (A_{11} + A_{22} + A_{33}) \cdot \lambda + Sp(A) \cdot \lambda^2 - \lambda^3$$

(здесь A_{kk} , $k = 1, 2, 3$ – алгебраические дополнения диагональных элементов матрицы).

Можно показать, что для квадратной матрицы A порядка n характеристический полином имеет степень n , причем старший коэффициент его равен $(-1)^n$, свободный член равен $\det(A)$, а коэффициент при λ^{n-1} равен $(-1)^{n-1} \cdot Sp(A)$.

Напомним (см. п.3.3 раздела "Математический анализ"), что всякий полином степени $n \geq 1$ может быть разложен на множители:

$$p_0 + p_1\lambda + \dots + p_n\lambda^n \equiv p_n \cdot (\lambda - \lambda_1)^{k_1} \cdot \dots \cdot (\lambda - \lambda_m)^{k_m}.$$

Здесь числа $\lambda_1, \dots, \lambda_m$ – попарно различные корни полинома, а натуральные числа k_1, \dots, k_m – их кратности. При этом $k_1 + \dots + k_m =$

n , т.е. полное количество корней полинома (с учетом их кратности) равно степени полинома. Далее, по формулам Виета сумма всех корней полинома (с учетом кратности) равна $-\left(\frac{p_{n-1}}{p_n}\right)$, а произведение равно $(-1)^n \left(\frac{p_0}{p_n}\right)$.

Эти свойства позволяют сформулировать ряд содержательных утверждений о собственных числах матрицы:

1. Каждая квадратная матрица порядка n имеет (с учетом возможной кратности) ровно n собственных чисел.

Замечание. Отметим, что даже у вещественной матрицы собственные числа не обязательно вещественны. Например:

$$B = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}; \quad P_B(\lambda) = \det \begin{bmatrix} -\lambda & 1 \\ -1 & -\lambda \end{bmatrix} = \lambda^2 + 1; \quad \lambda_{1,2} = \pm i.$$

2. Сумма всех (с учетом их кратности) собственных чисел матрицы равна ее следу. Произведение же всех собственных чисел матрицы равно ее определителю.

$$\sum_{r=1}^n \lambda_r(A) = Sp(A); \quad \prod_{r=1}^n \lambda_r(A) = \det(A).$$

Следствие. Вырожденность матрицы равносильна наличию у нее нулевого собственного числа.

Остановимся теперь на вопросе о количестве собственных векторов матрицы. Прежде всего отметим, что умножив собственный вектор на *отличное от нуля* число, получим вновь собственный вектор:

$$\begin{aligned} x \neq \theta \wedge Ax = \lambda x \wedge \alpha \neq 0 &\implies \\ &\implies \alpha x \neq \theta \wedge A(\alpha x) = \alpha(Ax) = \alpha(\lambda x) = \lambda(\alpha x). \end{aligned}$$

Поэтому имеет смысл говорить не о количестве собственных векторов матрицы вообще, а лишь о количестве ее *линейно независимых* собственных векторов.

Имеет место следующая важнейшая

Теорема. Собственные векторы матрицы, соответствующие ее *по-парно различным* собственным числам, линейно независимы.

Доказательство. Пусть $k \leq n$, $\lambda_1, \dots, \lambda_k$ – попарно различные собственные числа $(n \times n)$ -матрицы A , а $x^{(1)}, \dots, x^{(k)}$ – соответствующие им собственные векторы.

Составим уравнение

$$\alpha_1 x^{(1)} + \dots + \alpha_k x^{(k)} = \theta \quad (6.2.3)$$

и покажем, что оно имеет только нулевое решение.

Умножим (6.2.3) слева на матрицу A . По определению собственных векторов получим

$$\alpha_1 \lambda_1 x^{(1)} + \alpha_2 \lambda_2 x^{(2)} + \dots + \alpha_k \lambda_k x^{(k)} = \theta. \quad (6.2.4)$$

С другой стороны, умножая обе части (6.2.3) на λ_1 , имеем

$$\alpha_1 \lambda_1 x^{(1)} + \alpha_2 \lambda_1 x^{(2)} + \dots + \alpha_k \lambda_1 x^{(k)} = \theta.$$

Вычитание полученного равенства из (6.2.4) дает

$$\alpha_2 (\lambda_2 - \lambda_1) x^{(2)} + \dots + \alpha_k (\lambda_k - \lambda_1) x^{(k)} = \theta. \quad (6.2.5)$$

Количество слагаемых в левой части (6.2.5) уменьшилось по сравнению с (6.2.3) на единицу. Умножая (6.2.5) слева на матрицу A , затем на λ_2 и вычитая из первого произведения второе, уменьшим количество слагаемых в левой части еще на единицу. Повторяя этот прием, придем к равенству

$$\alpha_k (\lambda_k - \lambda_1) \cdot (\lambda_k - \lambda_2) \cdot \dots \cdot (\lambda_k - \lambda_{k-1}) x^{(k)} = \theta,$$

из которого, учитывая, что $x^{(k)} \neq \theta$, а собственные числа попарно различны, получим, что $\alpha_k = 0$.

Так как порядок собственных векторов в уравнении (6.2.3) произволен, мы показали, на самом деле, что равны нулю все числа α_r ($r = 1, \dots, k$), и линейная независимость собственных векторов, соответствующих попарно различным собственным числам, доказана. ■

Эта теорема имеет важное

Следствие. Если все n собственных чисел $(n \times n)$ -матрицы A попарно различны, то соответствующие им собственные векторы образуют базис в \mathbb{C}^n . Такой базис принято называть *собственным базисом* матрицы A .

Вопрос о существовании собственного базиса в случае наличия кратных корней характеристического полинома (кратных собственных чисел матрицы) оказывается более сложным.

Пример. Пусть $A = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, $B = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$.

Характеристические полиномы у этих матриц совпадают:

$$P_A(\lambda) = \det \begin{bmatrix} 2 - \lambda & 0 \\ 0 & 2 - \lambda \end{bmatrix} = (2 - \lambda)^2 = \det \begin{bmatrix} 2 - \lambda & 1 \\ 0 & 2 - \lambda \end{bmatrix} = P_B(\lambda).$$

Итак, обе матрицы имеют собственные числа $\lambda = 2$ двойной кратности. Но у матрицы A есть собственный базис (например, стандартный $-e^{(1)}$ и $e^{(2)}$). А вот у матрицы B есть (с точностью до числового множителя) только один собственный вектор. Действительно, решая систему

$$(B - 2I)x = \theta \iff \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \text{получим } x = \gamma \begin{bmatrix} 1 \\ 0 \end{bmatrix} (\gamma \neq 0).$$

Однако еще Григорий Сковорода⁵⁶ сказал: "Слава Создателю, с творившему все ненужное трудным, а все трудное – ненужным". Наиболее часто встречающийся в приложениях класс *самосопряженных* матриц избавлен от отмеченных сложностей. У таких матриц, как будет показано в п.8.1, всегда есть собственный базис.

Теперь мы можем закончить рассмотрение примера, с которого начинается эта глава:

$$A = \begin{bmatrix} 5 & -6 \\ 3 & -4 \end{bmatrix}, \quad P_A(\lambda) = \det \begin{bmatrix} 5 - \lambda & -6 \\ 3 & -4 - \lambda \end{bmatrix} = \lambda^2 - \lambda - 2.$$

Собственные числа матрицы A : $\lambda_1 = 2$, $\lambda_2 = -1$. Находим соответствующие им собственные векторы, решая однородные системы линейных уравнений $(A - \lambda_r I)x^{(r)} = \theta$, $r = 1, 2$.

$$r = 1; \quad \begin{bmatrix} 3 & -6 \\ 3 & -6 \end{bmatrix} \cdot x^{(1)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \iff x^{(1)} = \gamma_1 \begin{bmatrix} 2 \\ 1 \end{bmatrix},$$

где γ_1 – произвольное, отличное от нуля число (отметим, что при $\gamma_1 = 1$ мы получаем уже упоминавшийся в п.6.1 собственный вектор).

$$r = 2; \quad \begin{bmatrix} 6 & -6 \\ 3 & -3 \end{bmatrix} \cdot x^{(2)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \iff x^{(2)} = \gamma_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \gamma_2 \neq 0.$$

⁵⁶Григорий Саввич СКОВОРОДА (1722-1794) – украинский философ, поэт, педагог, вел жизнь странствующего нищего. Его сочинения распространялись в списках.

Векторы $x^{(1)}$ и $x^{(2)}$ образуют собственный базис матрицы A .

Итак, построен алгоритм решения полной проблемы собственных значений для квадратной матрицы:

1. Вычислить коэффициенты характеристического полинома $P_A(\lambda) = \det(A - \lambda I)$.
2. Найти все *попарно различные* корни этого полинома – собственные числа $\lambda_1, \dots, \lambda_m$ ($m \leq n$).
3. Для каждого собственного числа найти все соответствующие ему *линейно независимые* собственные векторы – ненулевые решения однородной линейной системы $(A - \lambda_r I)x^{(r)} = \theta$, $r = 1, \dots, m$.

Серьезное предупреждение. Этот, казалось бы, естественный алгоритм обладает одним убийственным недостатком: не существует численно устойчивых методов его реализации. Поэтому на практике используют другие методы решения полной проблемы собственных значений. Один из таких методов будет рассмотрен в п.8.2.

Отметим еще некоторые свойства собственных векторов и собственных чисел матрицы.

1. Если матрицу умножить на отличное от нуля число, то множество ее собственных векторов не изменится, а собственные числа умножатся на это же число.

Доказательство. Пусть $\alpha \neq 0$. Тогда

$$Ax = \lambda x \iff (\alpha A)x = \alpha(Ax) = \alpha(\lambda x) = (\alpha\lambda)x. \quad \blacksquare$$

2. Если матрица A обратима, то собственные векторы матриц A и A^{-1} совпадают, а их собственные числа взаимно обратны.

Доказательство. Из обратимости A следует, что $\det(A) \neq 0$ и, таким образом, все собственные числа отличны от нуля. Далее,

$$Ax = \lambda x \implies x = A^{-1}Ax = A^{-1}(\lambda x) = \lambda(A^{-1}x) \implies A^{-1}x = \frac{1}{\lambda}x. \quad \blacksquare$$

3. Собственные векторы сопряженных квадратных матриц – сопряженные комплексные числа.

Доказательство.

$$P_{A^*}(\bar{\lambda}) = \det(A^* - \bar{\lambda}I) = \det((A - \lambda I)^*) = \overline{\det(A - \lambda I)} = \overline{P_A(\lambda)}.$$

Поэтому, если $P_A(\lambda) = 0$, то $\overline{P_A(\lambda)} = 0$ и $P_{A^*}(\bar{\lambda}) = 0$. ■

Замечание. В отличие от взаимно обратных матриц собственные векторы эрмитово сопряженных матриц, вообще говоря, никак между собой не связаны.

6.3. Подобные матрицы

Определение. Говорят, что $(n \times n)$ -матрица A подобна $(n \times n)$ -матрице B , если существует такая обратимая $(n \times n)$ -матрица S , что

$$A = S^{-1}BS.$$

Очевидно, что если A подобна B , то и B подобна A , так как

$$A = S^{-1}BS \iff B = SAS^{-1} = (S^{-1})^{-1}BS^{-1}.$$

Поэтому говорят, что матрицы A и B подобны друг другу, Очевидно также, что всякая квадратная матрица подобна самой себе. Далее, если A подобна B и B подобна C , то A подобна C . Действительно,

$$A = S_1^{-1}BS_1 \bigwedge B = S_2^{-1}CS_2 \implies A = S_1^{-1}S_2^{-1}CS_2S_1 = (S_2S_1)^{-1}C(S_2S_1).$$

Рассмотрим теперь свойства собственных чисел подобных матриц.

Теорема. Характеристические полиномы подобных матриц равны.

Доказательство. Пусть $A = S^{-1}BS$. Тогда с учетом (3.4.1) имеем

$$\begin{aligned} P_A(\lambda) &= \det(A - \lambda I) = \det(S^{-1}BS - \lambda I) = \det(S^{-1}BS - \lambda S^{-1}IS) = \\ &= \det(S^{-1}(B - \lambda I)S) = \det(S^{-1}) \cdot \det(B - \lambda I) \cdot \det(S) = \\ &= \det(B - \lambda I) = P_B(\lambda). \end{aligned}$$

Следствие. Собственные числа подобных матриц попарно равны.

Для дальнейшего большое значение имеет следующая

Теорема. Если у матрицы есть собственный базис, то среди подобных ей есть диагональная.

Доказательство. Пусть $s^{(1)}, \dots, s^{(n)}$ – линейно независимые собственные векторы матрицы A ; $\lambda_1, \dots, \lambda_n$ – собственные числа, которым они соответствуют:

$$As^{(r)} = \lambda_r s^{(r)}, \quad r = 1, \dots, n. \tag{6.3.1}$$

Перепишем эту систему равенств в матричной форме

$$AS = S\Lambda, \quad (6.3.2)$$

где $S = [s^{(1)}, \dots, s^{(n)}]$, $\Lambda = \text{diag}[\lambda_1, \dots, \lambda_n]$ (обратите внимание на порядок сомножителей в правой части (6.3.2)!).

Так как векторы $s^{(1)}, \dots, s^{(n)}$ образуют базис, матрица S обратима. Домножив (6.3.2) слева на S^{-1} , получим $S^{-1}AS = \Lambda$. ■

Верно и обратное утверждение: если среди матриц, подобных матрице A , есть диагональная, то на ее диагонали стоят собственные числа матрицы A , и у A есть собственный базис.

Доказательство. Пусть $S^{-1}AS = \Lambda$. Домножив это равенство слева на S , получим $AS = S\Lambda$, что в векторной форме переписывается как (6.3.1). Таким образом, λ_r – собственные числа матрицы A , а $s^{(r)}$ – ее собственные векторы.

Осталось заметить, что в силу обратимости матрицы S ее столбцы – собственные векторы матрицы A – образуют базис. ■

Глава 7. СКАЛЯРНОЕ ПРОИЗВЕДЕНИЕ ВЕКТОРОВ

7.1. Определение и свойства скалярного произведения

Определение. Скалярным произведением векторов $x \in \mathbb{C}^n$ (левый сомножитель) и $y \in \mathbb{C}^n$ (правый сомножитель) называется комплексное число, которое обозначается $\langle x, y \rangle$ и находится по правилу

$$\langle x, y \rangle = x_1 \bar{y}_1 + \dots + x_n \bar{y}_n = \sum_{r=1}^n x_r \bar{y}_r. \quad (7.1.1)$$

Скалярное произведение векторов (одностолбцовых матриц) можно записать и в терминах матричного умножения:

$$\langle x, y \rangle = y^* x. \quad (7.1.2)$$

Рассмотрим свойства скалярного произведения.

1. Скалярный квадрат любого вектора – вещественное неотрицательное число. Более того, он может равняться нулю, только если вектор нулевой.

$$\boxed{\langle x, x \rangle \geq 0; \quad \langle x, x \rangle = 0 \iff x = \theta.}$$

Доказательство. Первая часть утверждения очевидна:

$$\langle x, x \rangle = \sum_{r=1}^n x_r \bar{x}_r = \sum_{r=1}^n |x_r|^2 \geq 0.$$

Далее, скалярный квадрат нулевого вектора равен, очевидно, нулю. С другой стороны, если сумма неотрицательных чисел – квадратов модулей координат вектора – равна нулю, то все координаты равны нулю, т.е. вектор – нулевой. ■

2. При изменении порядка сомножителей скалярное произведение векторов заменяется на сопряженное комплексное число.

$$\boxed{\langle y, x \rangle = \overline{\langle x, y \rangle}.}$$

Доказательство.

$$\langle y, x \rangle = y_1 \bar{x}_1 + \dots + y_n \bar{x}_n = \overline{\bar{y}_1 x_1} + \dots + \overline{\bar{y}_n x_n} = \overline{x_1 \bar{y}_1 + \dots + x_n \bar{y}_n} = \overline{\langle x, y \rangle}. \quad ■$$

3. Скалярное произведение линейно относительно *левого* сомножителя.

$$\langle (x + y), z \rangle = \langle x, z \rangle + \langle y, z \rangle; \quad \alpha \in \mathbb{C} \implies \langle \alpha x, y \rangle = \alpha \langle x, y \rangle.$$

Доказательство. Вычисление по определению. ■

Замечания. 1. Относительно правого сомножителя скалярное произведение линейным *не является*. Действительно, из свойств **2** и **3** следует

$$\langle x, \alpha y \rangle = \overline{\langle \alpha y, x \rangle} = \overline{\alpha} \overline{\langle y, x \rangle} = \overline{\alpha} \langle x, y \rangle.$$

2. В \mathbb{R}^n скалярное произведение вводится также по формуле (7.1.1), но знак комплексного сопряжения становится излишним.

3. В конечномерных линейных пространствах, отличных от \mathbb{C}^n и \mathbb{R}^n , можно ввести скалярное произведение, сопоставив каждой упорядоченной паре векторов x, y ("природа" которых не играет роли) число, обозначаемое $\langle x, y \rangle$. Свобода "назначения" этого числа ограничена следующими *аксиомами скалярного произведения*, которые были проверены выше для \mathbb{C}^n .

1. $\langle x, x \rangle \geq 0; \quad \langle x, x \rangle = 0 \iff x = \theta.$
2. $\langle y, x \rangle = \overline{\langle x, y \rangle}.$
3. $\langle (x + y), z \rangle = \langle x, z \rangle + \langle y, z \rangle; \quad \alpha \in \mathbb{C} \implies \langle \alpha x, y \rangle = \alpha \langle x, y \rangle.$

Убедившись в выполнении этих аксиом, можно использовать все результаты построенной теории. Пример скалярного произведения в пространстве полиномов \mathbb{P}^n будет рассмотрен в п.11.3.

4. Комплексное линейное пространство со скалярным произведением называется *унитарным* пространством, вещественное – *евклидовым*.

Докажем еще два важных свойства скалярного произведения.

4. Если $x \in \mathbb{C}^n$, $y \in \mathbb{C}^m$ и $A - (m \times n)$ -матрица, то

$$\langle x, A^*y \rangle = \langle Ax, y \rangle.$$

Доказательство. Используя (7.1.2), получаем

$$\langle x, A^*y \rangle = (A^*y)^*x = y^*A^{**}x = y^*Ax = \langle Ax, y \rangle. \quad ■$$

5. Для любых двух векторов $x, y \in \mathbb{C}^n$

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \cdot \langle y, y \rangle.$$

Свойство 5 именуется *неравенством Коши–Буняковского–Шварца*⁵⁷ (КБШ).

Доказательство. Если $y = \theta$, то $|\langle x, \theta \rangle|^2 = 0 = \langle x, x \rangle \cdot \langle \theta, \theta \rangle$.

Если $y \neq \theta$, то обозначим для краткости $\langle y, y \rangle = \beta > 0$, $\langle x, y \rangle = \gamma$ и распишем скалярный квадрат вектора, используя свойства 2 и 3:

$$\begin{aligned} \langle \beta x - \gamma y, \beta x - \gamma y \rangle &= \beta^2 \langle x, x \rangle - \beta \bar{\gamma} \langle x, y \rangle - \gamma \bar{\beta} \langle y, x \rangle + \gamma \bar{\gamma} \langle y, y \rangle = \\ &= \beta^2 \langle x, x \rangle - \beta \bar{\gamma} \gamma - \gamma \bar{\beta} \bar{\gamma} + \gamma \bar{\gamma} \beta = \beta(\beta \langle x, x \rangle - \gamma \bar{\gamma}) = \\ &= \beta(\langle y, y \rangle \cdot \langle x, x \rangle - |\langle x, y \rangle|^2). \end{aligned}$$

В силу свойства 1 это выражение неотрицательно. Деля его на *положительное* число β , получаем доказываемое неравенство. ■

Как известно из школьного курса, скалярное произведение векторов в \mathbb{R}^2 (\mathbb{R}^3) равно произведению длин соответствующих им направленных отрезков и косинуса угла между этими отрезками:

$$\langle x, y \rangle = |\vec{x}| \cdot |\vec{y}| \cdot \cos(\widehat{\vec{x}, \vec{y}}).$$

При этом неравенство КБШ становится тривиальным. Действительно, если x и y – ненулевые векторы, то

$$\begin{aligned} |\langle x, y \rangle|^2 &\leq \langle x, x \rangle \cdot \langle y, y \rangle \Leftrightarrow \\ \Leftrightarrow |\vec{x}|^2 \cdot |\vec{y}|^2 \cdot \cos^2(\widehat{\vec{x}, \vec{y}}) &\leq |\vec{x}|^2 \cdot |\vec{y}|^2 \Leftrightarrow \cos^2(\widehat{\vec{x}, \vec{y}}) \leq 1. \end{aligned}$$

7.2. Норма вектора

Известно, что в \mathbb{R}^3 $\langle x, x \rangle = |\vec{x}|^2$ или $|\vec{x}| = \langle x, x \rangle^{1/2}$, т.е. длина направленного отрезка равна корню квадратному из скалярного квадрата соответствующего вектора. Поскольку скалярный квадрат неотрицателен и для векторов из \mathbb{C}^n , можно ввести в \mathbb{C}^n *норму вектора* – обобщение понятия длины направленного отрезка.

Определение. Нормой вектора называется число, которое обозначается $\|x\|$ и находится по правилу

$$\|x\| = \langle x, x \rangle^{1/2}.$$

⁵⁷Виктор Яковлевич БУНЯКОВСКИЙ (1804-1889) – русский математик, член Петербургской АН.

Карл Герман Амандус ШВАРЦ (K.H.A. Schwarz, 1843-1921) – немецкий математик, член Берлинской АН и Петербургской АН.

Используя понятие нормы, можно записать неравенство КБШ в виде

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\|.$$

Рассмотрим свойства нормы.

1. Норма любого вектора – вещественное неотрицательное число.
Более того, она может равняться нулю, *только* если вектор нулевой.

$$\boxed{\|x\| \geq 0; \quad \|x\| = 0 \iff x = \theta.}$$

Доказательство. Следует из свойства 1 скалярного произведения. ■

2. Если $\alpha \in \mathbb{C}$, то

$$\boxed{\|\alpha x\| = |\alpha| \cdot \|x\|.}$$

Перед тем, как доказывать это утверждение, отметим, что часто удобнее работать не с нормой, а с ее квадратом.

Доказательство. $\|\alpha x\|^2 = \langle \alpha x, \alpha x \rangle = \alpha \bar{\alpha} \langle x, x \rangle = |\alpha|^2 \|x\|^2$. ■

Определение. Вектор, норма которого равна единице, называется *нормированным*.

Любой *ненулевой* вектор можно нормировать, разделив его на его собственную норму:

$$\left\| \frac{x}{\|x\|} \right\| = \frac{1}{\|x\|} \cdot \|x\| = 1.$$

3. Для любых $x, y \in \mathbb{C}^n$

$$\boxed{\|x + y\| \leq \|x\| + \|y\|.}$$

Доказательство. В силу аксиом скалярного произведения

$$\|x + y\|^2 = \langle x + y, x + y \rangle = \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle.$$

Учтем теперь, что модуль суммы двух чисел не превышает суммы модулей слагаемых. При этом у заведомо неотрицательных чисел знак модуля опустим.

$$\|x + y\|^2 \leq \|x\|^2 + 2|\langle x, y \rangle| + \|y\|^2.$$

По неравенству КБШ $|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$, откуда

$$\|x + y\|^2 \leq \|x\|^2 + 2\|x\| \cdot \|y\| + \|y\|^2 = (\|x\| + \|y\|)^2. \quad ■$$

Для случая \mathbb{R}^3 свойство **3** хорошо известно: длина стороны треугольника не больше суммы длин остальных его сторон. По этой причине доказанное для произвольного унитарного пространства неравенство называют *неравенством треугольника*.

Замечания. 1. Введенная нами норма вектора называется обычно *евклидовой нормой* или *нормой, порожденной скалярным произведением*.

В \mathbb{C}^n можно задать другие неотрицательные функционалы, обладающие свойствами **1 – 3** евклидовой нормы. Все они также называются нормами. Кроме евклидовой, обозначаемой $\|x\|_2$, чаще всего используются следующие две нормы:

$$\|x\|_1 = |x_1| + \dots + |x_n|; \quad \|x\|_\infty = \max_{r=1,\dots,n} |x_r|.$$

Проверьте выполнение свойств **1 – 3** для норм $\|x\|_1$ и $\|x\|_\infty$.

2. Как и скалярное произведение норма может быть введена в произвольном линейном пространстве: каждому вектору ставится в соответствие неотрицательное вещественное число – норма этого вектора. "Свобода назначения" нормы ограничивается лишь обязательностью выполнения свойств **1 – 3**, которые для евклидовой нормы в \mathbb{C}^n были доказаны, а теперь выступают в роли *аксиом нормы*. После проверки выполнения аксиом могут использоваться все выводы построенной теории.

7.3. Матрица Грама

Пусть A – произвольная $(m \times n)$ -матрица. Рассмотрим квадратную матрицу порядка n $G_A = A^*A$. По определению произведения матриц

$$(g_A)_{km} = \sum_{r=1}^n a_{kr}^* a_{rm} = \sum_{r=1}^n a_{rm} \overline{a_{rk}}.$$

Если обозначить, как принято, k -й столбец матрицы A $a^{(k)} \in \mathbb{C}^m$, то элементы матрицы G_A можно записать в терминах скалярного произведения:

$$(g_A)_{km} = \langle a^{(m)}, a^{(k)} \rangle.$$

Таким образом, матрица G_A содержит все попарные скалярные произведения векторов – столбцов матрицы A .

Эта матрица называется *матрицей Грама*⁵⁸ упорядоченного набора векторов $a^{(1)}, \dots, a^{(n)}$.

⁵⁸ Йорген Педерсен ГРАМ (J.P. Gram, 1850-1916) – датский математик.

Рассмотрим свойства матрицы Грама.

1. Матрица Грама для любого набора векторов – самосопряженная.

Доказательство. $G_A^* = (A^*A)^* = A^*A^{**} = A^*A = G_A$. ■

2. Линейная зависимость набора векторов равносильна вырожденности его матрицы Грама.

Доказательство. Пусть набор векторов $a^{(1)}, \dots, a^{(n)}$ линейно зависим. Тогда имеет ненулевое решение однородное матричное уравнение $Ax = \theta_m$. Умножив это уравнение слева на A^* , получим $A^*Ax = A^*\theta_m$ или $G_Ax = \theta_n$ – однородное уравнение с ненулевым решением. Следовательно, G_A – вырожденная.

Пусть теперь дано, что $\det(G_A) = 0$. Тогда однородное уравнение $G_Ax = \theta_n$ будет иметь ненулевое решение. Обозначим его \tilde{x} и умножим равенство $G_A\tilde{x} = \theta_n$ скалярно на \tilde{x} . Используя свойства **4** и **1** скалярного произведения, получим

$$\langle G_A\tilde{x}, \tilde{x} \rangle = 0 \iff \langle A^*A\tilde{x}, \tilde{x} \rangle = 0 \iff \langle A\tilde{x}, A\tilde{x} \rangle = 0 \iff A\tilde{x} = \theta.$$

Мы получили однородную систему с ненулевым решением, что и доказывает линейную зависимость столбцов матрицы A . ■

7.4. Ортогональность векторов

Определение. Векторы $x, y \in \mathbb{C}^n$ (\mathbb{R}^n) называются *ортогональными*, если $\langle x, y \rangle = 0$. Множество векторов называется ортогональным, если все его векторы попарно ортогональны.

Замечание. В \mathbb{R}^3 ортогональным *ненулевым* векторам соответствуют перпендикулярные направленные отрезки.

Рассмотрим некоторые свойства ортогональных векторов.

1. Нулевой вектор ортогонален любому вектору: $\langle x, \theta \rangle = 0$.

Доказательство. Вычисление по определению. ■

2. Матрица Грама ортогонального набора векторов $a^{(1)}, \dots, a^{(k)}$ диагональна:

$$G_A = \text{diag}[\|a^{(1)}\|^2, \dots, \|a^{(k)}\|^2].$$

Доказательство. $(g_A)_{jm} = \langle a^{(m)}, a^{(j)} \rangle = \delta_{jm} \cdot \|a^{(j)}\|^2$. ■

3. Если ортогональный набор векторов линейно зависим, то он содержит нулевой вектор.

Доказательство. Определитель диагональной матрицы Грама равен произведению ее диагональных элементов – квадратов норм векторов нашего набора. Но по свойству 2 матрицы Грама этот определитель для линейно зависимого набора векторов равен нулю. Следовательно, квадрат нормы хотя бы одного из векторов набора равен нулю. ■

Следующее свойство столь важно, что ему придается ранг теоремы.

Теорема. Любое ортогональное множество векторов в \mathbb{C}^n , не содержащее нулевого вектора, можно дополнить до ортогонального базиса.

Доказательство. Пусть $a^{(1)}, \dots, a^{(k)}$ – ортогональный и не содержащий нулевого вектора набор векторов. Если $k = n$, то этот набор уже является базисом. Если же $k < n$, то построим еще один ненулевой вектор, ортогональный уже имеющемуся набору.

Записывая условия ортогональности искомого вектора всем векторам набора, получим систему уравнений

$$\langle x, a^{(1)} \rangle = 0, \dots, \langle x, a^{(k)} \rangle = 0,$$

или, в матричном виде, $A^*x = \theta_k$, где $A = [a^{(1)}, \dots, a^{(k)}]$.

Матрица A^* имеет размер $k \times n$, и в силу $k < n$ эта система имеет ненулевое решение. Обозначив его $a^{(k+1)}$, получаем ортогональный (по построению) набор из $k + 1$ векторов, не содержащий нулевого вектора. Повторяя эту операцию, получим ортогональный базис \mathbb{C}^n . ■

Замечание. Разложение вектора b в \mathbb{C}^n по базису $a^{(1)}, \dots, a^{(n)}$ сводится, как известно, к решению системы линейных уравнений с квадратной невырожденной матрицей коэффициентов

$$x_1a^{(1)} + \dots + x_na^{(n)} = b \iff Ax = b. \quad (7.4.1)$$

Умножив в случае *ортогонального* базиса обе части матричного уравнения (7.4.1) на A^* слева, получим равносильную систему

$$G_Ax = A^*b.$$

Поскольку матрица G_A диагональна, решение этой системы имеет вид

$$x_k = \frac{a^{(k)*}b}{\|a^{(k)}\|^2} = \frac{\langle b, a^{(k)} \rangle}{\|a^{(k)}\|^2}, \quad k = 1, \dots, n,$$

и требует выполнения существенно меньшего ($\approx 2n^2$) количества арифметических операций, чем в общем случае ($\approx \frac{n^3}{3}$).

7.5. Унитарная матрица

Определение. Ортогональный набор нормированных векторов называется *ортонормированным*.

Определение. Квадратная матрица, столбцы которой образуют ортогональный набор, называется *унитарной*. *Вещественная* унитарная матрица называется *ортогональной*.

Пример. Рассмотрим матрицу второго порядка

$$U_\varphi = \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{bmatrix}.$$

Очевидно, что $\|u_\varphi^{(1)}\| = \|u_\varphi^{(2)}\| = 1$, $\langle u_\varphi^{(1)}, u_\varphi^{(2)} \rangle = 0$. Следовательно, эта матрица ортогональна.

Пусть x – ненулевой вектор в \mathbb{R}^2 , а $y = U_\varphi x$. Легко видеть, что направленный отрезок \vec{y} получается из \vec{x} поворотом на угол φ против часовой стрелки (рис.7.1). Поэтому матрицу U_φ называют *матрицей поворота*. В частности, $U_0 = I$ – матрица поворота на нулевой угол, $U_\pi = -I$ (поворот на угол π – это центральная симметрия).

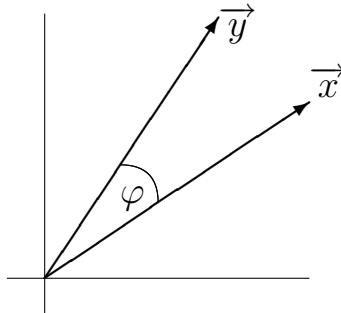


Рис.7.1

Рассмотрим свойства унитарных матриц.

1. Унитарность матрицы U равносильна тому, что $U^*U = I$.

Доказательство. Следует из определения матрицы Грама. ■

2. Умножение векторов из \mathbb{C}^n на унитарную матрицу не меняет их скалярных произведений и норм.

Доказательство. $\langle Ux, Uy \rangle = \langle U^*Ux, y \rangle = \langle Ix, y \rangle = \langle x, y \rangle$. В частности, $\langle Ux, Ux \rangle = \langle x, x \rangle$, т.е. $\|Ux\| = \|x\|$. ■

3. Если U – унитарная матрица, то и U^* – унитарная матрица.

Доказательство. Из $U^*U = I$ следует, что матрицы U^* и U взаимно обратны, а тогда $(U^*)^*U^* = UU^{-1} = I$. ■

4. Если V – унитарная матрица того же порядка, что U , то их произведение – унитарная матрица.

Доказательство. $(UV)^*(UV) = V^*(U^*U)V = V^*V = I$. ■

Пример. Покажите, что $U_\varphi U_\psi = U_{\varphi+\psi}$; $U_\varphi^{-1} = U_{-\varphi}$.

5. Модули всех собственных чисел унитарной матрицы равны единице: $|\lambda(U)| = 1$.

Доказательство. Пусть x – собственный вектор матрицы U , соответствующий собственному числу λ . Тогда по свойству 2

$$Ux = \lambda x \implies \|Ux\| = |\lambda| \cdot \|x\| \implies \|x\| = |\lambda| \cdot \|x\| \implies |\lambda| = 1.$$

Поскольку определитель матрицы равен произведению ее собственных чисел, отсюда, в частности, следует, что $|det(U)| = 1$. ■

Пример. Покажите, что собственные числа матрицы U_φ равны $exp(i\varphi)$ и $exp(-i\varphi)$.

6. Собственные векторы, соответствующие различным собственным числам унитарной матрицы, ортогональны.

Доказательство. Если $Ux = \lambda x$, $Uy = \mu y$, то по свойству 2

$$\langle x, y \rangle = \langle Ux, Uy \rangle = \langle \lambda x, \mu y \rangle = \lambda \bar{\mu} \langle x, y \rangle \implies (1 - \lambda \bar{\mu}) \langle x, y \rangle = 0. \quad (7.5.1)$$

Но по свойству 5 $|\mu| = 1$, т.е. $\bar{\mu} = \frac{1}{\mu}$. По условию $\lambda \neq \mu$. Отсюда $\lambda \bar{\mu} = \frac{\lambda}{\mu} \neq 1$, и из (7.5.1) вытекает $\langle x, y \rangle = 0$. ■

Пример. Найдите собственные векторы матрицы U_φ в \mathbb{C}^2 и проверьте их ортогональность.

Обратите внимание на то, что при $\varphi \neq k\pi$ ($k \in \mathbb{Z}$) матрица U_φ не имеет собственных векторов в \mathbb{R}^2 . Дайте этому факту геометрическую интерпретацию.

7.6. Площадь параллелограмма и объем параллелепипеда

Свойство 2 унитарных матриц имеет в \mathbb{R}^3 важную геометрическую интерпретацию: если x, y, z – три линейно независимых вектора (соответствующие им направленные отрезки $\vec{x}, \vec{y}, \vec{z}$ некомпланарны), и $x' = Ux, y' = Uy, z' = Uz$, где U – ортогональная матрица, то длины отрезков $\vec{x}', \vec{y}', \vec{z}'$ и углы между ними те же, что у тройки $\vec{x}, \vec{y}, \vec{z}$.

Докажем, что справедливо и обратное утверждение: если длины отрезков и углы между ними одинаковы для некомпланарных троек $\vec{x}, \vec{y}, \vec{z}$ и $\vec{x}', \vec{y}', \vec{z}'$, то

$$[x' y' z'] = U \cdot [x y z],$$

где U – ортогональная матрица.

Действительно, из условия следует равенство скалярных произведений соответствующих пар векторов

$$\begin{aligned} \langle x', x' \rangle &= \langle x, x \rangle; & \langle y', y' \rangle &= \langle y, y \rangle; & \langle z', z' \rangle &= \langle z, z \rangle; \\ \langle x', y' \rangle &= \langle x, y \rangle; & \langle x', z' \rangle &= \langle x, z \rangle; & \langle y', z' \rangle &= \langle y, z \rangle, \end{aligned}$$

т.е.

$$[x' y' z']^* \cdot [x' y' z'] = [x y z]^* \cdot [x y z]. \quad (7.6.1)$$

Обозначим $U = [x' y' z'] \cdot [x y z]^{-1}$ (матрица $[x y z]$ обратима, так как тройка $\vec{x}, \vec{y}, \vec{z}$ некомпланарна и, следовательно, векторы x, y, z линейно независимы). Домножив равенство (7.6.1) справа на $[x y z]^{-1}$, а слева на $([x y z]^*)^{-1}$, получим $U^* U = I$. ■

Замечание. Аналогичное утверждение верно для неколлинеарных пар направленных отрезков в \mathbb{R}^2 .

Пример. Поворот на угол φ вокруг оси x_3 в \mathbb{R}^3 , очевидно, сохраняет длины отрезков и углы между ними. Направленные отрезки $\vec{e}^{(1)}, \vec{e}^{(2)}, \vec{e}^{(3)}$ – орты – переходят при этом повороте соответственно в отрезки $\vec{g}^{(1)}, \vec{g}^{(2)}, \vec{g}^{(3)}$, где

$$g^{(1)} = [\cos(\varphi), \sin(\varphi), 0]^T, \quad g^{(2)} = [-\sin(\varphi), \cos(\varphi), 0]^T, \quad g^{(3)} = e^{(3)}.$$

Поэтому матрица

$$V_\varphi = \begin{bmatrix} g^{(1)} & g^{(2)} & g^{(3)} \end{bmatrix} \cdot \begin{bmatrix} e^{(1)} & e^{(2)} & e^{(3)} \end{bmatrix}^{-1} = \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) & 0 \\ \sin(\varphi) & \cos(\varphi) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

ортогональна (проверьте это по определению). Ее называют матрицей поворота в плоскости x_1Ox_2 (или матрицей плоского вращения).

Получим теперь формулы для вычисления площади параллелограмма и объема параллелепипеда.

1. Пусть параллелограмм в \mathbb{R}^2 построен на направленных отрезках \vec{x} и \vec{y} . Покажем, что его площадь равна $|det[x y]|$.

Рассмотрим сначала параллелограмм, изображенный на рис.7.2 (одна сторона лежит на оси абсцисс).

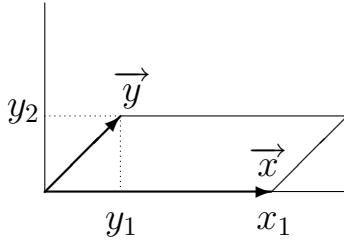


Рис.7.2

Очевидно, что

$$x = \begin{bmatrix} x_1 \\ 0 \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}; \quad S = |y_2| \cdot |x_1| = \left| \det \begin{bmatrix} x_1 & y_1 \\ 0 & y_2 \end{bmatrix} \right| = |\det[x \ y]|.$$

Пусть теперь неколлинеарные отрезки \vec{x} , \vec{y} расположены произвольно. Для вычисления площади построенного на них параллелограмма рассмотрим конгруэнтный параллелограмм, одна из сторон которого лежит на оси абсцисс (рис.7.3).

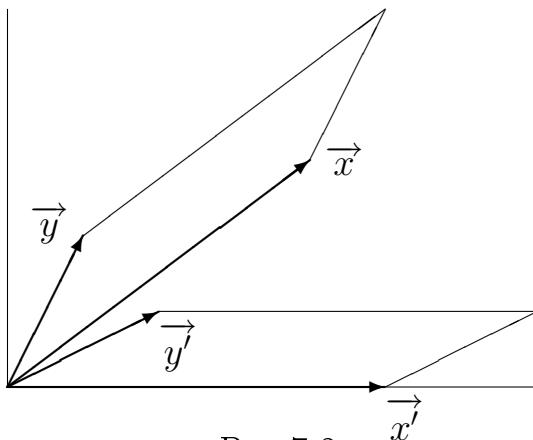


Рис.7.3

Мы доказали существование такой ортогональной матрицы U , что $x' = Ux$, $y' = Uy$. Поэтому

$$S = |\det[x' \ y']| = |\det(U \cdot [x \ y])| = \det(U) \cdot |\det[x \ y]| = |\det[x \ y]|. \quad (7.6.2)$$

Легко видеть, что формула (7.6.2) верна и в случае коллинеарных отрезков, когда площадь равна нулю.

2. Пусть параллелепипед в \mathbb{R}^3 построен на направленных отрезках \vec{x} , \vec{y} и \vec{z} . Покажем, что его объем равен $|\det[x \ y \ z]|$.

Рассмотрим сначала параллелепипед, изображенный на рис.7.4 (одна грань лежит в координатной плоскости):

$$x = [x_1, x_2, 0]^T, \quad y = [y_1, y_2, 0]^T, \quad z = [z_1, z_2, z_3]^T.$$

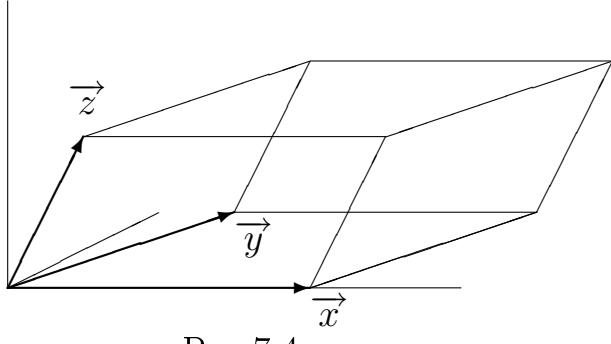


Рис.7.4

Объем параллелепипеда равен произведению высоты на площадь основания:

$$V = |z_3| \cdot \left| \det \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \end{bmatrix} \right| = \left| \det \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ 0 & 0 & z_3 \end{bmatrix} \right| = |\det[x \ y \ z]|.$$

Пусть теперь некомпланарные отрезки \vec{x} , \vec{y} , \vec{z} расположены произвольно. Для вычисления объема построенного на них параллелепипеда рассмотрим конгруэнтный параллелепипед с гранью, лежащей в координатной плоскости.

Мы доказали существование такой ортогональной матрицы U , что $x' = Ux$, $y' = Uy$, $z' = Uz$. Поэтому

$$\begin{aligned} V &= |\det[x' \ y' \ z']| = |\det(U \cdot [x \ y \ z])| = \\ &= |\det(U)| \cdot |\det[x \ y \ z]| = |\det[x \ y \ z]|. \end{aligned} \quad (7.6.3)$$

Убедитесь, что формула (7.6.3) верна и в случае компланарных отрезков, когда объем равен нулю.

3. Вычислим, наконец, площадь параллелограмма в \mathbb{R}^3 для случая, когда он не лежит в координатной плоскости. Пусть этот параллелограмм построен на неколлинеарных отрезках \vec{x} и \vec{y} .

Построим третий отрезок \vec{w} , перпендикулярный \vec{x} и \vec{y} . Координаты вектора w удовлетворяют условиям ортогональности

$$\begin{cases} x_1 w_1 + x_2 w_2 + x_3 w_3 = 0 \\ y_1 w_1 + y_2 w_2 + y_3 w_3 = 0 \end{cases}.$$

Одно из ненулевых решений этой системы $w = [\Delta_{23}, -\Delta_{13}, \Delta_{12}]^T$, где

$$\Delta_{12} = \det \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \end{bmatrix}, \quad \Delta_{23} = \det \begin{bmatrix} x_2 & y_2 \\ x_3 & y_3 \end{bmatrix}, \quad \Delta_{13} = \det \begin{bmatrix} x_1 & y_1 \\ x_3 & y_3 \end{bmatrix}.$$

Действительно,

$$\langle x, w \rangle = x_1 \Delta_{23} - x_2 \Delta_{13} + x_3 \Delta_{12} = \det \begin{bmatrix} x_1 & x_1 & y_1 \\ x_2 & x_2 & y_2 \\ x_3 & x_3 & y_3 \end{bmatrix} = 0.$$

Аналогично,

$$\langle y, w \rangle = \det \begin{bmatrix} y_1 & x_1 & y_1 \\ y_2 & x_2 & y_2 \\ y_3 & x_3 & y_3 \end{bmatrix} = 0.$$

Кроме того, $w \neq \theta$ в силу линейной независимости векторов x и y .

Нормируем вектор w :

$$z = \frac{w}{\|w\|} = \frac{1}{(\Delta_{12}^2 + \Delta_{23}^2 + \Delta_{13}^2)^{1/2}} \begin{bmatrix} \Delta_{23} \\ -\Delta_{13} \\ \Delta_{12} \end{bmatrix}.$$

Очевидно, что объем параллелепипеда с *единичной* высотой, построенного на направленных отрезках \vec{x} , \vec{y} , \vec{z} , численно равен площади его основания – параллелограмма, построенного на направленных отрезках \vec{x} и \vec{y} :

$$\begin{aligned} S = V &= |\det[x \ y \ z]| = \frac{1}{\|w\|} \cdot |\det[x \ y \ w]| = \\ &= \frac{1}{(\Delta_{12}^2 + \Delta_{23}^2 + \Delta_{13}^2)^{1/2}} \cdot \left| \det \begin{bmatrix} x_1 & y_1 & \Delta_{23} \\ x_2 & y_2 & -\Delta_{13} \\ x_3 & y_3 & \Delta_{12} \end{bmatrix} \right| = \\ &= \frac{\Delta_{12}^2 + \Delta_{23}^2 + \Delta_{13}^2}{(\Delta_{12}^2 + \Delta_{23}^2 + \Delta_{13}^2)^{1/2}} = (\Delta_{12}^2 + \Delta_{23}^2 + \Delta_{13}^2)^{1/2}. \end{aligned} \quad (7.6.4)$$

Убедитесь, что в случае, когда отрезки \vec{x} и \vec{y} лежат в координатной плоскости, формула (7.6.4) превращается в (7.6.2).

Терминологическое замечание. В "векторной алгебре т.е. в алгебре направленных отрезков, отрезок \vec{w} , соответствующий вектору $w = [\Delta_{23}, -\Delta_{13}, \Delta_{12}]^T$ называют *векторным произведением* отрезков \vec{x} и \vec{y} . Отметим свойства векторного произведения:

1. $\vec{w} \perp \vec{x}$, $\vec{w} \perp \vec{y}$.
2. Площадь параллелограмма, построенного на \vec{x} и \vec{y} , равна $|\vec{w}|$.
3. Направленные отрезки \vec{x} , \vec{y} , \vec{w} образуют правую тройку. В этом можно убедиться, положив $x = e^{(1)}$, $y = e^{(2)}$. Тогда $w = e^{(3)}$.

Число $\det[x \ y \ z]$ называют *смешанным* (векторно-скалярным) произведением направленных отрезков \vec{x} , \vec{y} , \vec{z} .

Так как $\det[e^{(1)} e^{(2)} e^{(3)}] = 1$, смешанное произведение положительно, если сомножители образуют правую тройку, и отрицательно, если левую.

7.7. Алгоритм Грама–Шмидта. *QR*-разложение матрицы

В заключение этой главы рассмотрим алгоритм Грама–Шмидта⁵⁹, который позволяет, имея линейно независимый набор из k векторов в \mathbb{C}^n ($k \leq n$), построить ортонормированный набор из k векторов.

Итак, пусть $a^{(1)}, \dots, a^{(k)}$ – линейно независимые векторы. Положим $b^{(1)} = a^{(1)}$ и $b^{(2)} = a^{(2)} - \alpha_{12}b^{(1)}$.

Число α_{12} выберем так, чтобы $\langle b^{(2)}, b^{(1)} \rangle = 0$, т.е. чтобы $b^{(2)}$ и $b^{(1)}$ были ортогональны:

$$\langle b^{(2)}, b^{(1)} \rangle = \langle a^{(2)}, b^{(1)} \rangle - \alpha_{12}\langle b^{(1)}, b^{(1)} \rangle = 0 \iff \alpha_{12} = \frac{\langle a^{(2)}, b^{(1)} \rangle}{\langle b^{(1)}, b^{(1)} \rangle}.$$

Далее, если уже построены попарно ортогональные векторы $b^{(1)}, \dots, b^{(m)}$, и $m < k$, то положим

$$b^{(m+1)} = a^{(m+1)} - \alpha_{1,m+1}b^{(1)} - \dots - \alpha_{m,m+1}b^{(m)}. \quad (7.7.1)$$

Умножая (7.7.1) скалярно на $b^{(r)}$, $1 \leq r \leq m$, получим уравнение

$$\langle b^{(m+1)}, b^{(r)} \rangle = \langle a^{(m+1)}, b^{(r)} \rangle - \alpha_{r,m+1}\langle b^{(r)}, b^{(r)} \rangle = 0$$

(остальные слагаемые исчезнут вследствие попарной ортогональности уже построенных векторов). Отсюда $\alpha_{r,m+1} = \langle a^{(m+1)}, b^{(r)} \rangle / \langle b^{(r)}, b^{(r)} \rangle$.

Осталось показать, что среди построенных ортогональных векторов $b^{(1)}, \dots, b^{(k)}$ нет нулевого. Предположим, напротив, что $b^{(1)} \neq \theta, \dots, b^{(m)} \neq \theta$, но $b^{(m+1)} = \theta$. Подставив в равенство

$$\theta = a^{(m+1)} - \alpha_{1,m+1}b^{(1)} - \dots - \alpha_{m,m+1}b^{(m)}$$

выражения векторов $b^{(1)}, \dots, b^{(m)}$ через векторы $a^{(1)}, \dots, a^{(m)}$, получим

$$\theta = a^{(m+1)} + \gamma_1 a^{(1)} + \dots + \gamma_m a^{(m)},$$

где γ_r , $r = 1, \dots, m$ – некоторые числа.

Поскольку коэффициент при $a^{(m+1)}$ отличен от нуля, полученное равенство противоречит линейной независимости исходного набора векторов. Таким образом, среди построенных векторов нулевых нет. Нормировав эти векторы, мы закончим работу алгоритма Грама–Шмидта.

⁵⁹Эрхард ШМИДТ (E. Schmidt, 1876-1959) – немецкий математик.

Перепишем равенство (7.7.1) в виде

$$a^{(m+1)} = \alpha_{1,m+1} b^{(1)} + \dots + \alpha_{m,m+1} b^{(m)} + b^{(m+1)}; \quad m = 1, \dots, k-1. \quad (7.7.2)$$

и объединим наборы векторов в матрицы:

$$A = \begin{bmatrix} a^{(1)} & \dots & a^{(k)} \end{bmatrix}; \quad B = \begin{bmatrix} b^{(1)} & \dots & b^{(k)} \end{bmatrix}.$$

Из формулы (7.7.2) видно, что $(m+1)$ -й столбец матрицы A может быть получен из матрицы B умножением справа на столбец $[\alpha_{1,m+1}, \dots, \alpha_{m,m+1}, 1, 0, \dots, 0]^T$, и, следовательно, вся матрица A получается умножением матрицы B справа на верхнюю треугольную матрицу с единичной диагональю:

$$A = B\alpha, \quad (7.7.3)$$

где

$$\alpha = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \dots & \alpha_{1k} \\ 0 & 1 & \alpha_{23} & \dots & \alpha_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

Нормирование построенных векторов можно осуществить с помощью умножения справа на матрицу D^{-1} , где $D = \text{diag} [\|b^{(1)}\|, \dots, \|b^{(1)}\|]$. Тогда (7.7.3) перейдет в $A = B(D^{-1}D)\alpha = QR$, где $Q = BD^{-1}$ – матрица с ортонормированными столбцами, $R = D\alpha$ – верхняя треугольная матрица.

Представление $(n \times k)$ -матрицы A с линейно независимыми столбцами в виде произведения $(n \times k)$ -матрицы Q с ортонормированными столбцами и верхней треугольной $(k \times k)$ -матрицы R называется *QR-разложением* матрицы A .

В частности, если A – квадратная невырожденная матрица, то Q – унитарная матрица.

Серьезное предупреждение. Необходимо отметить, что изложенный выше алгоритм, с помощью которого была доказана возможность получения *QR*-разложения, численно неустойчив и не может быть использован для вычислений. Численно устойчивые алгоритмы, выполняющие *QR*-разложение, реализованы в средах конечного пользователя и в виде стандартных программ на Фортране.

Глава 8. САМОСОПРЯЖЕННАЯ МАТРИЦА

8.1. Свойства собственных чисел и собственных векторов самосопряженной матрицы

Напомним, что квадратная матрица A называется самосопряженной (эрмитовой), если $A^* = A$.

Изучим свойства собственных чисел и собственных векторов самосопряженной матрицы.

1. Собственные числа самосопряженной матрицы вещественны.

Доказательство. Для начала отметим, что условие $A^* = A$ и свойство 4 скалярного произведения дают

$$\langle Ax, x \rangle = \langle x, A^*x \rangle = \langle x, Ax \rangle. \quad (8.1.1)$$

Пусть теперь $A^* = A$, $Ax = \lambda x$, $x \neq 0$. Тогда из (8.1.1) имеем

$$\lambda \langle x, x \rangle = \langle \lambda x, x \rangle = \langle Ax, x \rangle = \langle x, Ax \rangle = \langle x, \lambda x \rangle = \bar{\lambda} \langle x, x \rangle.$$

Сокращая на $\langle x, x \rangle \neq 0$, получим $\lambda = \bar{\lambda}$, т.е. $\lambda \in \mathbb{R}$. ■

2. Собственные векторы самосопряженной матрицы, соответствующие попарно различным собственным числам, ортогональны.

Доказательство. Пусть $A^* = A$, $Ax = \lambda x$, $Ay = \mu y$. Тогда из (8.1.1) с учетом $\mu \in \mathbb{R}$ имеем

$$\begin{aligned} \lambda \langle x, y \rangle &= \langle \lambda x, y \rangle = \langle Ax, y \rangle = \langle x, Ay \rangle = \langle x, \mu y \rangle = \mu \langle x, y \rangle \implies \\ &\implies (\lambda - \mu) \langle x, y \rangle = 0. \end{aligned}$$

Но $\lambda \neq \mu$ и, следовательно, $\langle x, y \rangle = 0$. ■

Замечания. 1. Напомним, что в случае произвольной квадратной матрицы попарное различие собственных чисел обеспечивает лишь линейную независимость соответствующих собственных векторов.

2. Сравните доказанные свойства самосопряженной матрицы со свойствами 6 и 7 унитарных матриц.

Важнейшее свойство самосопряженной матрицы устанавливает

Теорема. Для любой самосопряженной матрицы существует ортонормированный базис, состоящий из ее собственных векторов.

Доказательство проведем индукцией по порядку матрицы A .

Для матрицы порядка 1 утверждение теоремы очевидно. Пусть оно доказано для матриц порядка $k - 1$. Рассмотрим произвольную эрмитову матрицу A порядка k и найдем какой-нибудь корень λ_1 ее характеристического полинома $P_A(\lambda)$. Пусть $s^{(1)}$ – нормированный собственный вектор, соответствующий λ_1 .

Дополним "набор состоящий из одного вектора $s^{(1)}$, до ортонормированного базиса в \mathbb{C}^k векторами $g^{(2)}, \dots, g^{(k)}$. Собственные векторы $s^{(2)}, \dots, s^{(k)}$ будем искать в виде

$$s^{(r)} = \alpha_2^{(r)} g^{(2)} + \dots + \alpha_k^{(r)} g^{(k)} \quad \text{или} \quad s^{(r)} = D\alpha^{(r)},$$

где $D = [s^{(1)}, g^{(2)}, \dots, g^{(k)}]$, $\alpha^{(r)} = [0, \alpha_2^{(r)}, \dots, \alpha_k^{(r)}]^T$. Отметим, что матрица D унитарна по построению, и потому $D^* = D^{-1}$.

По определению собственного вектора

$$As^{(r)} = \lambda_r s^{(r)} \quad \text{или} \quad AD\alpha^{(r)} = \lambda_r D\alpha^{(r)},$$

откуда

$$D^*AD\alpha^{(r)} = \lambda_r \alpha^{(r)}. \quad (8.1.2)$$

Матрица AD имеет вид

$$AD = [As^{(1)}, Ag^{(2)}, \dots, Ag^{(k)}] = [\lambda_1 s^{(1)}, Ag^{(2)}, \dots, Ag^{(k)}]. \quad (8.1.3)$$

Поскольку матрица D^*AD эрмитова (проверьте это!), ее можно записать в виде

$$D^*AD = \begin{bmatrix} c & d^* \\ \hline d & B \end{bmatrix},$$

где B – эрмитова матрица порядка $k - 1$, c – матрица первого порядка (число), d – столбец высоты $k - 1$. При этом из (8.1.3) следует

$$c = \langle \lambda_1 s^{(1)}, s^{(1)} \rangle = \lambda_1; \quad d_r = \langle Ag^{(r)}, s^{(1)} \rangle = \langle g^{(r)}, As^{(1)} \rangle = \lambda_1 \langle g^{(r)}, s^{(1)} \rangle = 0,$$

так как $g^{(r)}$ ортогональны $s^{(1)}$ по построению. Итак,

$$D^*AD = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \dots & & B & \\ 0 & & & \end{bmatrix}.$$

Поскольку $D^* = D^{-1}$, матрицы D^*AD и A подобны, и их характеристические полиномы совпадают:

$$P_A(\lambda) = P_{D^*AD}(\lambda) = (\lambda_1 - \lambda) \cdot P_B(\lambda).$$

Поэтому собственные числа матрицы B являются также собственными числами матрицы A . Верно и обратное (за исключением, может быть, числа λ_1).

Система уравнений (8.1.2) для определения собственных чисел λ_r и векторов $\alpha^{(r)}$ имеет вид

$$D^*AD\alpha^{(r)} = \left[\begin{array}{c|cccc} \lambda_1 & 0 & \dots & 0 \\ \hline 0 & & & & \\ \dots & & B & & \\ 0 & & & & \end{array} \right] \cdot \begin{bmatrix} 0 \\ \alpha_2^{(r)} \\ \dots \\ \alpha_k^{(r)} \end{bmatrix} = \lambda_r \cdot \begin{bmatrix} 0 \\ \alpha_2^{(r)} \\ \dots \\ \alpha_k^{(r)} \end{bmatrix}.$$

Она, очевидно, равносильна системе

$$B \cdot \begin{bmatrix} \alpha_2^{(r)} \\ \dots \\ \alpha_k^{(r)} \end{bmatrix} = \lambda_r \cdot \begin{bmatrix} \alpha_2^{(r)} \\ \dots \\ \alpha_k^{(r)} \end{bmatrix}.$$

По индукционному предположению матрица B имеет ортонормированный собственный базис в \mathbb{C}^{k-1} . Обозначим его векторы $a^{(2)}, \dots, a^{(k)}$ и положим

$$\alpha^{(r)} = \begin{bmatrix} 0 \\ a^{(r)} \end{bmatrix}, \quad r = 2, \dots, k.$$

Из $Ba^{(r)} = \lambda_r a^{(r)}$ имеем $D^*AD\alpha^{(r)} = \lambda_r \alpha^{(r)}$. Отсюда следует $A(D\alpha^{(r)}) = \lambda_r(D\alpha^{(r)})$, т.е. $As^{(r)} = \lambda_r s^{(r)}$.

Но $a^{(r)}$ ортонормированы, следовательно, и $\alpha^{(r)}$ ортонормированы. Поскольку умножение на унитарную матрицу D сохраняет скалярное произведение и норму, векторы $s^{(2)}, \dots, s^{(k)}$ также ортонормированы. Нормированный вектор $s^{(1)}$ ортогонален векторам $g^{(2)}, \dots, g^{(k)}$ и, следовательно, ортогонален векторам $s^{(2)}, \dots, s^{(k)}$. Ортонормированный собственный базис матрицы A построен, и теорема доказана. ■

Следствие. Всякая эрмитова матрица A подобна диагональной матрице Λ , на диагонали которой стоят собственные числа A .

Матрица S , с помощью которой осуществляется подобие, унитарна, ибо ее столбцы – ортонормированные собственные векторы матрицы A . Говорят, что матрицы A и $\Lambda = \text{diag}[\lambda_1, \dots, \lambda_n]$ *унитарно подобны*.

Домножив равенство $S^{-1}AS = \Lambda$ на S слева и на $S^* = S^{-1}$ справа, получим

$$A = S\Lambda S^*.$$

Такое представление самосопряженной матрицы называется ее *спектральным разложением*.

Пример. $A = \begin{bmatrix} 0 & i & 1 \\ -i & 0 & -i \\ 1 & i & 0 \end{bmatrix}$. Легко видеть, что $A^* = A$. Прямым вычислением получаем

$$P_A(\lambda) = \det(A - \lambda I) = -\lambda^3 + 3\lambda + 3.$$

Корни характеристического полинома $\lambda_1 = 2$, $\lambda_2 = \lambda_3 = -1$ – собственные числа матрицы A .

Для нахождения собственного вектора $s^{(1)}$ решим однородную систему $(A - \lambda_1 \cdot I)s^{(1)} = \theta$:

$$\left| \begin{array}{ccc|c} -2 & i & 1 & 0 \\ -i & -2 & -i & 0 \\ 1 & i & -2 & 0 \end{array} \right| \Leftrightarrow \left| \begin{array}{ccc|c} 1 & -i/2 & -1/2 & 0 \\ 0 & -3/2 & -3i/2 & 0 \\ 0 & 3i/2 & -3/2 & 0 \end{array} \right| \Leftrightarrow \left| \begin{array}{ccc|c} 1 & 0 & -1 & 0 \\ 0 & 1 & i & 0 \\ 0 & 0 & 0 & 0 \end{array} \right|$$

откуда $s^{(1)} = \alpha[1 \ -i \ 1]^T$.

Подберем α из условия $\|s^{(1)}\| = 1$:

$$\alpha = \frac{1}{\sqrt{3}}; \quad s^{(1)} = \frac{1}{\sqrt{3}}[1 \ -i \ 1]^T.$$

Для нахождения собственного вектора $s^{(2)}$ решим однородную систему $(A - \lambda_2 \cdot I)s^{(2)} = \theta$:

$$\left| \begin{array}{ccc|c} 1 & i & 1 & 0 \\ -i & 1 & -i & 0 \\ 1 & i & 1 & 0 \end{array} \right| \Leftrightarrow \left| \begin{array}{ccc|c} 1 & i & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right|$$

Одно из решений этой системы $s^{(2)} = \beta[1 \ 0 \ -1]^T$. Подберем β из условия $\|s^{(2)}\| = 1$:

$$\beta = \frac{1}{\sqrt{2}}; \quad s^{(2)} = \frac{1}{\sqrt{2}}[1 \ 0 \ -1]^T.$$

Третье собственное число равно второму. Следовательно, система для определения $s^{(3)}$ совпадает с системой для определения $s^{(2)}$. Но если $s^{(2)}$ и $s^{(3)}$ ортогональны $s^{(1)}$ "автоматически" (свойство 2 самосопряженной матрицы), то условие ортогональности $s^{(2)}$ и $s^{(3)}$ дает дополнительное уравнение. Итак,

$$\begin{array}{ccc|c} 1 & i & 1 & 0 \\ -i & 1 & -i & 0 \\ 1 & i & 1 & 0 \\ 1 & 0 & -1 & 0 \end{array} \iff \begin{array}{ccc|c} 1 & 0 & -1 & 0 \\ 0 & 1 & -2i & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}$$

откуда $s^{(3)} = \gamma [1 \ 2i \ 1]^T$. Условие нормировки дает $s^{(3)} = \frac{1}{\sqrt{6}} [1 \ 2i \ 1]^T$.

Запишем спектральное разложение матрицы A :

$$A = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ -\frac{i}{\sqrt{3}} & 0 & \frac{2i}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \end{bmatrix} \cdot \begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{i}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & \frac{-2i}{\sqrt{6}} & \frac{1}{\sqrt{6}} \end{bmatrix} = S \Lambda S^*$$

8.2. Решение полной проблемы собственных значений для самосопряженной матрицы. Метод Якоби

В п.6.2 было указано, что очевидный алгоритм решения полной проблемы собственных значений, вытекающий из определения собственных чисел и собственных векторов, неприменим из-за его численной неустойчивости. Практическое применение получили методы, основанные на других идеях. Один из них – метод Якоби – рассматривается ниже. Мы ограничимся случаем, когда A – вещественная симметричная матрица.

Если найдется такая ортогональная матрица U , что $\Lambda = U^T A U$ – диагональная матрица, то (см. п.6.3) на диагонали Λ будут стоять собственные числа матрицы A , а столбцы матрицы U образуют собственный базис матрицы A . Будем строить матрицы, унитарно подобные матрице A , добиваясь превращения всех внедиагональных элементов в нули.

На первом шаге алгоритма Якоби мы добьемся того, чтобы *наибольший по модулю внедиагональный элемент* (назовем его ведущим элементом первого шага) обратился в нуль. Вследствие симметрии матрицы таких элементов четное число. Если их больше двух, можно выбрать любую пару a_{ik}, a_{ki} (например, пару с наименьшей суммой $i + k$). Будем называть строки и столбцы с номерами i и k *отмеченными*.

На рисунке схематически изображена исходная матрица $A^{(0)} = A$ с отмеченной парой ведущих элементов первого шага $a_{ik}^{(0)}$ и $a_{ki}^{(0)}$ ($|a_{ik}^{(0)}| = \max_{p \neq m} |a_{pm}^{(0)}|$).

$$A^{(0)} = \begin{bmatrix} & \vdots & & & \vdots \\ \cdots & \vdots & \cdots & a_{ik}^{(0)} & \cdots \\ & \vdots & & \vdots & \\ \cdots & a_{ki}^{(0)} & \cdots & \vdots & \cdots \\ & \vdots & & \vdots & \end{bmatrix}.$$

Представим результат первого шага – матрицу $A^{(1)}$, унитарно подобную A , – в виде $A^{(1)} = U^{(1)T} A^{(0)} U^{(1)}$, где $U^{(1)}$ – ортогональная матрица, изображенная схематически ниже.

$$U^{(1)} = \begin{bmatrix} \mathbb{I} & \vdots & \mathbb{O} & \vdots & \mathbb{O} \\ \cdots & c_1 & \cdots & -s_1 & \cdots \\ \mathbb{O} & \vdots & \mathbb{I} & \vdots & \mathbb{O} \\ \cdots & s_1 & \cdots & c_1 & \cdots \\ \mathbb{O} & \vdots & \mathbb{O} & \vdots & \mathbb{I} \end{bmatrix}.$$

Здесь символом \mathbb{I} обозначена единичная подматрица, символом \mathbb{O} – нулевая подматрица. Таким образом, матрица $U^{(1)}$ получается из единичной путем замены двух диагональных и двух внедиагональных элементов, стоящих на пересечении отмеченных строк и столбцов:

$$u_{ii}^{(1)} = u_{kk}^{(1)} = c_1 = \cos(\varphi_1); \quad u_{ki}^{(1)} = -u_{ik}^{(1)} = s_1 = \sin(\varphi_1)$$

$$(\varphi_1 – угол, подлежащий определению из условия $a_{ik}^{(1)} = a_{ik}^{(0)} = 0$).$$

По определению умножения матриц из $A^{(1)} = U^{(1)T} A^{(0)} U^{(1)}$ следует

$$a_{pm}^{(1)} = \sum_{r=1}^n u_{pr}^{(1)T} \sum_{j=1}^n a_{rj}^{(0)} u_{jm}^{(1)}. \quad (8.2.1)$$

Покажем, что элементы матрицы, *не стоящие в отмеченных строках и столбцах*, не изменяются. Рассмотрим внутреннюю сумму в (8.2.1). Если m -й столбец не отмечен, то эта сумма состоит из одного слагаемого:

$$\sum_{j=1}^n a_{rj}^{(0)} u_{jm}^{(1)} = a_{rm}^{(0)} u_{mm}^{(1)} = a_{rm}^{(0)}.$$

Если p -я строка не отмечена, то внешняя сумма тоже состоит из одного слагаемого:

$$a_{pm}^{(1)} = \sum_{r=1}^n u_{pr}^{(1)T} a_{rm}^{(0)} = u_{pp}^{(1)T} a_{pm}^{(0)} = a_{pm}^{(0)}.$$

Вычислим теперь элемент матрицы $A^{(1)}$, стоящий в *отмеченном столбце* и в *неотмеченной строке*. Пусть, например, $m = i$, $p \neq i$, $p \neq k$. Тогда внутренняя сумма содержит два слагаемых:

$$\sum_{j=1}^n a_{rj}^{(0)} u_{ji}^{(1)} = a_{ri}^{(0)} u_{ii}^{(1)} + a_{rk}^{(0)} u_{ki}^{(1)} = a_{ri}^{(0)} c_1 + a_{rk}^{(0)} s_1,$$

а внешняя – одно:

$$a_{pi}^{(1)} = u_{pp}^{(1)T} (a_{pi}^{(0)} c_1 + a_{pk}^{(0)} s_1) = a_{pi}^{(0)} c_1 + a_{pk}^{(0)} s_1.$$

Повторяя рассуждения для $m = k$, получим

$$\sum_{j=1}^n a_{rj}^{(0)} u_{jk}^{(1)} = a_{rk}^{(0)} u_{kk}^{(1)} + a_{ri}^{(0)} u_{ik}^{(1)} = a_{rk}^{(0)} c_1 - a_{ri}^{(0)} s_1,$$

$$a_{pk}^{(1)} = u_{pp}^{(1)T} (a_{pk}^{(0)} c_1 - a_{pi}^{(0)} s_1) = a_{pk}^{(0)} c_1 - a_{pi}^{(0)} s_1.$$

Заметим, что

$$(a_{pi}^{(1)})^2 + (a_{pk}^{(1)})^2 = (a_{pi}^{(0)})^2 + (a_{pk}^{(0)})^2,$$

т.е. сумма квадратов элементов отмеченных столбцов, стоящих в одной (не отмеченной) строке, не меняется. В силу симметрии матриц не меняется и сумма квадратов элементов отмеченных строк, стоящих в одном (не отмеченном) столбце.

Осталось найти элемент $a_{ik}^{(1)}$, получающийся на месте ведущего элемента первого шага:

$$\begin{aligned} a_{ik}^{(1)} &= \sum_{r=1}^n u_{ir}^{(1)T} \sum_{j=1}^n a_{rj}^{(0)} u_{jk}^{(1)} = \\ &= \sum_{r=1}^n u_{ir}^{(1)T} (a_{ri}^{(0)} u_{ik}^{(1)} + a_{rk}^{(0)} u_{kk}^{(1)}) = \sum_{r=1}^n u_{ir}^{(1)T} (-a_{ri}^{(0)} s_1 + a_{rk}^{(0)} c_1) = \\ &= u_{ii}^{(1)} (-a_{ii}^{(0)} s_1 + a_{ik}^{(0)} c_1) + u_{ik}^{(1)} (-a_{ki}^{(0)} s_1 + a_{kk}^{(0)} c_1) = \\ &= -a_{ii}^{(0)} c_1 s_1 + a_{ik}^{(0)} c_1^2 - a_{ki}^{(0)} s_1^2 + a_{kk}^{(0)} c_1 s_1. \end{aligned}$$

Приравняв полученное выражение нулю и вспоминая, что $c_1 = \cos(\varphi_1)$, а $s_1 = \sin(\varphi_1)$, получим уравнение для определения φ_1 :

$$(\cos^2(\varphi_1) - \sin^2(\varphi_1)) a_{ik}^{(0)} = (a_{ii}^{(0)} - a_{kk}^{(0)}) \cos(\varphi_1) \sin(\varphi_1)$$

(здесь учтено, что $a_{ik}^{(0)} = a_{ki}^{(0)}$). По условию $a_{ik}^{(0)} \neq 0$ (как наибольший по модулю внедиагональный элемент), и это уравнение приводится к виду

$$\operatorname{ctg}(2\varphi_1) = (a_{ii}^{(0)} - a_{ik}^{(0)}) / 2a_{ik}^{(0)}.$$

Найдя из этого уравнения φ_1 , получим матрицу $A^{(1)}$, которая унитарно подобна матрице $A^{(0)}$ и обладает следующими свойствами:

1. $a_{ik}^{(1)} = a_{ki}^{(1)} = 0$.

2. Сумма квадратов остальных *внедиагональных* элементов не изменилась.

На втором шаге алгоритма Якоби мы сконструируем матрицу $A^{(2)} = U^{(2)*}A^{(1)}U^{(2)}$, унитарно подобную $A^{(1)}$ (а, следовательно, и $A^{(0)}$), у которой будут равны нулю ведущие элементы матрицы $A^{(1)}$.

Вообще, $A^{(p)} = V^{(p)*}AV^{(p)}$, где $V^{(p)} = U^{(1)} \cdot \dots \cdot U^{(p)}$.

Здесь можно было бы поставить слова "и так далее но..." к сожалению, на втором шаге те элементы, которые на первом шаге были обнулены, вообще говоря, станут снова отличными от нуля! Поэтому, в отличие, скажем, от алгоритма Гаусса–Йордана, процесс преобразования по алгоритму Якоби, вообще говоря, бесконечен. Однако сейчас мы покажем, что сумма квадратов внедиагональных элементов на каждом шаге алгоритма уменьшается.

Обозначим $Q^{(p)} = \sum_{r \neq m} (a_{rm}^{(p)})^2$. Тогда в силу свойств **1** и **2**

$$Q^{(p+1)} = Q^{(p)} - 2(a_{ik}^{(p)})^2 = Q^{(p)} \cdot \left(1 - 2 \frac{(a_{ik}^{(p)})^2}{Q^{(p)}}\right).$$

Но $(a_{ik}^{(p)})^2 = \max_{r \neq m} (a_{rm}^{(p)})^2$. Поэтому $Q^{(p)} \leq (a_{ik}^{(p)})^2 \cdot n(n-1)$, где n – порядок матрицы A . Отсюда

$$\frac{(a_{ik}^{(p)})^2}{Q^{(p)}} \geq \frac{1}{n(n-1)} \quad \text{и} \quad Q^{(p+1)} \leq Q^{(p)} \cdot \left(1 - \frac{2}{n(n-1)}\right).$$

Полученное неравенство показывает, что при итерациях по методу Якоби сумма квадратов внедиагональных элементов матрицы убывает не медленнее, чем геометрическая прогрессия:

$$Q^{(p)} \leq Q^{(0)} \cdot \left(1 - \frac{2}{n(n-1)}\right)^p,$$

и, следовательно, может быть сделана как угодно малой.

Итак, мы получили последовательность матриц $A^{(p)}$, унитарно подобных матрице A и приближающихся с ростом p к диагональной матрице. Поэтому можно ожидать, что диагональные элементы матриц $A^{(p)}$ с ростом p приближаются к собственным числам матрицы A .

Действительно, можно показать, что в интервале

$$\left[a_{ii}^{(p)} - \sqrt{Q^{(p)}}, a_{ii}^{(p)} + \sqrt{Q^{(p)}} \right]$$

содержится хотя бы одно собственное число матрицы A . Естественно назвать этот интервал *оценкой собственного числа*. Длина интервала оценки, очевидно, может быть сделана как угодно малой при достаточном количестве итераций.

Матрицы $V^{(p)} = U^{(1)} \cdot \dots \cdot U^{(p)}$ унитарны, и их столбцы являются приближениями для собственных векторов матрицы A .

Терминологическое замечание. Матрица $U^{(1)}$ осуществляет поворот на угол φ_1 в плоскости x_iOx_k , проходящей через i -ю и k -ю координатные оси в пространстве \mathbb{R}^n (сравните с примером в п.7.6). Матрица $V^{(p)}$ – произведение нескольких матриц плоских вращений. Поэтому метод Якоби иногда называют *методом вращений*.

Эффективные численные алгоритмы, реализованные в средах конечного пользователя и в библиотеках стандартных Фортран-программ, обеспечивают решение полной проблемы собственных значений для *самосопряженных* матриц с машинной точностью.

Серьезное предупреждение. В случае *несамосопряженных* матриц ситуация осложняется. Как было показано, такие матрицы могут и не иметь полного набора линейно независимых собственных векторов. Реализованные в средах конечного пользователя и в библиотеках стандартных Фортран-программ алгоритмы решения полной проблемы собственных значений для произвольных матриц *не гарантируют* получение результата. Мы настоятельно рекомендуем следовать в этом случае совету Хемминга⁶⁰: не жечь зря машинное время, а обращаться за консультацией к специалистам.

⁶⁰Ричард Уэсли ХЕММИНГ (R.W. Hamming, 1915-1998), американский математик, участник Манхэттенского проекта, автор фундаментальных результатов в численном анализе, теории информации, теории кодирования ("код Хемминга"), теории цифровых фильтров. В 1988 г. IEEE учредил медаль в его честь.

9. ПРОСТЕЙШИЕ ФУНКЦИОНАЛЫ НА ПРОСТРАНСТВАХ \mathbb{C}^n И \mathbb{R}^n

9.1. Линейные формы

В п.6.1 было показано, что всякая матрица A размера $m \times n$ порождает линейное отображение \mathbb{C}^n в \mathbb{C}^m : $x \rightarrow Ax$. В частном случае, когда $m = 1$, значения этого отображения – комплексные числа. Такое отображение называют *линейным функционалом (линейной формой)*.

Итак, $(1 \times n)$ -матрица (матрица-строка) порождает линейный функционал – отображение \mathbb{C}^n в \mathbb{C} .

Пусть $A = [a_1, \dots, a_n]$. Тогда $Ax = a_1x_1 + \dots + a_nx_n$.

Введем вектор-столбец $a = A^* = [\bar{a}_1, \dots, \bar{a}_n]^T$. Тогда наш линейный функционал может быть записан в терминах скалярного произведения

$$Ax = a^*x = \langle x, a \rangle.$$

Этот способ записи мы и будем, как правило, использовать.

Замечание. Вектор с вещественными компонентами, очевидно, порождает также вещественный линейный функционал на \mathbb{R}^n .

Рассмотрим теперь геометрическую интерпретацию линейной формы $y = \langle x, a \rangle$, заданной на \mathbb{R}^n ($a \in \mathbb{R}^n$, $n = 1, 2, 3$).

Если $a = \theta$, то $y \equiv 0$. Поэтому в дальнейшем мы считаем, что $a \neq \theta$.

Для $n = 1$ $a = [a]$, $x = [x]$, $y = ax$.

График этой формы – прямая на плоскости, проходящая через начало координат.

Для $n = 2$ $a = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$, $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $y = \langle x, a \rangle = a_1x_1 + a_2x_2$.

График этой формы – плоскость, проходящая через начало координат.

Полезно рассмотреть также *линии уровня* этого функционала, т.е. линии в \mathbb{R}^2 , на которых функционал сохраняет постоянное значение. Полагая $y = const$, получим уравнение

$$a_1x_1 + a_2x_2 = const. \tag{9.1.1}$$

Как известно, это уравнение определяет прямую.

Итак, график вещественного линейного функционала, заданного на \mathbb{R}^2 , – это плоскость, проходящая через начало координат, а его линии уровня образуют однопараметрическое семейство прямых на плоскости.

Рассмотрим прямую из этого семейства

$$\langle x, a \rangle = c, \quad c \in \mathbb{R}. \quad (9.1.2)$$

Зафиксируем на ней точку $x^{(0)}$. Вычитая из (9.1.2) равенство $\langle x^{(0)}, a \rangle = c$, получим $\langle x - x^{(0)}, a \rangle = 0$. Таким образом, векторы $x - x^{(0)}$ и a ортогональны, и соответствующие им направленные отрезки перпендикулярны.

Но отрезок $\overrightarrow{x - x^{(0)}}$, очевидно, параллелен нашей прямой. Поэтому все прямые семейства перпендикулярны \vec{a} , и, следовательно, параллельны между собой (рис.9.1).

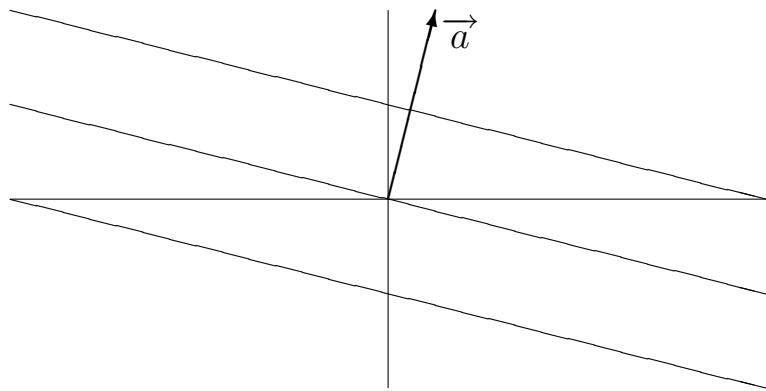


Рис.9.1

Замечания. 1. Как известно из школьного курса, любая прямая на плоскости задается уравнением (9.1.1). Поэтому любая прямая на плоскости является линией уровня некоторого линейного функционала.

2. Множество точек \mathbb{R}^2 , удовлетворяющих линейному неравенству $\langle x, a \rangle \leq c$, очевидно, есть объединение линий уровня $\langle x, a \rangle = \gamma$ при любых $\gamma \leq c$. Это одна из двух *полуплоскостей*, на которые прямая $\langle x, a \rangle = c$ делит плоскость. Множество точек, удовлетворяющих неравенству $\langle x, a \rangle \geq c$, образует вторую полуплоскость (рис.9.2).

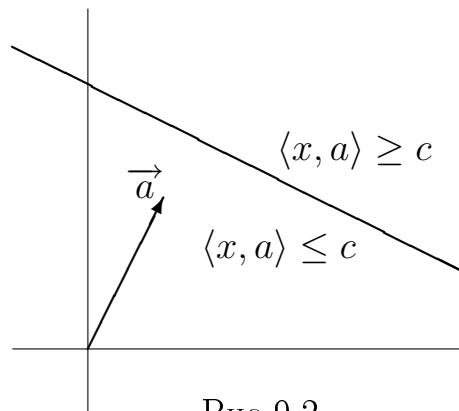


Рис.9.2

Для $n = 3$

$$a = [a_1, a_2, a_3]^T, \quad x = [x_1, x_2, x_3]^T, \quad y = \langle x, a \rangle = a_1 x_1 + a_2 x_2 + a_3 x_3.$$

График этого функционала построить невозможно, ибо три отпущенные нам природой оси декартовой системы координат заняты значениями компонент вектора x , и значения функционала давать уже некуда. Приходится ограничиться рассмотрением его *поверхностей уровня*, уравнения которых получаем, фиксируя значения функционала. Полагая $y = \text{const}$, имеем

$$\langle x, a \rangle = a_1 x_1 + a_2 x_2 + a_3 x_3 = c$$

— уравнение плоскости.

Итак, поверхности уровня нашего функционала образуют однопараметрическое семейство плоскостей.

Зафиксировав на одной из плоскостей этого семейства точку $x^{(0)}$, для любой другой точки x этой плоскости имеем $\langle x, a \rangle = c = \langle x^{(0)}, a \rangle$. Отсюда $\langle x - x^{(0)}, a \rangle = 0$ и, следовательно, направленный отрезок $\overrightarrow{x - x^{(0)}}$ перпендикулярен \overrightarrow{a} .

Поскольку x и $x^{(0)}$ лежат в нашей плоскости, отрезок $\overrightarrow{x - x^{(0)}}$ компланарен ей. Следовательно, эта плоскость перпендикулярна \overrightarrow{a} , и, значит, все плоскости семейства параллельны между собой.

Как и в двумерном случае, каждая плоскость в \mathbb{R}^3 является поверхностью уровня некоторого линейного функционала; множество точек, удовлетворяющих линейному неравенству $\langle x, a \rangle \leq c$, и множество точек, удовлетворяющих линейному неравенству $\langle x, a \rangle \geq c$, образуют два *полупространства*, разделяемых плоскостью $\langle x, a \rangle = c$.

9.2. Квадратичные формы

Пусть A — самосопряженная матрица порядка n , $x \in \mathbb{C}^n$. Тогда по формуле (8.1.1) $\langle Ax, x \rangle = \overline{\langle x, Ax \rangle}$, а так как по свойству 2 скалярного произведения $\langle Ax, x \rangle = \overline{\langle x, Ax \rangle}$, то $\langle Ax, x \rangle \in \mathbb{R}$.

Определение. Пусть A — самосопряженная матрица порядка n . вещественная числовая функция, заданная на \mathbb{C}^n правилом $x \rightarrow \langle Ax, x \rangle$, называется *квадратичной формой* (*квадратичным функционалом*).

Запишем квадратичную форму через координаты вектора x :

$$(Ax)_i = \sum_{j=1}^n a_{ij} x_j;$$

$$\langle Ax, x \rangle = \sum_{i=1}^n (Ax)_i \bar{x}_i = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_j \bar{x}_i = \sum_{i=1}^n a_{ii} |x_i|^2 + \sum_{j \neq i} a_{ij} x_i \bar{x}_j.$$

Наличие слагаемых с попарными произведениями координат затрудняет исследование квадратичной формы. Покажем, что от этого затруднения можно избавиться, если использовать для представления вектора ортонормированный собственный базис самосопряженной (!) матрицы A .

Как известно (п.8.1), $A = SAS^*$, где $\Lambda = \text{diag}[\lambda_1, \dots, \lambda_n]$ – диагональная матрица, диагональ которой состоит из собственных чисел матрицы A , $S = [s^{(1)}, \dots, s^{(n)}]$ – унитарная матрица, столбцы которой – соответствующие этим собственным числам ортонормированные собственные векторы (собственный базис матрицы A в \mathbb{C}^n).

Разложим вектор x по этому базису:

$$x = \alpha_1 s^{(1)} + \dots + \alpha_n s^{(n)} \quad \text{или} \quad x = S\alpha, \quad \text{где} \quad \alpha = [\alpha_1, \dots, \alpha_n]^T.$$

Отсюда

$$\langle Ax, x \rangle = x^* Ax = (S\alpha)^* A(S\alpha) = \alpha^* (S^* AS)\alpha = \alpha^* \Lambda \alpha = \langle \Lambda \alpha, \alpha \rangle,$$

или, в координатной записи,

$$\langle Ax, x \rangle = \sum_{j=1}^n \lambda_j |\alpha_j|^2. \tag{9.2.1}$$

Выражение (9.2.1) называется *каноническим представлением* квадратичной формы.

При исследовании квадратичной формы, как явствует из (9.2.1), важную роль играют собственные числа матрицы A . Введем в связи с этим несколько полезных терминов.

Определение. Если все собственные числа матрицы A положительны (отрицательны), то вследствие (9.2.1) квадратичная форма положительна (отрицательна) на \mathbb{C}^n , за исключением нулевого вектора, на котором она равна нулю. Такую квадратичную форму (как и ее матрицу) называют *положительно определенной* (*отрицательно определенной*).

Если все собственные числа матрицы A неотрицательны (неположительны), то и соответствующая квадратичная форма всюду неотрицательна (неположительна). Такую квадратичную форму (как и ее матрицу) называют *неотрицательно определенной* (*неположительно определенной*).

Пример. Пусть B – произвольная $(m \times n)$ -матрица. Рассмотрим матрицу Грама $G_B = B^*B$ (см. п.7.3). Для $x \in \mathbb{C}^n$ имеем

$$\langle G_Bx, x \rangle = \langle B^*Bx, x \rangle = \langle Bx, Bx \rangle = \|Bx\|^2 \geq 0.$$

Таким образом, матрица Грама любого набора векторов неотрицательно определена.

Если векторы (столбцы матрицы B) линейно независимы, то по свойству 2 матрицы Грама $\det(G_B) \neq 0$. Поэтому ее собственные числа не могут равняться нулю и, следовательно, положительны. Таким образом, матрица Грама линейно независимого набора векторов положительно определена.

Если среди собственных чисел матрицы A есть и положительные, и отрицательные, то соответствующая квадратичная форма принимает как положительные, так и отрицательные значения, и называется *знакопеременной*.

Докажем теперь одно важное для приложений неравенство:

$$\lambda_{\min}(A)\|x\|^2 \leq \langle Ax, x \rangle \leq \lambda_{\max}(A)\|x\|^2. \quad (9.2.2)$$

Действительно, из (9.2.1) имеем

$$\begin{aligned} \langle Ax, x \rangle &= \sum_{k=1}^n \lambda_k |\alpha_k|^2 \leq \sum_{k=1}^n \lambda_{\max} |\alpha_k|^2 = \lambda_{\max} \sum_{k=1}^n |\alpha_k|^2 = \lambda_{\max} \|\alpha\|^2; \\ \langle Ax, x \rangle &= \sum_{k=1}^n \lambda_k |\alpha_k|^2 \geq \sum_{k=1}^n \lambda_{\min} |\alpha_k|^2 = \lambda_{\min} \sum_{k=1}^n |\alpha_k|^2 = \lambda_{\min} \|\alpha\|^2. \end{aligned}$$

Остается только заметить, что умножение на унитарную матрицу не меняет норму вектора. Следовательно, $\|x\| = \|S\alpha\| = \|\alpha\|$, и (9.2.2) доказано. ■

Определенная для любого ненулевого вектора $x \in \mathbb{C}^n$ функция

$$x \rightarrow \frac{\langle Ax, x \rangle}{\|x\|^2},$$

где A – самосопряженная $(n \times n)$ -матрица, называется *отношением Рэлея*⁶¹. Из (9.2.2) следует, что значения отношения Рэлея заключены между наименьшим и наибольшим собственными числами матрицы A .

⁶¹Джон Вильям СТРЕТТ, барон РЭЛЕЙ (J.W. Rayleigh, 1842-1919) – английский физик и математик, президент Лондонского Королевского общества, лауреат Нобелевской премии.

9.3. Геометрическая интерпретация квадратичных форм

В этом пункте рассматриваются квадратичные формы с вещественной симметричной матрицей, заданные на \mathbb{R}^n , $n = 1, 2, 3$.

Замечание. Легко видеть, что собственные векторы вещественной симметричной матрицы вещественны. Поэтому координаты вектора из \mathbb{R}^n в собственном базисе такой матрицы тоже вещественны.

Далее мы считаем, что матрица A ненулевая ("исследование" случая $A = \Theta$ тривиально – квадратичная форма тождественно равна нулю).

1. Квадратичная форма на $\mathbb{R}^1 = \mathbb{R}$:

$$A = [a] – \text{матрица 1-го порядка, } y = \langle Ax, x \rangle = ax^2.$$

График этой квадратичной формы – парабола (на рис.9.3 $a > 0$).

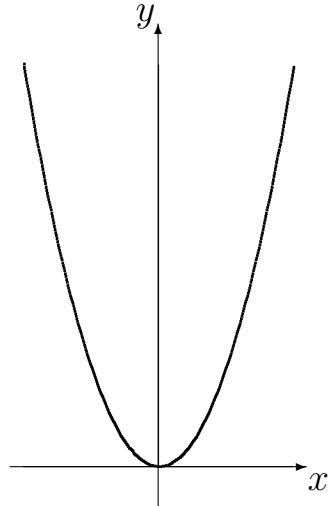


Рис.9.3

2. Квадратичная форма на \mathbb{R}^2 :

Условимся сразу записывать квадратичную форму в ортонормированном собственном базисе ее матрицы. Тогда

$$y = \langle Ax, x \rangle = \lambda_1 x_1^2 + \lambda_2 x_2^2.$$

График этого функционала – поверхность в \mathbb{R}^3 . Для исследования вида этой поверхности рассмотрим линии уровня квадратичной формы.

а) Пусть квадратичная форма *положительно определена* ($\lambda_1 > 0$, $\lambda_2 > 0$). Тогда, положив $y = c > 0$ (значение формы всюду неотрицательно и равно нулю только в начале координат), получим уравнение линии, во всех точках которой квадратичная форма равна c :

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 = c, \quad \text{или} \quad \frac{x_1^2}{c/\lambda_1} + \frac{x_2^2}{c/\lambda_2} = 1.$$

Это уравнение эллипса с полуосами $\sqrt{c/\lambda_1}$ и $\sqrt{c/\lambda_2}$ (рис.9.4).

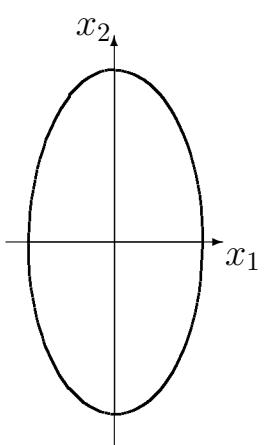


Рис.9.4

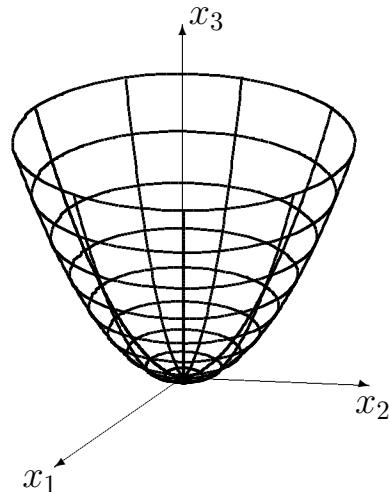


Рис.9.5

Таким образом, сечение нашей поверхности горизонтальной плоскостью есть эллипс, и полуоси этого эллипса неограниченно растут по мере удаления секущей плоскости от начала координат. Если учесть, что сечения поверхности вертикальными координатными плоскостями представляют собой параболы $y = \lambda_1 x_1^2$ и $y = \lambda_2 x_2^2$, то форма поверхности станет очевидной. Эта поверхность называется *эллиптическим параболоидом* (рис.9.5).

Если квадратичная форма отрицательно определена, то ее график, очевидно, также является эллиптическим параболоидом, но перевернутым "вверх ногами".

При $\lambda_1 = \lambda_2$ в сечениях поверхности горизонтальными плоскостями получаются окружности. Такой параболоид может быть получен вращением параболы $y = \lambda_1 x_1^2$, лежащей в плоскости x_1Oy , вокруг оси Oy . Он называется *параболоидом вращения*.

б) Пусть квадратичная форма знакопеременна. Примем для определенности, что $\lambda_1 > 0$, $\lambda_2 < 0$. Уравнение линии уровня $y = c$ имеет вид

$$\lambda_1 \cdot x_1^2 - |\lambda_2| \cdot x_2^2 = c.$$

При $c > 0$ (сечение горизонтальной плоскостью, расположенной над координатной) это гипербола

$$\frac{x_1^2}{c/\lambda_1} - \frac{x_1^2}{c/|\lambda_2|} = 1$$

с полуосами $\sqrt{c/\lambda_1}$ и $\sqrt{c/|\lambda_2|}$, а при $c < 0$ (сечение горизонтальной плоскостью, расположенной под координатной) – гипербола

$$-\frac{x_1^2}{|c|/\lambda_1} + \frac{x_1^2}{c/\lambda_2} = 1$$

с полуосами $\sqrt{|c|/\lambda_1}$ и $\sqrt{c/\lambda_2}$.

В сечении горизонтальной координатной плоскостью ($c = 0$) получается линия уровня $\lambda_1 \cdot x_1^2 = |\lambda_2| \cdot x_2^2$, представляющая собой пару пересекающихся прямых

$$x_2 = \pm \sqrt{\frac{\lambda_1}{|\lambda_2|}} x_1,$$

разделяющих два семейства гипербол (рис.9.6).

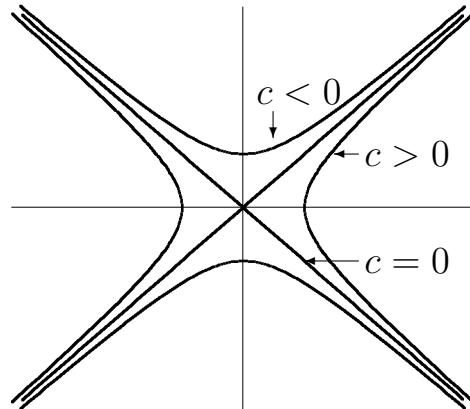


Рис.9.6

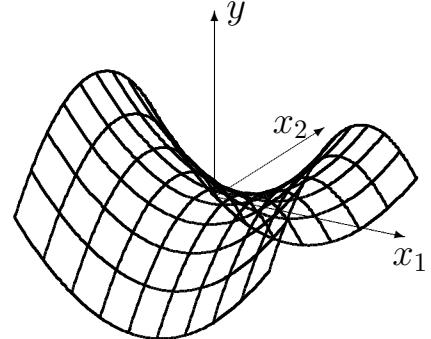


Рис.9.7

Сечения вертикальными координатными плоскостями – параболы $y = \lambda_1 x_1^2$ и $y = -|\lambda_2| x_2^2$. График знакопеременной квадратичной формы имеет вид бесконечного "седла" и называется *гиперболическим параболоидом* (рис.9.7).

в) При наличии у матрицы нулевого собственного числа (пусть, например, $\lambda_1 = 0$) график квадратичной формы $y = \lambda_2 x_2^2$ есть поверхность, во всех сечениях которой вертикальными плоскостями, перпендикулярными осям Ox_1 , получается одна и та же парабола. Эта поверхность называется *параболическим цилиндром* (рис.9.8).

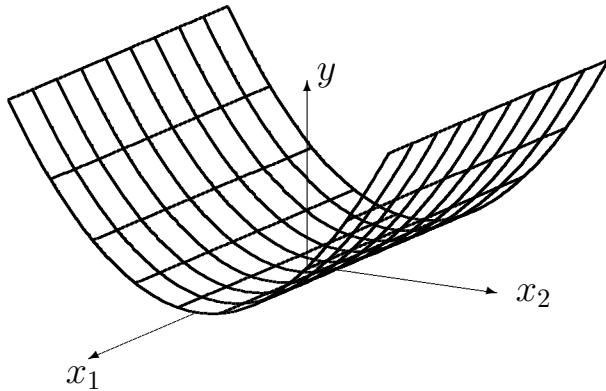


Рис.9.8

Образующая цилиндра параллельна оси Ox_1 . Линии уровня квадратичной формы – это пары параллельных прямых $x_2 = \pm\sqrt{c/\lambda_2}$ (знак c совпадает со знаком λ_2).

3. Квадратичная форма на \mathbb{R}^3 .

Записывая эту форму по-прежнему в собственном базисе матрицы A , получим

$$y = \langle Ax, x \rangle = \lambda_1 x_1^2 + \lambda_2 x_2^2 + \lambda_3 x_3^2.$$

График этой квадратичной формы, очевидно, изобразить невозможно, на что указывалось уже при рассмотрении линейных форм. Рассмотрим поверхности уровня.

а) Если квадратичная форма положительно (отрицательно) определена, то задавая положительное (отрицательное) ее значение $y = c$, получим уравнение поверхности уровня

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 + \lambda_3 x_3^2 = c, \quad \text{или} \quad \frac{x_1^2}{c/\lambda_1} + \frac{x_2^2}{c/\lambda_2} + \frac{x_3^2}{c/\lambda_3} = 1.$$

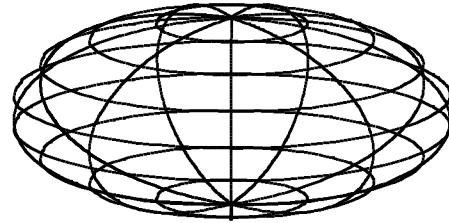


Рис.9.9

Это уравнение эллипсоида с полуосами $\sqrt{c/\lambda_1}$, $\sqrt{c/\lambda_2}$ и $\sqrt{c/\lambda_3}$ (рис.9.9). В сечениях эллипса плоскостями, параллельными координатным, получаются эллипсы (проверьте это!).

Если две из трех полуосей равны между собой (например, $\lambda_1 = \lambda_2$), то эллипсоид может быть получен вращением эллипса $\lambda_1 x_1^2 + \lambda_3 x_3^2 = c$, лежащего в координатной плоскости $x_1 O x_3$, вокруг оси $O x_1$. Такой эллипсоид называется *эллипсоидом вращения*. Если же все полуоси равны, то поверхности уровня квадратичной формы – сферы.

б) Если квадратичная форма знакопеременна, и среди ее собственных чисел нет нуля, то будем считать, что $\lambda_1 > 0$, $\lambda_2 > 0$, $\lambda_3 < 0$. Рассмотрим сначала поверхности уровня, на которых эта квадратичная форма положительна. Уравнение такой поверхности имеет вид

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 - |\lambda_3| x_3^2 = c > 0, \quad \text{или} \quad \frac{x_1^2}{c/\lambda_1} + \frac{x_2^2}{c/\lambda_2} - \frac{x_3^2}{c/|\lambda_3|} = 1.$$

Это *однополосный гиперболоид* (рис.9.10). Проверьте, что в сечениях горизонтальными плоскостями получаются эллипсы (в частном случае – при $\lambda_1 = \lambda_2$ – окружности), а в сечениях координатными плоскостями $x_1 O x_3$ и $x_2 O x_3$ – гиперболы.

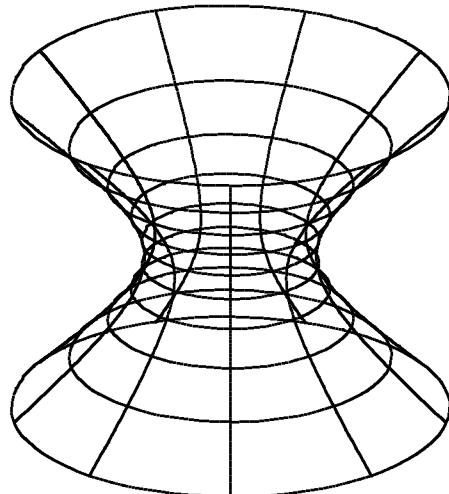


Рис.9.10

Поверхность уровня, на которой квадратичная форма отрицательна, – *двухполостный гиперболоид*⁶² (рис.9.11) – имеет уравнение

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 - |\lambda_3| x_3^2 = c < 0, \quad \text{или} \quad \frac{x_1^2}{|c|/\lambda_1} + \frac{x_2^2}{|c|/\lambda_2} - \frac{x_3^2}{c/\lambda_3} = -1.$$

Исследуйте ее сечения плоскостями, параллельными координатным.

⁶²По утверждению инженера Гарина, именно эта поверхность была взята им за основу при построении его смертоносного оружия, На самом же деле оптическим свойством, описанным в романе А.Н. Толстого, обладает параболоид вращения.

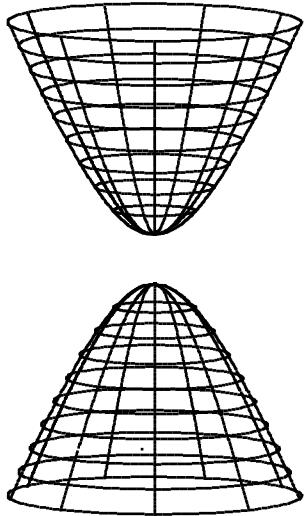


Рис.9.11

Уравнение поверхности, на которой квадратичная форма равна нулю,

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 - |\lambda_3| x_3^2 = 0.$$

Это *конус*, разделяющий семейства однополостных и двухполостных гиперболоидов (рис.9.12). Проверьте, что в сечениях горизонтальными плоскостями получаются эллипсы, а в сечениях координатными плоскостями x_1Ox_3 и x_2Ox_3 – пары пересекающихся прямых. Если $\lambda_1 = \lambda_2$, то конус называется *прямым круговым конусом*.

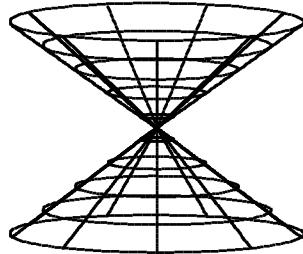


Рис.9.12

Прелагаем читателю самостоятельно убедиться, что случай одного положительного и двух отрицательных собственных чисел не дает поверхностей уровня, отличных от уже рассмотренных.

в) Пусть теперь одно собственное число (например, λ_3) равно нулю. Тогда остается рассмотреть случаи $\lambda_1 \cdot \lambda_2 > 0$ и $\lambda_1 \cdot \lambda_2 < 0$.

Если λ_1, λ_2 (а также c) положительны, то уравнение

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 = c$$

определяет эллиптический (при $\lambda_1 = \lambda_2$ – круговой) цилиндр с образующей, параллельной оси Ox_3 (рис.9.13). Такой же цилиндр получается, если λ_1, λ_2 и c отрицательны.

Если $\lambda_1 > 0, \lambda_2 < 0$, то уравнение

$$\lambda_1 x_1^2 - |\lambda_2| x_2^2 = c$$

определяет при $c \neq 0$ гиперболический цилиндр с образующей, параллельной оси Ox_3 (рис.9.14), а при $c = 0$ – пару пересекающихся плоскостей

$$\sqrt{\lambda_1} x_1 \pm \sqrt{|\lambda_3|} x_3 = 0.$$

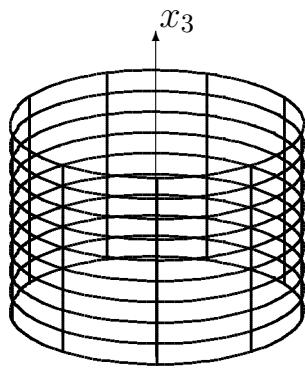


Рис.9.13

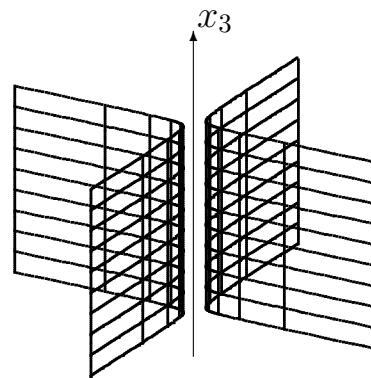


Рис.9.14

г) Если, наконец, равны нулю два собственных числа из трех, то уравнение поверхности уровня имеет вид

$$\lambda_1 x_1^2 = c \quad (\lambda_1 \cdot c \geq 0).$$

Это уравнение пары параллельных плоскостей $x_1 = \pm \sqrt{\frac{c}{\lambda_1}}$ (при $c = 0$ – одной плоскости).

Глава 10. ЛИНЕЙНЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

10.1. Сингулярные числа и сингулярные базисы матрицы

Пусть A – матрица размера $m \times n$. Как известно, она порождает линейный оператор $x \rightarrow Ax$, действующий из \mathbb{C}^n в \mathbb{C}^m .

Оператор $y \rightarrow A^*y$, порождаемый сопряженной к A матрицей A^* , действует из \mathbb{C}^m в \mathbb{C}^n .

Рассмотрим матрицы Грама

$$P = G_A = A^*A \quad \text{и} \quad Q = G_{A^*} = (A^*)^*A^* = AA^*.$$

Как известно, они эрмитовы и неотрицательно определены. Их порядки равны n и m соответственно.

Пусть $\lambda_1, \dots, \lambda_n \geq 0$ – собственные числа матрицы P , $v^{(1)}, \dots, v^{(n)}$ – соответствующие им ортонормированные собственные векторы в пространстве \mathbb{C}^n . Рассмотрим свойства образов этих векторов в \mathbb{C}^m .

Теорема 1. Векторы $Av^{(i)}$ и $Av^{(j)}$ ортогональны при $i \neq j$.

2. Если $\lambda_i = 0$, то $Av^{(i)} = \theta_m$.

3. Если $\lambda_i > 0$, то λ_i является также собственным числом матрицы Q , а $Av^{(i)}$ – соответствующий ему собственный вектор этой матрицы.

Доказательство. Обозначим $h^{(i)} = Av^{(i)}$. Тогда

$$\begin{aligned} \langle h^{(i)}, h^{(j)} \rangle &= \langle Av^{(i)}, Av^{(j)} \rangle = \langle A^*Av^{(i)}, v^{(j)} \rangle = \\ &= \langle Pv^{(i)}, v^{(j)} \rangle = \lambda_i \langle v^{(i)}, v^{(j)} \rangle. \end{aligned} \quad (10.1.1)$$

Полагая в (10.1.1) $i \neq j$, получим доказательство утверждения 1. Полагая там же $i = j$, получим

$$\|h^{(i)}\|^2 = \lambda_i \|v^{(i)}\|^2 = \lambda_i.$$

Следовательно, при $\lambda_i = 0$ $\|h^{(i)}\|^2 = 0$, т.е. $h^{(i)} = \theta_m$. Доказано утверждение 2. Далее,

$$\begin{aligned} Qh^{(i)} &= (AA^*)(Av^{(i)}) = A(A^*A)v^{(i)} = A(Pv^{(i)}) = \\ &= A(\lambda_i v^{(i)}) = \lambda_i (Av^{(i)}) = \lambda_i h^{(i)}. \end{aligned} \quad (10.1.2)$$

Из (10.1.2) следует утверждение 3 (при $\lambda_i \neq 0$ $h^{(i)} \neq \theta_m$). ■

Поменяв местами матрицы P и Q и повторив доказательство, получим

Следствие. Ненулевые собственные числа матриц P и Q попарно совпадают.

В дальнейшем будем считать, что r – общее количество ненулевых собственных чисел этих матриц (с учетом кратности). Тогда кратности нулевого собственного числа у матриц P и Q будут соответственно равны $n - r$ и $m - r$. В частности, при разных порядках матриц у "меньшей" может вообще не быть нулевых собственных чисел.

Упорядочим теперь положительные собственные числа по убыванию:

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0.$$

Введем векторы $u^{(j)}$, $j = 1, \dots, m$, следующим образом:

При $j \leq r$ $u^{(j)} = \frac{1}{\sqrt{\lambda_j}} h^{(j)}$ (в силу (10.1.3) $\|u^{(j)}\| = 1$);

При $r < j \leq m$ $u^{(j)}$ – ортонормированные собственные векторы матрицы Q , соответствующие нулевому собственному числу.

Из формулы (10.1.1) видно, что векторы образуют ортонормированный собственный базис матрицы Q . Далее, по построению

$$\begin{aligned} Av^{(j)} &= h^{(j)} = \sqrt{\lambda_j} u^{(j)} \quad (j \leq r); \\ Av^{(j)} &= \theta_m \quad (r < j \leq n). \end{aligned} \tag{10.1.4}$$

Аналогично,

$$\begin{aligned} A^*u^{(j)} &= \frac{1}{\sqrt{\lambda_j}} (A^*A)v^{(j)} = \sqrt{\lambda_j} v^{(j)} \quad (j \leq r); \\ A^*u^{(j)} &= \theta_n \quad (r < j \leq m). \end{aligned} \tag{10.1.5}$$

Введем теперь важное новое понятие.

Определение. Квадратные корни из общих ненулевых собственных чисел матриц $P = A^*A$ и $Q = AA^*$ ($\sigma_j = \sqrt{\lambda_j}$, $j = 1, \dots, r$) называются *сингулярными числами* матрицы A . Ортонормированные базисы, состоящие из собственных векторов матрицы P (в пространстве \mathbb{C}^n) и из собственных векторов матрицы Q (в пространстве \mathbb{C}^m) называются *сингулярными базисами* матрицы A (соответственно *правым* и *левым*).

Сведем векторы сингулярных базисов в унитарные матрицы

$$U = [u^{(1)}, \dots, u^{(m)}] \quad \text{и} \quad V = [v^{(1)}, \dots, v^{(n)}].$$

Теорема. $(m \times n)$ -матрица $\Sigma = U^*AV$ имеет структуру

$$\Sigma = U^*AV = \begin{bmatrix} \Sigma_r & \vdots & \mathbb{O}_{r \times (n-r)} \\ \dots & \ddots & \dots \\ \mathbb{O}_{(m-r) \times r} & \vdots & \mathbb{O}_{(m-r) \times (n-r)} \end{bmatrix}, \quad (10.1.6)$$

где $\Sigma_r = \text{diag}[\sigma_1, \dots, \sigma_r]$, а \mathbb{O} – нулевые матрицы, размеры которых обозначены в виде индексов.

Доказательство. Из (10.1.4) имеем

$$\begin{aligned} AV &= A \cdot [v^{(1)}, \dots, v^{(r)}, v^{(r+1)}, \dots, v^{(m)}] = \\ &= [Av^{(1)}, \dots, Av^{(r)}, Av^{(r+1)}, \dots, Av^{(m)}] = \\ &= [\sigma_1 u^{(1)}, \dots, \sigma_r u^{(r)}, \underbrace{\theta_m, \dots, \theta_m}_{(n-r)}]. \end{aligned}$$

Поэтому

$$\begin{aligned} U^*AV &= U^* \cdot [\sigma_1 u^{(1)}, \dots, \sigma_r u^{(r)}, \underbrace{\theta_m, \dots, \theta_m}_{(n-r)}] = \\ &= [u^{(1)}, \dots, u^{(r)}, u^{(r+1)}, \dots, u^{(m)}]^* \cdot [\sigma_1 u^{(1)}, \dots, \sigma_r u^{(r)}, \underbrace{\theta_m, \dots, \theta_m}_{(n-r)}] = \Sigma. \quad \blacksquare \end{aligned}$$

Равенство $U^*AV = \Sigma$ можно переписать в виде $A = U\Sigma V^*$. Такое представление матрицы называется ее *сингулярным разложением*.

Серьезное предупреждение. Сингулярное разложение матрицы несколько напоминает спектральное разложение эрмитовой матрицы. Однако в общем случае правый и левый сингулярные базисы не совпадают. Даже если матрица эрмитова, можно утверждать лишь, что ее сингулярные числа равны *модулям* ненулевых собственных чисел. Только для неотрицательно определенных эрмитовых матриц сингулярные числа совпадают с ненулевыми собственными числами, правый и левый сингулярные базисы одинаковы и совпадают с собственным базисом матрицы.

Пример.

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad A^* = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \end{bmatrix}, \quad A^*A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}, \quad AA^* = \begin{bmatrix} 2 & 2 & 0 \\ 2 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix};$$

$$\det(A^*A - \lambda I) = \lambda^2 - 6\lambda + 8; \quad \lambda_1(A^*A) = 4, \quad \lambda_2(A^*A) = 2.$$

Сингулярные числа матрицы A : $\sigma_1 = \sqrt{\lambda_1} = 2$, $\sigma_2 = \sqrt{\lambda_2} = \sqrt{2}$.

Найдем собственные векторы матрицы A^*A :

$$(A^*A - 4I)x = \theta \iff x = \alpha[1 \ 1]^T, \alpha \neq 0; \quad v^{(1)} = \frac{1}{\sqrt{2}}[1 \ 1]^T;$$

$$(A^*A - 2I)x = \theta \iff x = \beta[1 \ -1]^T, \beta \neq 0; \quad v^{(2)} = \frac{1}{\sqrt{2}}[1 \ -1]^T.$$

Два собственных числа матрицы AA^* совпадают с ненулевыми собственными числами матрицы A^*A , а третье обязано быть нулем. Итак, $\lambda_1(AA^*) = 4$, $\lambda_2(AA^*) = 2$, $\lambda_3(AA^*) = 0$.

Собственные векторы матрицы AA^* , соответствующие ее ненулевым собственным числам, найдем по формуле (10.1.4):

$$u^{(1)} = \frac{1}{\sigma_1}Av^{(1)} = \frac{1}{\sqrt{2}}[1 \ 1 \ 0]^T, \quad v^{(2)} = \frac{1}{\sigma_2}Av^{(2)} = \frac{1}{\sqrt{2}}[0 \ 0 \ 1]^T.$$

Третий собственный вектор найдем из соответствующей однородной системы и условия нормировки:

$$AA^*x = \theta \iff x = \alpha[1 \ -1 \ 0]^T, \alpha \neq 0; \quad u^{(3)} = \frac{1}{\sqrt{2}}[1 \ -1 \ 0]^T.$$

Сведем собственные векторы в матрицы

$$V = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}, \quad U = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \end{bmatrix}.$$

Выпишем сингулярное разложение $A = U\Sigma V^*$:

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 2 & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

10.2. Псевдорешение системы линейных алгебраических уравнений

Рассмотрим систему уравнений

$$Ax = b \tag{10.2.1}$$

с матрицей A размера $m \times n$, столбцом свободных членов $b \in \mathbb{C}^m$ и переменным вектором $x \in \mathbb{C}^n$.

Назовем *вектором невязок* системы (10.2.1) вектор

$$d(x) = b - Ax.$$

Тогда, очевидно, решением системы (10.2.1) будет такой вектор $x^{(0)} \in \mathbb{C}^n$, что $d(x^{(0)}) = \theta_m$.

При решении содержательных прикладных задач часто встречаются ситуации, в которых система линейных уравнений несовместна, хотя по "физическому смыслу" решение должно существовать. Объясняется это, как правило, тем, что коэффициенты и свободные члены системы, полученные из эксперимента, содержат погрешности.

В таких ситуациях разумно выбирать переменный вектор в системе (10.2.1) так, чтобы норма невязки, которую мы не можем сделать равной нулю, оказалась бы минимально возможной.

Определение. *Псевдорешением* системы линейных алгебраических уравнений (10.2.1) называется такой вектор $\tilde{x} \in \mathbb{C}^n$, что $\|d(\tilde{x})\| \leq \|d(x)\|$ для всех $x \in \mathbb{C}^n$, т.е. вектор, минимизирующий евклидову норму невязки.

Отметим, что в случае совместной системы любое ее решение $x^{(0)}$ является псевдорешением, поскольку $0 = \|d(x^{(0)})\| \leq \|d(x)\|$ для всех $x \in \mathbb{C}^n$.

Теорема. Всякая система линейных алгебраических уравнений имеет псевдорешение.

Доказательство. Пусть $A = U\Sigma V^*$ – сингулярное разложение матрицы коэффициентов системы (напомним, что U и V – унитарные матрицы сингулярных базисов – имеют порядки m и n соответственно, а $(m \times n)$ -матрица Σ имеет вид (10.1.6)).

Разложим вектор свободных членов b по сингулярному базису в \mathbb{C}^m :

$$b = c_1 u^{(1)} + \dots + c_m u^{(m)}, \quad \text{или} \quad b = Uc. \quad (10.2.2)$$

Вектор-псевдорешение \tilde{x} будем искать в виде разложения по сингулярному базису в \mathbb{C}^n :

$$\tilde{x} = \alpha_1 v^{(1)} + \dots + \alpha_n v^{(n)}, \quad \text{или} \quad \tilde{x} = V\alpha, \quad (10.2.3)$$

где α – новый искомый вектор.

Подставив (10.2.2) и (10.2.3) в уравнение (10.2.1), получим

$$AV\alpha = Uc.$$

Умножая это уравнение на $U^* = U^{-1}$ слева и учитывая, что $U^*AV = \Sigma$, имеем

$$\Sigma\alpha = c. \quad (10.2.4)$$

Покажем, что нормы невязок систем (10.2.1) и (10.2.4) равны, т.е. задача свелась к отысканию псевдорешения системы (10.2.4). Действительно, из (10.2.3) имеем $\alpha = V^*\tilde{x}$ и, следовательно,

$$b - A\tilde{x} = Uc - U\Sigma V^*\tilde{x} = U \cdot (c - \Sigma\alpha).$$

Умножение на унитарную матрицу сохраняет норму вектора. Поэтому

$$\|b - A\tilde{x}\| = \|U \cdot (c - \Sigma\alpha)\| = \|c - \Sigma\alpha\|.$$

Но

$$\Sigma\alpha = [\sigma_1\alpha_1, \dots, \sigma_r\alpha_r, \underbrace{0, \dots, 0}_{(m-r)}]^T,$$

и квадрат нормы невязки для системы (10.2.4) равен

$$\|c - \Sigma\alpha\|^2 = |c_1 - \sigma_1\alpha_1|^2 + \dots + |c_r - \sigma_r\alpha_r|^2 + |c_{r+1}|^2 + \dots + |c_m|^2. \quad (10.2.5)$$

Из (10.2.5) видно, что минимум нормы невязки достигается при

$$\alpha_j = \frac{c_j}{\sigma_j}, \quad j = 1, \dots, r, \quad (10.2.6)$$

и равен $(|c_{r+1}|^2 + \dots + |c_m|^2)^{1/2}$. Теорема доказана. ■

Рассмотрим частные случаи.

1. $r = n$. В этом случае равенства (10.2.6) однозначно определяют все компоненты вектора α . По формуле (10.2.3) получаем $\tilde{x} = V\alpha$ – единственное псевдорешение системы (10.2.1).

2. $r < n$. В этом случае из (10.2.6) определяются только первые r компонент вектора α . Однако формула (10.2.5) показывает, что оставшиеся компоненты не влияют на величину невязки и могут быть выбраны произвольно. В этом случае псевдорешение *не единствено*. *Обычно* полагают

$$\alpha_{r+1} = \dots = \alpha_n = 0. \quad (10.2.7)$$

Соответствующее псевдорешение $\tilde{x} = V\alpha$ называют *нормальным псевдорешением* системы (10.2.1).

Очевидно, всякая система линейных уравнений имеет единственное нормальное псевдорешение. Среди всех возможных псевдорешений оно выделяется наименьшей нормой.

3. $r = m$. В этом случае любое псевдорешение обеспечивает нулевую норму невязки, т.е. является *решением* системы (10.2.1).

Серьезное предупреждение. Не следует путать понятия "псевдорешение" и "приближенное решение" системы линейных алгебраических уравнений. О приближенном решении можно говорить только в случае совместной системы (вектор приближенного решения в каком-то смысле близок к существующему "точному" решению). Но в этом случае псевдорешение совпадает с решением.

В отличие от решения псевдорешение существует и у несовместной системы, когда говорить о приближенном решении бесмысленно, так как решение отсутствует и приближаться не к чему!

Терминологическое замечание. Псевдорешение системы линейных уравнений часто называют *решением в смысле наименьших квадратов*, а метод отыскания псевдорешения – *методом наименьших квадратов*. Это историческое название объясняется тем, что минимизируется квадрат евклидовой нормы невязки, который сводится к сумме квадратов модулей невязок всех уравнений системы.

Почему из всевозможных мер близости выбрана именно *евклидова* норма? Потому, что такой выбор приводит к простому алгоритму построения псевдорешения. Никаких "более глубоких" обоснований у этого метода нет. Заметим, что иногда пользуются и другими нормами.

Замечание. Геометрическая интерпретация псевдорешения весьма проста: это такой вектор $\tilde{x} \in \mathbb{C}^n$, который при умножении на матрицу коэффициентов системы переходит в вектор пространства \mathbb{C}^m , "ближайший" к вектору-свободному члену. При этом расстояние между векторами определяется по теореме Пифагора.

Пример. Найдем нормальное псевдорешение системы линейных уравнений

$$\begin{cases} x_1 + x_2 = 1 \\ x_1 + x_2 = 2, \\ x_1 - x_2 = 3 \end{cases}$$

или, в матричном виде,

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

Сингулярное разложение матрицы коэффициентов этой системы было получено в п.10.1. Используя его, перепишем систему:

$$\begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{\sqrt{2}}{2} & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 2 & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

Умножив это равенство слева на $U^* = U^{-1}$, получим

$$\begin{bmatrix} 2 & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \frac{3}{\sqrt{2}} \\ 3 \\ -\frac{1}{\sqrt{2}} \end{bmatrix},$$

или

$$2\alpha_1 = \frac{3}{\sqrt{2}}; \quad \sqrt{2}\alpha_2 = 3; \quad 0 = -\frac{1}{\sqrt{2}}.$$

Из первых двух уравнений имеем

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \frac{3}{2\sqrt{2}} \\ \frac{3}{\sqrt{2}} \end{bmatrix}.$$

Умножив обе части равенства на $V = (V^*)^{-1}$, получим нормальное псевдорешение системы $\tilde{x} = [9/4 \quad -3/4]^T$.

Рассмотрим геометрическую интерпретацию этого примера.

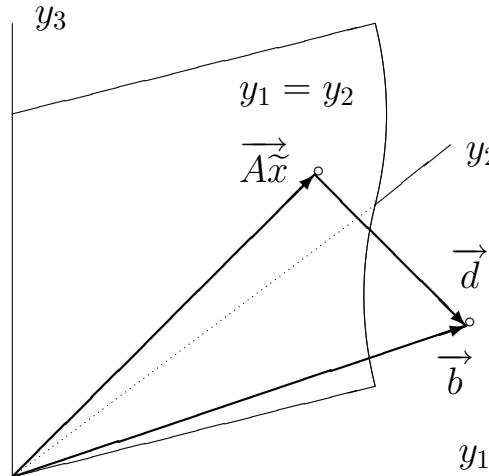


Рис.10.1

Пусть $y = [y_1, y_2, y_3]^T$ – образ вектора $x = [x_1, x_2]^T$ при отображении $y = Ax$. Тогда $y_1 = x_1 + x_2 = y_2$, т.е. множество значений этого отображения есть плоскость $y_1 = y_2$ в \mathbb{R}^3 (рис.10.1). Точка с координатами $(1, 2, 3)$ в этой плоскости не лежит, т.е. система уравнений не имеет решения. Ближайшая к этой точке точка плоскости $y_1 = y_2$ есть образ псевдорешения (см. Замечание):

$$A\tilde{x} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} 9/4 \\ -3/4 \end{bmatrix} = \begin{bmatrix} 3/2 \\ 3/2 \\ 3 \end{bmatrix}.$$

Вектор невязки $d(\tilde{x}) = [-\frac{1}{2}, \frac{1}{2}, 0]^T$, а его норма $\|d(\tilde{x})\| = \frac{1}{\sqrt{2}}$.

10.3. Псевдообратная матрица. Нормальные уравнения

В предыдущем пункте было построено псевдорешение системы (10.2.1). Запишем формулы для вычисления псевдорешения в матричной форме. Для этого введем $(n \times m)$ -матрицу следующей структуры

$$\Sigma^+ = \begin{bmatrix} \Sigma_r^+ & : & \mathbb{O}_{r \times (m-r)} \\ \dots & & \dots \\ \mathbb{O}_{(n-r) \times r} & : & \mathbb{O}_{(n-r) \times (m-r)} \end{bmatrix},$$

где $\Sigma_r^+ = diag \left[\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_r} \right]$, а \mathbb{O} – нулевые матрицы, размеры которых обозначены в виде индексов.

Нетрудно убедиться в том, что формулы (10.2.6) и (10.2.7) можно переписать в виде $\alpha = \Sigma^+ c$ (проверьте это!). Тогда нормальное псевдорешение системы (10.2.1) запишется в виде

$$\tilde{x} = V\alpha = V\Sigma^+ c = V\Sigma^+ U^* b.$$

Если матрица коэффициентов системы A *квадратная и обратимая*, то решение этой системы получается, как известно, умножением свободного члена слева на обратную матрицу: $x = A^{-1}b$.

Назовем матрицу $A^+ = V\Sigma^+ U^*$ *псевдообратной*. Тогда нормальное псевдорешение получается умножением свободного члена слева на псевдообратную матрицу:

$$\tilde{x} = A^+ b. \tag{10.3.1}$$

Замечание. Если $r = m = n$, то матрица Σ , очевидно, обратима, и $\Sigma^+ = \Sigma^{-1}$. Поэтому

$$A^+ = V\Sigma^+U^* = V\Sigma^{-1}U^{-1} = (U\Sigma V^{-1})^{-1} = (U\Sigma V^*)^{-1} = A^{-1},$$

т.е. если матрица обратима, то ее псевдообратная совпадает с ее обратной. Это согласуется с утверждением (см. п.10.2), что в этом случае псевдорешение единственно ($r = n$) и является решением системы ($r = m$).

Серьезное предупреждение. Формула (10.3.1) есть не более, чем удобная форма записи *результатата*. Так же как для нахождения решения системы незачем вычислять обратную матрицу, для нахождения нормального псевдорешения нет смысла вычислять псевдообратную матрицу, а следует пользоваться формулами п.10.2.

Докажем теперь, что $A^*AA^+ = A^*$. Действительно,

$$\begin{aligned} \Sigma \cdot \Sigma^+ &= \begin{bmatrix} \Sigma_r & : & \mathbb{O}_{r \times (n-r)} \\ \dots & & \dots \\ \mathbb{O}_{(m-r) \times r} & : & \mathbb{O}_{(m-r) \times (n-r)} \end{bmatrix} \cdot \begin{bmatrix} \Sigma_r^+ & : & \mathbb{O}_{r \times (m-r)} \\ \dots & & \dots \\ \mathbb{O}_{(n-r) \times r} & : & \mathbb{O}_{(n-r) \times (m-r)} \end{bmatrix} = \\ &= \begin{bmatrix} I_r & : & \mathbb{O}_{r \times (m-r)} \\ \dots & & \dots \\ \mathbb{O}_{(m-r) \times r} & : & \mathbb{O}_{(m-r) \times (m-r)} \end{bmatrix}; \\ \Sigma^* \cdot \Sigma \cdot \Sigma^+ &= \begin{bmatrix} \Sigma_r & : & \mathbb{O}_{r \times (n-r)} \\ \dots & & \dots \\ \mathbb{O}_{(m-r) \times r} & : & \mathbb{O}_{(m-r) \times (n-r)} \end{bmatrix} \cdot \begin{bmatrix} I_r & : & \mathbb{O}_{r \times (m-r)} \\ \dots & & \dots \\ \mathbb{O}_{(m-r) \times r} & : & \mathbb{O}_{(m-r) \times (m-r)} \end{bmatrix} = \Sigma^*. \end{aligned}$$

Поэтому

$$\begin{aligned} A^*AA^+ &= (U\Sigma V^*)^* \cdot (U\Sigma V^*) \cdot (V\Sigma^+U^*) = \\ &= V\Sigma^*(U^*U)\Sigma(V^*V)\Sigma^+U^* = V \cdot (\Sigma^*\Sigma\Sigma^+) \cdot U^* = V\Sigma^*U^* = A^*. \quad \blacksquare \end{aligned}$$

Умножив теперь обе части (10.3.1) слева на A^*A , получим

$$A^*A\tilde{x} = (A^*AA^+)b = A^*b.$$

Таким образом, *нормальное псевдорешение* системы (10.2.1) будет *решением* системы $(A^*A)x = A^*b$, (10.3.2)

получающейся из (10.2.1) умножением обеих частей слева на матрицу A^* . Уравнения (10.2.3) называются *нормальными уравнениями* метода наименьших квадратов.

Серьезное предупреждение. Возникает вопрос: для чего же было мучиться столько времени, вводить новые понятия "сингулярное разложение" и "псевдорешение"? Не проще ли построить систему нормальных уравнений и решить ее?

Оказывается, все не так просто.

Во-первых, никто не гарантирует нам невырожденности матрицы $G_A = A^*A$ (которая равносильна, как известно, линейной независимости столбцов матрицы A). Если матрица Грама окажется вырожденной, то нам все равно придется искать псевдорешение, но уже системы (10.3.2)!

Во-вторых, если даже матрица системы (10.3.2) не вырождена, то при переходе к нормальным уравнениям резко возрастают вычислительные трудности (см. главу 13).

Поэтому мы настоятельно рекомендуем *не пользоваться нормальными уравнениями* (тем более, что численно устойчивые алгоритмы сингулярного разложения и построения нормального псевдорешения реализованы и в средах конечного пользователя, и в библиотеках Фортрана).

Замечание. Отметим один случай, когда все-таки можно переходить к нормальным уравнениям. Если столбцы матрицы попарно ортогональны, то матрица Грама окажется диагональной, и решение системы (10.3.2) – нормальное псевдорешение системы (10.2.1) – выписывается в явном виде:

$$\tilde{x}_k = \frac{(A^*b)_k}{\|a^{(k)}\|^2} = \frac{\langle b, a^{(k)} \rangle}{\|a^{(k)}\|^2}, \quad k = 1, \dots, n. \quad (10.3.3)$$

Глава 11. СГЛАЖИВАНИЕ РЕЗУЛЬТАТОВ ИЗМЕРЕНИЙ МЕТОДОМ НАИМЕНЬШИХ КВАДРАТОВ

11.1. Одна содержательная задача

Описание проблемы, которой посвящена эта глава, мы начнем с простейшего примера – измерения сопротивления резистора методом амперметра и вольтметра. Метод этот, как известно, состоит в том, что одновременно измеряются: ток J , текущий через резистор, и U – падение напряжения на нем (рис.11.1).

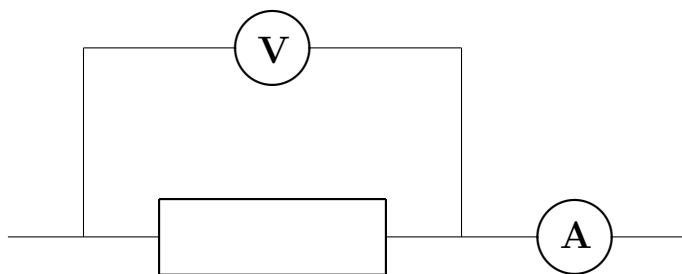


Рис.11.1

Если пренебречь током, текущим через вольтметр, то, в соответствии с законом Ома, искомое сопротивление находится по формуле $R = U/J$.

Дело, однако, усложняется тем, что, повторяя измерения, получают каждый раз новое значение сопротивления. Возможны два различных толкования этого экспериментального факта.

1. Сопротивление резистора не постоянно, а изменяется со временем.
2. Сопротивление постоянно, но данные измерений содержат ошибки (погрешности приборов, субъективные ошибки наблюдателя и пр.).

Следует ясно понимать, что
без выбора математической модели эксперимента
его результаты обрабатывать нельзя.

В нашей задаче мы *постулируем* неизменность сопротивления резистора во времени (выяснение правомерности такого постулата выходит, естественно, за рамки курса математики) и взваливаем ответственность за его наблюдающиеся изменения на погрешности. Тогда, проведя серию из n измерений, мы можем записать их результаты в виде

$$\begin{cases} J_1 R = U_1 \\ \dots \\ J_n R = U_n \end{cases}$$

т.е. в виде системы из n линейных уравнений с одной переменной – исключенным сопротивлением. Поскольку эта система, очевидно, несовместна, будем искать ее псевдорешение. Несмотря на простоту задачи, проделаем все операции подробно, чтобы еще раз продемонстрировать методику.

Запишем систему в виде

$$JR = U, \quad (11.1.1)$$

где $J = [J_1, \dots, J_n]^T$, $U = [U_1, \dots, U_n]^T$.

Матрица $J^*J = [J_1^2 + \dots + J_n^2]$ имеет размер 1×1 , а матрица

$$JJ^* = \begin{bmatrix} J_1^2 & J_1 J_2 & \dots & J_1 J_n \\ \dots & \dots & \dots & \dots \\ J_n J_1 & J_n J_2 & \dots & J_n^2 \end{bmatrix}$$

– размер $n \times n$.

Так как J^*J имеет размер 1×1 , сингулярное число оказывается единственным – это положительный корень уравнения

$$\det(J^*J - \sigma^2 I_1) = 0, \quad \text{или} \quad \sum_{k=1}^n J_k^2 - \sigma^2 = 0, \quad \text{т.е.} \quad \sigma = \left(\sum_{k=1}^n J_k^2 \right)^{1/2}.$$

Соответствующие сингулярные базисы находятся так: *правый* (состоящий из одного вектора) – путем решения однородной "системы"

$$(J^*J - \sigma^2 I_1)v^{(1)} = \theta_1, \quad \text{или} \quad \theta_1 \cdot v^{(1)} = \theta_1, \quad \text{т.е.} \quad v^{(1)} = [1]$$

(напомним, что сингулярный вектор берется нормированным). По известному из п.10.1 правилу находим соответствующий сингулярный вектор $w^{(1)} = \frac{1}{\sigma} J v^{(1)} = \frac{1}{\sigma} J$ и дополняем его до ортонормированного базиса.

Сингулярное разложение матрицы J имеет вид $J = W\Sigma V^*$, где $V^* = V = [1]$, $W = [w^{(1)}, \dots, w^n]$ – унитарная матрица порядка n , а $\Sigma = [\sigma, 0, \dots, 0]^T$ – столбец высоты n .

Матрица $\Sigma^+ = [1/\sigma, 0, \dots, 0]$ – строка ширины n . Поэтому

$$J^+ = V\Sigma^+W^* = \frac{1}{\sigma} w^{(1)*} = \frac{1}{\sigma} J^*.$$

Теперь можно по формуле (10.3.1) найти псевдорешение:

$$\tilde{R} = J^+U = \frac{1}{\sigma^2} J^*U = \frac{1}{\sigma^2} (J_1 U_1 + \dots + J_n U_n) = \frac{J_1 U_1 + \dots + J_n U_n}{J_1^2 + \dots + J_n^2}.$$

Внимательный читатель заметит, что проще было бы в уравнении (11.1.1) умножить обе части слева на J^* (перейти к нормальным уравнениям):

$$(J^* J) \tilde{R} = J^* U \implies \tilde{R} = \frac{J^* U}{J^* J} = \frac{J_1 U_1 + \dots + J_n U_n}{J_1^2 + \dots + J_n^2}.$$

Но мы, повторяя, хотели еще раз продемонстрировать методику построения псевдорешения с помощью сингулярного разложения матрицы системы.

Полученное число \tilde{R} можно истолковать как значение сопротивления резистора, "наилучшим образом согласующееся" сразу со всеми результатами измерения (при этом степень согласованности понимается так, как было сказано в замечании п.10.2). Можно полагать, что таким образом мы избавились от "случайных" погрешностей в результатах измерений – *сгладили* эти результаты.

Серьезное предупреждение. Подчеркнем еще раз, что приведенные выше рассуждения имеют смысл *только в рамках принятой математической модели* – сопротивление резистора предполагается постоянным во времени. Если же наша модель не верна, и сопротивление на самом деле изменялось во время измерений, то такое "сглаживание" превращается в искажение реально наблюдаемого явления.

11.2. Полиномиальное сглаживание

Рассмотрим (с меньшей конкретизацией содержательной постановки) еще одну распространенную прикладную задачу.

Информационно-измерительная система (ИИС) фиксирует в равнотстоящие моменты времени t_1, \dots, t_n значения некоторой измеряемой величины y_1, \dots, y_n .

Требуется (как обычно говорят прикладники) "подобрать какую-нибудь *простую и удобную* функцию, которая *хорошо описывала бы* полученные результаты".

Попытаемся придать точный смысл этой туманной фразе.

По-видимому слова "хорошо описывает" можно понимать только в одном смысле: поскольку ИИС никаких сведений об измеряемой величине, кроме пар чисел (t_i, y_i) , $i = 1, \dots, n$, не имеет, подобранная нами (мы будем называть ее *аппроксимирующей*) функция в точках t_1, \dots, t_n должна принимать значения y_1, \dots, y_n соответственно.

Второе требование – "простота" функции – обычно считается удовлетворенным, если предлагается полином не очень большой степени.

Итак, формулируем первый вариант постановки задачи: построить полином минимально возможной степени, который в заданных точках t_1, \dots, t_n принимает заданные значения y_1, \dots, y_n .

Это известная задача полиномиальной интерполяции. Ее решением будет (см. п.5.6) полином порядка n (степень его, естественно зависит от интерполируемой таблицы).

Обычно такая постановка задачи отвергается по двум причинам. Во-первых, при большом количестве измерений степень полинома оказывается также большой (он перестает быть "простой и удобной" функцией).

Во-вторых, известно, что результаты измерений всегда содержат погрешности, и естественно предполагать, что рост степени интерполярующего полинома при увеличении количества измерений объясняется стремлением этого полинома хорошо описывать ошибки!

Поэтому при обработке результатов измерений чаще применяется другая постановка задачи: заранее фиксируется порядок полинома (вопрос о выборе этого порядка лежит вне рамок нашего курса – он требует подробного рассмотрения содержательной задачи). Требование совпадения значений полинома в узлах со значениями измеренной величины заменяется требованием "достаточной их близости" (вопрос о достаточности достигаемой близости также требует рассмотрения содержательной постановки). Такой полином называют *сглаживающим*.

Итак, пусть m – *назначенный* порядок сглаживающего полинома ($m \leq n$). Для определения его коэффициентов получаем систему линейных уравнений

или $TP = y$, где

$$T = \begin{bmatrix} 1 & t_1 & \dots & t_1^{m-1} \\ \dots & \dots & \dots & \dots \\ 1 & t_n & \dots & t_n^{m-1} \end{bmatrix}$$

– матрица размера $n \times m$, $P = [p_1, \dots, p_m]^T$ – искомый столбец коэффициентов полинома, $y = [y_1, \dots, y_n]^T$ – столбец результатов измерений.

Эта система, как правило, бывает несовместной. Однако по постановке задачи совместность нам не обязательна. Поскольку требуется не совпадение значений полинома в узлах с результатами измерений, а лишь наивозможная их близость, будем искать псевдорешение этой системы.

Отметим, что целесообразно стандартизовать задачу, введя вместо времени t целочисленную переменную (номер измерения), связанную с t формулой $t_k = t_1 + (k - 1) \cdot \Delta t$, $k = 1, \dots, n$ (напомним, что моменты времени считаются равноотстоящими; Δt – шаг по времени). Тогда сглаживающий полином примет вид

$$s(k) = s_1 + s_2 k + \dots + s_m k^{m-1},$$

и столбец его коэффициентов s будет псевдорешением линейной системы $\tilde{T}S = y$, где

$$\tilde{T} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & 2^{m-1} \\ \dots & \dots & \dots & \dots \\ 1 & n & \dots & n^{m-1} \end{bmatrix}$$

– теперь уже стандартная (при фиксированных значениях n и m) матрица, для которой заранее может быть найдено сингулярное разложение.

Построенный сглаживающий полином служит для вычисления значений величины y в табличных точках t_1, \dots, t_n . Его применение позволяет хранить вместо n чисел y_1, \dots, y_n всего лишь m чисел s_1, \dots, s_m – коэффициенты полинома. При большой разнице между n и m экономия памяти может оказаться существенной.

Качество аппроксимации оценивается *среднеквадратической погрешностью*

$$\left(\frac{1}{n} \sum_{k=1}^n (s(k) - y_k)^2 \right)^{1/2}$$

(по построению сглаживающий полином минимизирует именно эту погрешность на множестве всех полиномов порядка m).

Серьезное предупреждение. Часто делаются попытки использовать сглаживающий полином для работы с ним *между* узлами таблицы. Не имея возможности запретить такого рода операции, хотим заранее снять с математики ответственность за возможное "качество" их результатов.

11.3. Сглаживание полиномами, ортогональными на сетке

Мы показали, что построение сглаживающего полинома на равномерной сетке сводится к нахождению нормального псевдорешения системы линейных уравнений

$$\tilde{T}s = y. \quad (11.3.1)$$

В п.10.3 указано, что нормальное псевдорешение этой системы является решением системы нормальных уравнений. Однако переход от системы (11.3.1) к системе нормальных уравнений невыгоден, как указано в том же п.10.3. Вот если бы столбцы матрицы \tilde{T} попарно ортогональны...

Но ведь мы знаем, как из линейно независимого набора векторов сделать ортогональный: следует применить алгоритм Грама–Шмидта!

Известно, что матрица Вандермонда

$$E = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & 2^{n-1} \\ \dots & \dots & \dots & \dots \\ 1 & n & \dots & n^{n-1} \end{bmatrix}$$

не вырождена. Следовательно, множество ее столбцов линейно независимо. Поэтому линейно независима и его часть – множество столбцов матрицы \tilde{T} . А любое линейно независимое множество векторов можно с помощью алгоритма Грама–Шмидта преобразовать в ортогональное.

Положим, согласно п.7.7, $F^{(1)} = E^{(1)} = [1, \dots, 1]^T$. Далее, положим $F^{(2)} = E^{(2)} - \alpha_{12}F^{(1)}$, где

$$\alpha_{12} = \frac{\langle E^{(2)}, F^{(1)} \rangle}{\langle F^{(1)}, F^{(1)} \rangle} = \frac{n+1}{2},$$

и т.д.

На языке полиномов этот процесс выглядит так.

Вектор $E^{(j)}$ – это значения полинома $e^{(j)}(t) = t^{j-1}$ на *стандартной сетке* $\{1, \dots, n\}$. Поэтому $F^{(j)}$ – значения на той же сетке полинома

$$f^{(j)}(t) = e^{(j)}(t) - \alpha_{1j}f^{(1)}(t) - \dots - \alpha_{j-1,j}f^{(j-1)}(t).$$

Например, $f^{(1)}(t) \equiv 1$, $f^{(2)}(t) = t - \frac{n+1}{2}$, и т.д.

Очевидно, что $f^{(j)}(t)$ – полином степени $j-1$ со старшим коэффициентом, равным единице. Условие $\langle F^{(i)}, F^{(j)} \rangle = 0$, $i \neq j$ на языке полиномов запишется так:

$$\sum_{k=1}^n f^{(i)}(k) \cdot \overline{f^{(j)}(k)} = 0.$$

Определение. Если $p^{(i)}, p^{(j)} \in \mathcal{P}_n$ (полиномы порядка n), то их *скалярным произведением на стандартной сетке* называется число

$$\langle p^{(i)}, p^{(j)} \rangle = \sum_{k=1}^n p^{(i)}(k) \cdot \overline{p^{(j)}(k)}. \quad (11.3.2)$$

Замечание. Очевидно, что $\langle p, p \rangle \geq 0$ для любого $p \in \mathcal{P}_n$. Если $\langle p, p \rangle = 0$, то, очевидно, $p(k) = 0$ при $k = 1, \dots, n$, т.е. полином порядка n имеет по крайней мере n корней. Но полином с такими свойствами только один – нулевой. Итак, если $\langle p, p \rangle = 0$, то $p(t) = \theta(t)$.

Читателю предоставляется возможность самому проверить, что введенное на пространстве \mathcal{P}_n по формуле (11.3.2) "скалярное произведение" удовлетворяет и остальным аксиомам скалярного произведения.

Естественно назвать полиномы $f^{(j)}$, $j = 1, \dots, n$, *полиномами, ортогональными на стандартной сетке*. Они, очевидно, образуют базис пространства \mathcal{P}_n (ортогональный в смысле скалярного произведения (11.3.2)).

Будем теперь искать сглаживающий полином в виде линейной комбинации полиномов построенного базиса:

$$p(t) = p_1 f^{(1)}(t) + \dots + p_m f^{(m)}(t)$$

(так как степень полинома $f^{(j)}$ равна $j - 1$, $p(t)$ – полином порядка m).

Для определения коэффициентов имеем, аналогично предыдущему пункту, систему линейных уравнений

$$FP = y, \quad (11.3.3)$$

где $F = [F^{(1)}, \dots, F^{(m)}]$ – матрица размера $n \times m$ с попарно ортогональными столбцами, p – искомый вектор коэффициентов, y – вектор результатов измерений.

Умножим теперь (11.3.3) слева на F^* (перейдем к нормальным уравнениям):

$$(F^*F)P = F^*y.$$

Решение этой системы (с диагональной матрицей), согласно формулам (10.3.3), имеет вид

$$p_k = \frac{\langle y, F^{(k)} \rangle}{\|F^{(k)}\|^2}, \quad k = 1, \dots, m.$$

Таким образом, при заготовленной заранее стандартной (при заданных n и m) матрице F построение сглаживающего полинома сводится фактически к умножению матрицы F^* на полученный в эксперименте столбец и делению координат столбца-произведения на также заранее заготовленные числа $\|F^{(k)}\|^2$.

Мы рассмотрели два способа построения сглаживающего полинома: с помощью сингулярного разложения матрицы коэффициентов системы и с помощью ортогональных на сетке полиномов.

И тот и другой способ требует выполнения большой подготовительной работы: в первом случае – построения сингулярного разложения, во втором – построения ортогональных на сетке полиномов. И та и другая операция поддерживается устойчивыми вычислительными алгоритмами, реализованными как в средах конечного пользователя, так и в библиотеках программ на Фортране.

Мы не будем пытаться дать сравнительную оценку рассмотренных методов. Отметим лишь два факта.

1. Если мы построили сглаживающий полином, но среднеквадратическая погрешность оказалась велика и следует увеличить порядок полинома, то в случае сингулярного разложения придется пересчитывать все заново. При использовании же ортогональных на сетке полиномов потребуется добавить лишь еще одно слагаемое – уже сосчитанные коэффициенты разложения сохраняются. Это – полезное свойство *ортогональных* систем функций.

2. При использовании сингулярного разложения сглаживающий полином получается в стандартной форме,

$$p(t) = a_1 + a_2 t + \dots + a_m t^{m-1},$$

которая допускает применение схемы Горнера.

При использовании ортогональных на сетке полиномов сглаживающий полином получается в виде их линейной комбинации

$$p(t) = p_1 f^{(1)}(t) + \dots + p_m f^{(m)}(t).$$

Не следует приводить подобные члены в правой части (преобразовывать этот полином в стандартную форму), так как существует простое обобщение схемы Горнера, позволяющее работать с полиномом, разложенными по ортогональным на сетке полиномам.

11.4. Дискретное преобразование Фурье

Идея построения сглаживающего полинома с помощью ортогональных на сетке полиномов может быть обобщена. В самом деле, кроме полиномов, существуют и другие системы "простых и удобных функций". В этом пункте мы рассмотрим одну такую систему.

Пусть $w^{(r)}(t) = \exp(i\frac{2\pi r}{n}t)$, $r = 1, \dots, n$. Покажем, что эти функции попарно ортогональны на стандартной сетке $\{1, \dots, n\}$. Действительно,

$$\begin{aligned}\langle w^{(j)}, w^{(r)} \rangle &= \sum_{k=1}^n w^{(j)}(k) \cdot \overline{w^{(r)}(k)} = \sum_{k=1}^n \exp\left(i\frac{2\pi j}{n}k\right) \cdot \overline{\exp\left(i\frac{2\pi r}{n}k\right)} = \\ &= \sum_{k=1}^n \exp\left(i\frac{2\pi}{n}(j-r)k\right) = \sum_{k=1}^n \left(\exp\left(i\frac{2\pi}{n}(j-r)\right)\right)^k.\end{aligned}$$

Если $j = r$, то $\exp\left(i\frac{2\pi}{n}(r-r)\right) = 1$, и, следовательно,

$$\langle w^{(r)}, w^{(r)} \rangle = n. \quad (11.4.1)$$

Если же $j \neq r$, то, просуммировав геометрическую прогрессию, имеем

$$\begin{aligned}\langle w^{(j)}, w^{(r)} \rangle &= \exp\left(i\frac{2\pi}{n}(j-r)\right) \cdot \frac{\left(\exp\left(i\frac{2\pi}{n}(j-r)\right)\right)^n - 1}{\exp\left(i\frac{2\pi}{n}(j-r)\right) - 1} = \\ &= \frac{\exp\left(i\frac{2\pi}{n}(j-r)\right)}{\exp\left(i\frac{2\pi}{n}(j-r)\right) - 1} \cdot \left(\exp\left(i\frac{2\pi n}{n}(j-r)\right) - 1\right) = 0. \quad \blacksquare\end{aligned}$$

Замечание. Функции $w^{(r)}(t) = \exp(i\frac{2\pi r}{n}t)$ можно определить при всех $r \in \mathbb{Z}$, и если рассматривать их как функции, заданные на \mathbb{R} , то все они различны. Однако мы рассматриваем их *только на стандартной сетке*, поэтому $t \in \{1, \dots, n\}$ и

$$\exp\left(i\frac{2\pi(r+n)}{n}t\right) = \exp\left(i\left(\frac{2\pi r}{n}t + 2\pi t\right)\right) = \exp\left(i\frac{2\pi r}{n}t\right).$$

Таким образом, $w^{(r+n)}(t) = w^{(r)}(t)$, и любая функция $w^{(r)}$, $r \in \mathbb{Z}$, на стандартной сетке совпадает с одной из функций $w^{(1)}, w^{(2)}, \dots, w^{(n)}$.

Отметим еще некоторые свойства рассматриваемых функций.

1. Функции $w^{(r)}$ n -периодичны по t :

$$\begin{aligned}w^{(r)}(t+n) &= \exp\left(i\frac{2\pi r}{n}(t+n)\right) = \\ &= \exp\left(i\left(\frac{2\pi r}{n}t + 2\pi r\right)\right) = \exp\left(i\frac{2\pi r}{n}t\right) = w^{(r)}(t).\end{aligned}$$

2. Функции $w^{(r)}$ и $w^{(n-r)}$ принимают на стандартной сетке комплексно сопряженные значения:

$$w^{(n-r)}(k) = w^{(-r)}(k) = \exp\left(-i\frac{2\pi r}{n}k\right) = \overline{w^{(r)}(k)}.$$

$$3. w^{(n)}(k) \equiv w^{(0)}(k) \equiv 1 \quad \text{при } k \in \mathbb{Z}.$$

Сгладим теперь результаты наблюдений (п.11.2) линейной комбинацией функций $w^{(1)}, \dots, w^{(m)}$, $m \leq n$. Для определения коэффициентов q_1, \dots, q_m этой линейной комбинации получим систему линейных уравнений

$$y_k = q_1 w^{(1)}(k) + \dots + q_m w^{(m)}(k), \quad k = 1, \dots, n. \quad (11.4.2)$$

В матричной форме система уравнений (11.4.2) имеет вид $Wq = y$, где

$$W = \begin{bmatrix} w^{(1)}(1) & \dots & w^{(m)}(1) \\ \dots & \dots & \dots \\ w^{(1)}(n) & \dots & w^{(m)}(n) \end{bmatrix}$$

– $(n \times m)$ -матрица с попарно ортогональными столбцами.

Переходя к нормальным уравнениям, получим $(W^*W)q = W^*y$, или, учитывая (11.4.1), $\text{diag}[n, \dots, n] \cdot q = W^*y$. Отсюда

$$q = \frac{1}{n} W^*y, \quad \text{или} \quad q_r = \frac{1}{n} \langle y, w^{(r)} \rangle, \quad r = 1, \dots, m. \quad (11.4.3)$$

Если в сглаживании участвуют все функции ($m = n$), то нормальное псевдорешение превращается в решение – получаем разложение вектора в \mathbb{C}^n по ортогональному базису:

$$y_k = \sum_{r=1}^n q_r \exp\left(i\frac{2\pi r}{n}k\right), \quad k = 1, \dots, n, \quad (11.4.4)$$

где

$$q_r = \sum_{k=1}^n y_k \exp\left(-i\frac{2\pi r}{n}k\right), \quad r = 1, \dots, n. \quad (11.4.5)$$

Формулы (11.4.4) и (11.4.5) задают так называемое *дискретное преобразование Фурье*⁶³, причем формула (11.4.5), с помощью которой определяются коэффициенты разложения q_r , называется *прямым преобразованием*, а формула (11.4.4), восстанавливающая исходный вектор y – *обратным преобразованием*.

⁶³Жан Батист Жозеф ФУРЬЕ (J.-B.-J. Fourier, 1768-1830) – французский математик, член Парижской АН, почетный член Петербургской АН.

Числа q_r называются *коэффициентами Фурье* вектора y или его *комплексным Фурье-спектром*. Отметим, что при аппроксимации *вещественного* вектора целесообразно выбирать попарно сопряженные на стандартной сетке функции $w^{(r)}$ (см. свойство 2).

Замечание. Так же, как при использовании ортогональных на сетке полиномов, при увеличении числа слагаемых в сглаживающей линейной комбинации уже сосчитанные коэффициенты q_r сохраняются.

11.5. Быстрое преобразование Фурье

Для вычисления одного коэффициента Фурье по формуле (11.4.5) или для восстановления одной компоненты исходного вектора по формуле (11.4.4) требуется (при готовой матрице W) выполнить n пар "условных операций" (умножение + сложение). Таким образом, весь процесс вычисления дискретного преобразования Фурье потребует выполнения n^2 условных операций. При $n \approx 10^7 - 10^8$ (а такие массивы в приложениях не редки) обработка результатов измерений становится недоступной для современных ЭВМ.

В этом пункте рассматривается одна из разновидностей так называемого *быстрого преобразования Фурье* (БПФ). Будем считать, что $n = 2^p$, $p \in \mathbb{N}$.

Обозначим $z = \exp(i\frac{2\pi}{n})$. Тогда, очевидно, $z^{n/2} = -1$, $z^n = 1$.

Формула (11.4.4) примет вид

$$y_k = \sum_{r=1}^n q_r z^{kr}, \quad k = 1, \dots, n. \quad (11.5.1)$$

Соберем в правой части (11.5.1) отдельно слагаемые с четными и с нечетными номерами (по предположению $n = 2^p$ – четное число):

$$\begin{aligned} y_k &= \sum_{r=1}^{n/2} (q_{2r} z^{2kr} + q_{2r-1} z^{k(2r-1)}) = \\ &= \sum_{r=1}^{n/2} (q_{2r} z^{2kr} + z^{-k} q_{2r-1} z^{2kr}) = x_k^{(e)} + z^{-k} x_k^{(o)}, \end{aligned} \quad (11.5.2)$$

где

$$x_k^{(e)} = \sum_{r=1}^{n/2} q_{2r} (z^2)^{kr}, \quad x_k^{(o)} = \sum_{r=1}^{n/2} q_{2r-1} (z^2)^{kr}. \quad (11.5.3)$$

Очевидно, что при $k = 1, \dots, n/2$

$$x_{k+n/2}^{(e)} = \sum_{r=1}^{n/2} q_{2r} z^{2(k+n/2)r} = \sum_{r=1}^{n/2} q_{2r} z^{nr} (z^2)^{kr} = x_k^{(e)}$$

и, аналогично, $x_{k+n/2}^{(o)} = x_k^{(o)}$.

Формулы (11.5.3) показывают, что столбцы $x_k^{(e)}$ и $x_k^{(o)}$ (высоты $n/2$) являются дискретными Фурье-образами столбцов, составленных соответственно из четных и нечетных компонент столбца q .

Преобразуем формулу (11.5.2):

$$\begin{cases} y_k = x_k^{(e)} + z^{-k} x_k^{(o)}, \\ y_{k+n/2} = x_{k+n/2}^{(e)} + z^{-n/2} z^{-k} x_{k+n/2}^{(o)} = x_k^{(e)} - z^{-k} x_k^{(o)}, \end{cases} \quad k = 1, \dots, n/2.$$

Видно, что для получения столбца y (высоты n) достаточно вычислить столбцы $x_k^{(e)}$ и $x_k^{(o)}$ (высоты $n/2$) и выполнить еще n условных операций.

Обозначим $f(p)$ количество условных операций, необходимых для вычисления дискретного преобразования Фурье столбца высоты $n = 2^p$. Тогда

$$f(p) = 2 \cdot f(p-1) + 2^p.$$

Учитывая, что $f(0) = 0$, методом математической индукции несложно убедиться (проверьте это!), что

$$f(p) = p \cdot 2^p = n \cdot \log_2(n).$$

При $n = 2^{24} \approx 1.7 \cdot 10^7$ имеем $n \cdot \log_2(n) = 24 \cdot 2^{24} < 4.1 \cdot 10^8$ – при помощи БПФ вычисления выполняются за несколько секунд, в то время как $n^2 = 2^{48} > 2.8 \cdot 10^{14}$ – без использования БПФ требуемое время увеличивается почти в 10^6 раз!

Алгоритм БПФ реализован в средах конечного пользователя и в библиотеках Фортран-программ. Имеются варианты алгоритма, в которых высота столбца не обязана быть степенью двойки.

Глава 12. ЭЛЕМЕНТЫ ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ

Терминологическое замечание. Мы считаем необходимым отметить, что следует различать *программирование*, т.е. составление компьютерных программ, и *математическое программирование*, т.е. исследование и решение задач оптимизации (минимизации или максимизации) вещественного функционала, заданного на \mathbb{R}^n или на его части.

Частный случай математического программирования есть *линейное программирование* – оптимизация *линейного* функционала на части \mathbb{R}^n , заданной *линейными* равенствами или неравенствами.

Линейное программирование ведет свою историю от работы Л.В. Канторовича⁶⁴, выполненной в 1938 году.

12.1. Одна содержательная задача

Описание проблемы, которой посвящена эта глава, мы начнем с простейшего примера.

Коммерсант, выехавший для закупки двух видов товара, имеет 18 денежных единиц (д.е.)⁶⁵, его автомобиль может вместить 10 единиц массы (е.м.). Одна е.м. товара первого вида стоит 1 д.е., второго вида – 3 д.е. При продаже 1 е.м. товара первого вида коммерсант рассчитывает получить 0.5 д.е. прибыли, при продаже 1 е.м. товара второго вида – 0.75 д.е. Как распределить имеющиеся деньги и вместимость автомобиля, чтобы ожидаемая прибыль была максимальной?

Пусть x_1 – закупаемое коммерсантом количество товара первого вида, x_2 – второго вида. Эти переменные должны удовлетворять следующим очевидным неравенствам:

- 1) $x_1 \geq 0, \quad x_2 \geq 0$ (количество товаров неотрицательны);
- 2) $x_1 + 3x_2 \leq 18$ (сумма затрат не может превышать наличность);
- 3) $x_1 + x_2 \leq 10$ (суммарная масса закупленных товаров не может превышать вместимость автомобиля).

На части \mathbb{R}^2 , где выполнены все эти неравенства, требуется найти наибольшее значение линейного функционала

⁶⁴Леонид Витальевич КАНТОРОВИЧ (1912-1986) – советский математик и экономист, лауреат Нобелевской премии, член АН СССР и ряда зарубежных академий, один из основоположников математической экономики.

⁶⁵Мы намеренно не уточняем, о каких единицах идет речь, чтобы сохранить коммерческую тайну.

$$f(x) = 0.5x_1 + 0.75x_2.$$

Рассмотрим геометрическую интерпретацию нашей задачи. Множество, задаваемое одним линейным неравенством, представляет собой полуплоскость в \mathbb{R}^2 (см. п.9.1). Множество, задаваемое системой неравенств 1)-3), есть выпуклый⁶⁶ четырехугольник (см. рис.12.1).

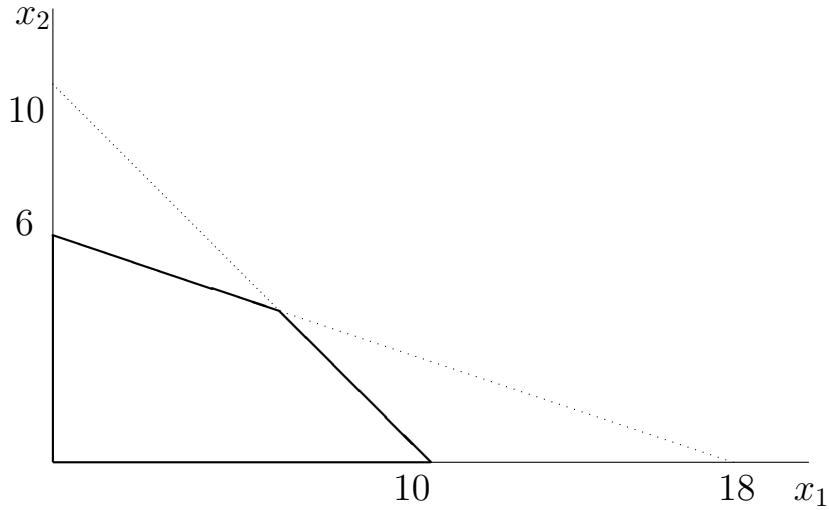


Рис.12.1

Линии уровня функционала f – это семейство параллельных прямых (все они перпендикулярны направленному отрезку, соответствующему вектору $[0.5, 0.75]^T$, см. п.9.1). Из прямых этого семейства, пересекающих наш четырехугольник, следует выбрать ту, которая соответствует наибольшему значению f . Из рис.12.2 видно, что глобальный максимум f достигается в вершине четырехугольника с координатами $x_1 = 6$, $x_2 = 4$, и $f_{max} = 0.5 \cdot 6 + 0.75 \cdot 4 = 6$.

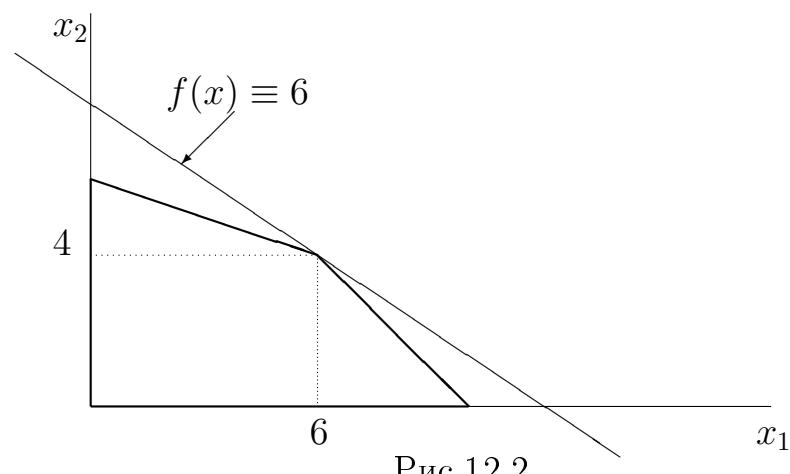


Рис.12.2

⁶⁶Множество называется *выпуклым*, если вместе с любыми двумя своими точками оно содержит соединяющий эти точки отрезок.

12.2. Каноническая задача линейного программирования

Обобщим пример, рассмотренный в предыдущем пункте. Задача линейного программирования состоит в поиске *глобального минимума* (наименьшего значения) линейного функционала

$$f(x) = \langle x, c \rangle = c_1 x_1 + \dots + c_n x_n \quad (12.2.1)$$

на части \mathbb{R}^n , все точки которой удовлетворяют следующим условиям:

$$a_{k1}x_1 + \dots + a_{kn}x_n \leq b_k, \quad k = 1, \dots, m1; \quad (12.2.2)$$

$$a_{k1}x_1 + \dots + a_{kn}x_n = b_k, \quad k = m1 + 1, \dots, m2; \quad (12.2.3)$$

$$a_{k1}x_1 + \dots + a_{kn}x_n \geq b_k, \quad k = m2 + 1, \dots, m. \quad (12.2.4)$$

Сделаем необходимые уточнения.

1. Существуют содержательные задачи (см. п.12.1), в которых функционал нужно не минимизировать, а максимизировать. Этот вариант, очевидно, укладывается в нашу схему при помощи замены вектора c на противоположный. Иначе говоря, *максимизация* функционала $f(x) = \langle x, c \rangle$ – это то же, что *минимизация* функционала $\varphi(x) = \langle x, -c \rangle$.

2. Мы будем считать все переменные неотрицательными ($x_k \geq 0$, $k = 1, \dots, n$) и выделим эти неравенства в особую группу. Покажем, что это условие не является стесняющим. Действительно, если переменная x_k ограничена снизу ($x_k \geq p$), можно ввести новую переменную по формуле $x_k^+ = x_k - p \geq 0$. Если переменная x_k ограничена сверху ($x_k \leq p$), можно ввести новую переменную по формуле $x_k^- = p - x_k \geq 0$. Если, наконец, переменная x_k не ограничена ни сверху, ни снизу, положим (увеличивая количество переменных) $x_k = x'_k - x''_k$, где $x'_k \geq 0$, $x''_k \geq 0$.

Множество векторов из \mathbb{R}^n с *неотрицательными* координатами будем обозначать \mathbb{R}_+^n .

3. Мы будем считать все свободные члены в системе (12.2.2)–(12.2.4) неотрицательными ($b \in \mathbb{R}_+^m$). Этого всегда можно добиться, умножая при необходимости уравнение или неравенство на (-1) .

Можно показать, что часть \mathbb{R}^n , задаваемая системой (12.2.2)–(12.2.4), есть

- 1) либо пустое множество (система несовместна),
- 2) либо точка,
- 3) либо выпуклый многогранник:

- а) ограниченный (пример – рис.12.3),
 б) неограниченный (пример – рис.12.4).

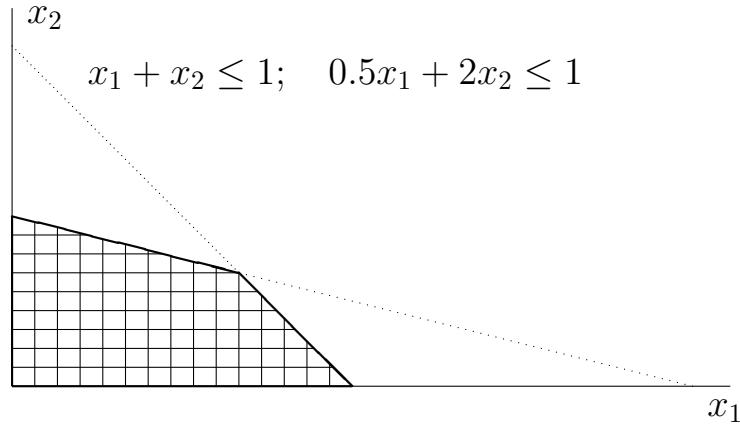


Рис.12.3

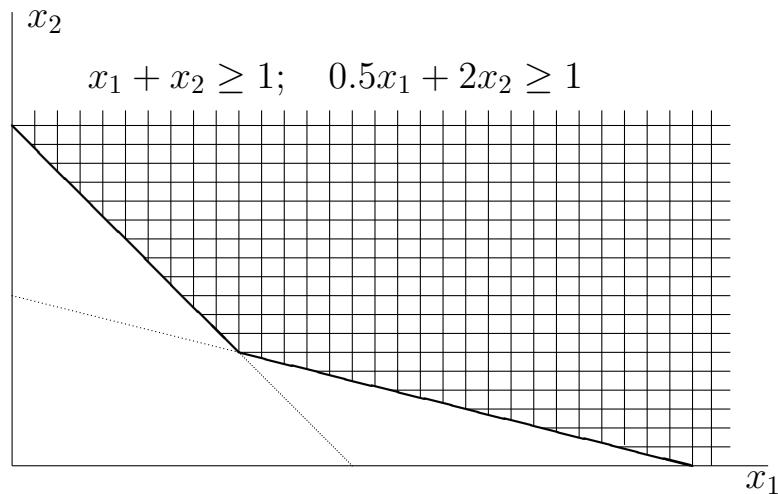


Рис.12.4

В случае **1** задача линейного программирования не имеет решения;
 В случае **2** решением задачи линейного программирования является, очевидно, единственное решение системы. Оба эти случая тривиальны.

Интерес представляет случай **3**.

В случае **3а)** рассуждением, подобным проведенному в п.12.1, можно показать, что глобальный минимум функционала существует и достигается хотя бы в одной из вершин многогранника.

В случае **3б)** возможны два варианта: либо множество значений функционала не ограничено снизу, либо минимум существует и также достигается в одной из вершин многогранника.

Покажем теперь, что все ограничения-неравенства, кроме неравенств $x_k \geq 0$, можно заменить ограничениями-равенствами (за счет увеличения количества переменных в задаче).

Действительно, можно ввести новую переменную $x_{n+1} \geq 0$ и заменить неравенство (12.2.2) на уравнение

$$a_{k1}x_1 + \dots + a_{kn}x_n + x_{n+1} = b_k,$$

из которого следует (12.2.2).

Аналогично, вводя новую переменную $x_{n+1} \geq 0$, можно заменить неравенство (12.2.4) на уравнение

$$a_{k1}x_1 + \dots + a_{kn}x_n - x_{n+1} = b_k,$$

из которого следует (12.2.4).

Конечно, дополнительные переменные не должны входить в минимизируемый функционал (соответствующие координаты вектора с должны равняться нулю).

Сформулируем теперь так называемую *каноническую задачу линейного программирования*:

Минимизировать линейный функционал (12.1.1) на части \mathbb{R}_+^n , все точки которой удовлетворяют системе линейных уравнений

$$Ax = b, \tag{12.2.5}$$

где A – матрица размера $m \times n$, а $b \in \mathbb{R}_+^n$.

Замечания. 1. Мы считаем, что система (12.2.5) имеет бесконечно много решений в \mathbb{R}_+^n , иначе задача тривиальна. Можно также полагать, что ни одно из уравнений не является следствием других (иначе его можно просто вычеркнуть). Отсюда, кстати, следует, что $m < n$.

2. Мы по-прежнему обозначаем буквой n размерность пространства (количество переменных) и буквой m – количество ограничений-равенств. Следует, однако, помнить, что теперь количество переменных может быть *больше*, чем в исходной задаче – за счет дополнительных переменных, появляющихся при замене ограничений-неравенств равенствами. Количество ограничений может оказаться *меньше*, чем в исходной задаче – за счет вычеркивания уравнений, являющихся следствием оставшихся.

12.3. Преобразование канонической задачи линейного программирования

Из замечания 1 в конце п.12.2 следует, что при решении системы (12.2.5) методом Гаусса–Йордана количество уравнений не уменьшается.

Поэтому значения некоторых m переменных однозначно определяются значениями оставшихся $n - m$ переменных, а те могут быть заданы произвольно (здесь мы пока не учитываем ограничения $x \in \mathbb{R}_+^n$). Отметим также, что указанные выше m переменных можно выбрать не единственным образом. Мы будем для удобства считать, что эти m переменных – x_1, \dots, x_m .

Разобьем матрицу A , вектор c и переменный вектор x на две части:

$$A = [B : N], \quad c = \begin{bmatrix} c^B \\ \vdots \\ c^N \end{bmatrix}, \quad x = \begin{bmatrix} x^B \\ \vdots \\ x^N \end{bmatrix}.$$

Здесь B – квадратная матрица порядка m ; N – $m \times (n - m)$ -матрица; x^B, c^B – столбцы высоты m ; x^N, c^N – столбцы высоты $n - m$.

Теперь система (12.2.5) перепишется в виде

$$Bx^B + Nx^N = b, \quad (12.3.1)$$

а функционал (12.2.1) – в виде

$$f(x) = (c^B)^T \cdot x^B + (c^N)^T \cdot x^N. \quad (12.3.2)$$

Поскольку система (12.3.1) при каждом значении x^N однозначно разрешима относительно x^B , то матрица B обратима, и мы получаем

$$x^B = B^{-1}b - B^{-1}Nx^N. \quad (12.3.3)$$

Подставляя полученный результат в (12.3.2), имеем

$$f(x) = (c^B)^T \cdot B^{-1}b + ((c^N)^T - (c^B)^T \cdot B^{-1} \cdot N) \cdot x^N.$$

Введем следующие обозначения:

$$\beta = B^{-1}b \text{ (столбец высоты } m\text{),}$$

$$\pi = (c^N)^T - (c^B)^T \cdot B^{-1} \cdot N \text{ (строка ширины } n - m\text{).}$$

Тогда получим так называемую *преобразованную задачу линейного программирования*:

Минимизировать функционал

$$\varphi(x^N) = (c^B)^T \cdot \beta + \pi \cdot x^N \quad (12.3.3)$$

при условиях

$$x^N \in \mathbb{R}_+^{n-m}, \quad x^B = \beta - B^{-1} \cdot N \cdot x^N \in \mathbb{R}_+^m. \quad (12.3.4)$$

Координаты вектора x^B принято называть *базисными* переменными, координаты вектора x^N – *небазисными*. Решение системы $Ax = b$ называется *базисным*, если $x^N = \theta_{n-m}$ ($x^B = \beta$). Базисное решение называется *допустимым*, если $x^B = \beta \in \mathbb{R}_+^m$. Если базисное решение допустимо, и $\pi^T \in \mathbb{R}_+^{n-m}$, то это базисное решение доставляет функционалу искомый минимум, так как при $x^N \in \mathbb{R}_+^{n-m}$

$$\varphi(x^N) = (c^B)^T \cdot \beta + \pi \cdot x^N \geq (c^B)^T \cdot \beta = \varphi(\theta_{n-m}).$$

Замечание. Столбец β и строка π зависят от выбора базисных переменных. Напомним, что этот выбор можно произвести не единственным образом.

Пример. Минимизировать функционал

$$f(x) = 3x_1 + x_2 + 2x_3$$

при условиях

$$x_1 + 2x_2 + 3x_3 = 6; \quad x_1 \geq 0, \quad x_2 \geq 0, \quad x_3 \geq 0.$$

Здесь $n = 3$, $m = 1$, $A = [1, 2, 3]$, $b = [6]$, $c = [3, 1, 2]^T$.

Если за базисную переменную принять x_1 , то

$$B = [1], \quad N = [2, 3], \quad \beta = [6], \quad \pi = [1, 2] - 3 \cdot [2, 3] = [-5, -7].$$

Базисное решение $[6, 0, 0]^T$ допустимо.

Если за базисную переменную принять x_3 , то

$$B = [3], \quad N = [1, 2], \quad \beta = [2], \quad \pi = [3, 1] - \frac{2}{3} \cdot [1, 2] = [\frac{7}{3}, -\frac{1}{3}].$$

Базисное решение $[0, 0, 2]^T$ также допустимо.

Если, наконец, за базисную переменную принять x_2 , то

$$B = [2], \quad N = [1, 3], \quad \beta = [3], \quad \pi = [3, 2] - \frac{1}{2} \cdot [1, 3] = [\frac{5}{2}, \frac{1}{2}].$$

Базисное решение $[0, 3, 0]^T$ допустимо и доставляет минимум функционалу, так как $\pi^T \in \mathbb{R}_+^2$.

Проиллюстрируем наш пример геометрически. Система ограничений, состоящая из одного уравнения $x_1 + 2x_2 + 3x_3 = 6$, задает плоскость в \mathbb{R}^3 . Пересечение этой плоскости с \mathbb{R}_+^3 – треугольник (рис.12.5). Легко видеть, что допустимые базисные решения $[6, 0, 0]^T$, $[0, 3, 0]^T$, $[0, 0, 2]^T$

соответствуют вершинам этого треугольника. Как уже указывалось, минимальное значение функционала f достигается хотя бы в одной из вершин. Сравнив значения функционала во всех вершинах $f(6, 0, 0) = 18$, $f(0, 3, 0) = 3$, $f(0, 0, 2) = 4$, видим, что базисное решение $[0, 3, 0]^T$ действительно доставляет функционалу глобальный минимум.

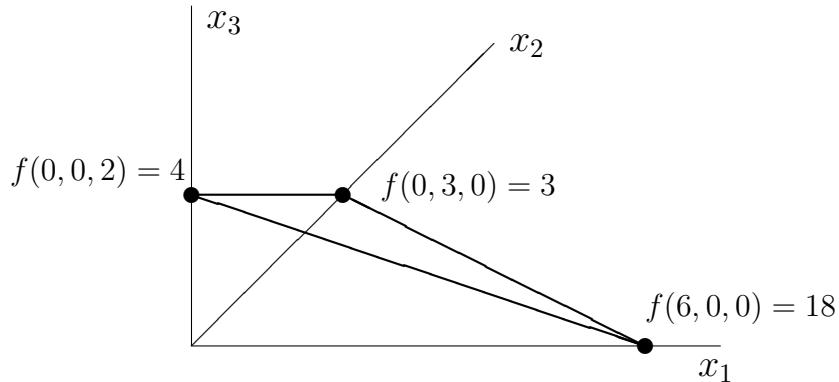


Рис.12.5

12.4. Понятие о симплекс-методе решения задачи линейного программирования

Этот метод, открытый в 1947 году Д. Данцигом⁶⁷, представляет собой *конечный* алгоритм, который, начиная с некоторого *допустимого базисного* решения, строит новые *допустимые базисные решения*, обеспечивая *уменьшение* значения функционала.

Мы опишем один шаг симплекс-метода, не вдаваясь, естественно, в технологические подробности.

Итак, пусть имеется допустимое базисное решение

$$x = \begin{bmatrix} x^B \\ \dots \\ x^N \end{bmatrix}, \quad x^B = \beta \in \mathbb{R}_+^m, \quad x^N = \theta_{n-m}.$$

Вычислим строку π , соответствующую нашему базисному решению. Если $\pi^T \in \mathbb{R}_+^{n-m}$, то, как указано в п.12.3, минимум функционала φ достигнут, и задача линейного программирования решена. *Работа алгоритма заканчивается*.

Если среди элементов строки π есть отрицательные, то, увеличивая соответствующие им координаты вектора x^N , мы будем уменьшать значение функционала. Скорость убывания функционала пропорциональна значениям отрицательных элементов строки π . Поэтому выберем

⁶⁷Джордж Бернард ДАНЦИГ (J.B. Dantzig, 1914-2005) – американский математик.

из них наименьший (наибольший по модулю). Пусть его порядковый номер q . Будем увеличивать q -ю координату вектора x^N , пока все координаты вектора x^B еще неотрицательны. Поскольку отлична от нуля только одна координата вектора x^N , (12.3.4) принимает вид

$$x^B = B^{-1}b - (B^{-1}N)^{(q)} x_q^N.$$

Если все элементы столбца $(B^{-1}N)^{(q)}$ неположительны, то координаты вектора x^B остаются неотрицательными при любых положительных значениях x_q^N . Это значит, что множество значений функционала φ не ограничено снизу, и задача не имеет решения. Работа алгоритма заканчивается.

Если среди элементов столбца $(B^{-1}N)^{(q)}$ есть положительные, то соответствующие им координаты вектора x^B будут убывать с увеличением x_q^N . Пусть s – номер той координаты вектора x^B , которая первой обратится в нуль в этом процессе. Тогда наибольшее возможное значение x_q^N равно

$$\max(x_q^N) = \frac{\beta_s}{(B^{-1}N)_s^{(q)}} = \min_j \left(\frac{\beta_j}{(B^{-1}N)_j^{(q)}} \right)$$

(минимум берется по тем индексам j , для которых $(B^{-1}N)_j^{(q)} > 0$).

При этом значении x_q^N получим $x_s^B = 0$. Теперь переменная x_q^N включается в состав базисных, а обнуленная переменная x_s^B исключается.

Дальнейшие действия можно было бы представить себе так: меняем местами столбцы $B^{(s)}$ и $N^{(q)}$ в матрице A и соответствующие координаты вектора c . Получим "новые" матрицы B и N . Вычислим новый вектор β и новую строку π . На полученном новом допустимом ($x^B = \beta \in \mathbb{R}_+^m$) базисном ($x^N = \theta_{n-m}$) решении функционал φ по построению имеет меньшее значение, чем на старом. Один шаг алгоритма закончен.

Замечание. Конечно, изложенная процедура неэффективна. Стандартные программы, реализующие алгоритм симплекс-метода, работают иначе. Но мы, напоминаем, не рассматриваем технические подробности.

Поскольку количество допустимых базисных решений не превосходит количества различных наборов базисных переменных, а оно, в свою очередь, не превосходит количества сочетаний из n столбцов матрицы A по m , т.е. $\frac{n!}{m!(n-m)!}$, то за конечное число шагов алгоритм либо находит глобальный минимум функционала, либо выявляет его отсутствие.

12.5. Построение начального допустимого базисного решения

Для начала работы алгоритма симплекс-метода необходимо иметь какое-нибудь допустимое базисное решение. Мы опишем один из возможных способов его построения.

Пусть требуется минимизировать функционал $f(x) = \langle x, c \rangle$ при условиях

$$x \in \mathbb{R}_+^n, \quad Ax = b \in \mathbb{R}_+^m. \quad (12.5.1)$$

Рассмотрим вспомогательную задачу о минимизации линейного функционала $\varphi(x, y) = y_1 + \dots + y_m$ при условиях

$$x \in \mathbb{R}_+^n, \quad y \in \mathbb{R}_+^m, \quad [A : I_m] \begin{bmatrix} x \\ \vdots \\ y \end{bmatrix} = b. \quad (12.5.2)$$

Для задачи (12.5.2) одно допустимое базисное решение очевидно: $x = \theta_n$, $y = b$. Поэтому можно применить к ней симплекс-метод.

Поскольку $\varphi(x, y) \geq 0$, функционал φ ограничен снизу, и алгоритм за конечное число шагов даст решение вспомогательной задачи (\tilde{x}, \tilde{y}) .

Если окажется, что $\tilde{y} = \theta_m$, то \tilde{x} – допустимое базисное решение системы $Ax = b$ (проверьте это!).

Если же $\tilde{y} \neq \theta_m$, то минимум во вспомогательной задаче положителен ($\varphi(\tilde{x}, \tilde{y}) > 0$). Но если для вектора x выполнены условия (12.5.1), то пара (x, θ_m) удовлетворяет условиям (12.5.2), и $\varphi(x, \theta_m) = 0$, т.е. минимум во вспомогательной задаче равен нулю. Поэтому результат $\tilde{y} \neq \theta_m$ указывает на несовместность условий (12.5.1), и исходная задача не имеет решения.

Глава 13. ЭЛЕМЕНТАРНЫЙ АНАЛИЗ ПОГРЕШНОСТЕЙ

13.1. Предварительные замечания

В предыдущих главах мы предполагали, что все исходные данные рассматриваемых задач заданы точно, и точно выполняются арифметические операции. Однако в действительности и в исходных данных обычно имеется неопределенность (так как они являются, как правило, либо результатами измерений, либо результатами вычислений), и арифметические операции выполняются либо с округлением, либо с усечением. Это приводит к появлению в результатах неопределенности, без оценки которой пользоваться этими результатами нельзя. Могут появляться и несуществующие на самом деле "решения" задач.

Пример. Попробуем решить очевидно несовместную систему

$$\begin{cases} 7x + 9y = 16 \\ 14x + 18y = 33 \end{cases}$$

методом Гаусса–Йордана без выбора ведущего элемента (вычисления ведутся с тремя значащими цифрами):

$$\begin{array}{c} \left[\begin{array}{cc|c} 7 & 9 & 16 \\ 14 & 18 & 33 \end{array} \right] \iff \left[\begin{array}{cc|c} 1.00 & 1.29 & 2.29 \\ 14. & 18. & 33. \end{array} \right] \iff \left[\begin{array}{cc|c} 1.00 & 1.29 & 2.29 \\ 0. & -1. & .9 \end{array} \right] \iff \\ \iff \left[\begin{array}{cc|c} 1.00 & 1.29 & 2.29 \\ -0. & 1.00 & -9.00 \end{array} \right] \iff \left[\begin{array}{cc|c} 1.00 & 0. & 13.9 \\ -0. & 1.00 & -9.00 \end{array} \right]. \end{array}$$

Получено "решение": $x = 13.9$, $y = -9.00$.

Для тех, кто думает, что такой эффект возможен только при решении "вручную" приводим пример решения несовместной (проверьте это!) системы с помощью среды конечного пользователя MAPLE:

```
> restart: with(linalg):  
  
Warning, the protected names norm and trace  
have been redefined and unprotected  
> fsolve({77777777777.*x+99999999999.*y=1777777777776,  
> 15555555554*x+199999999998*y=1777777777777},{x,y});  

$$\{x = -0.4444444445 \cdot 10^{10}, y = 0.3456790126 \cdot 10^{10}\}$$

```

Проиллюстрируем проблему геометрически. Как известно, линейное алгебраическое уравнение с двумя переменными задает на плоскости прямую, а система из двух таких уравнений – пару прямых. Если матрица коэффициентов системы невырождена, то прямые непараллельны

и имеют единственную точку пересечения, координаты которой – единственное решение системы (рис.13.1).

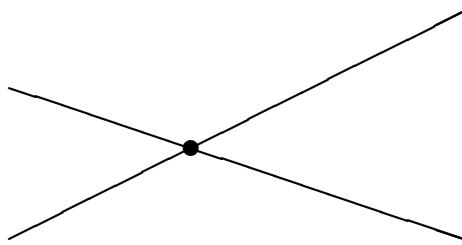


Рис.13.1

Наличие неопределенности в исходных данных (коэффициентах и свободных членах системы) превращает каждую прямую в целое семейство прямых. В простейшем случае, когда неопределенность содержится только в свободных членах, прямые каждого семейства параллельны между собой, и каждое уравнение системы порождает "толстую" прямую (полосу), толщина которой растет с ростом неопределенности. "Решением" системы теперь является пересечение двух "толстых" прямых – "толстая" точка, размеры которой характеризуют неопределенность решения.

Если матрица коэффициентов системы ортогональна, то прямые перпендикулярны, и размеры "решения" будут того же порядка, что и ширина полос (рис.13.2).

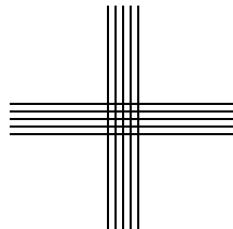


Рис.13.2

Если угол между полосами близок к нулю, то размеры "решения" могут во много раз превышать ширину полос (рис.13.3).

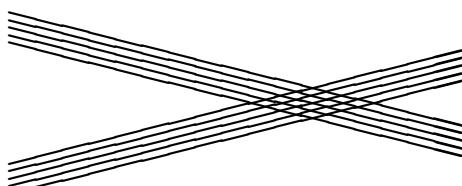


Рис.13.3

При параллельных прямых может возникнуть "решение не существующее в точной арифметике (рис.13.4).

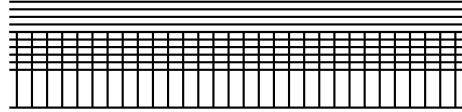


Рис.13.4

Эти геометрические построения показывают (хотя и не доказывают), что "плохими" являются системы, близкие к вырожденным. Чтобы придать этому утверждению точный смысл, нам потребуется ввести некоторые новые понятия.

13.2. Норма матрицы

Для квадратичной формы, порожденной эрмитовой матрицей A , известно неравенство (см. п.9.2)

$$\langle Ax, x \rangle \leq \lambda_{max} \cdot \|x\|^2, \quad (13.2.1)$$

где λ_{max} – наибольшее собственное число матрицы A .

Пусть теперь B – произвольная матрица размера $m \times n$. Из (13.2.1) следует, что

$$\|Bx\|^2 = \langle Bx, Bx \rangle = \langle B^*Bx, x \rangle \leq \sigma_{max}^2 \cdot \|x\|^2,$$

или

$$\|Bx\| \leq \sigma_{max} \cdot \|x\|, \quad (13.2.2)$$

где σ_{max} – наибольшее сингулярное число матрицы B . Если вектор x *ненулевой*, то, разделив на его норму обе части (13.2.2), получим

$$\frac{\|Bx\|}{\|x\|} \leq \sigma_{max}. \quad (13.2.3)$$

Покажем, что в (13.2.3) равенство достигается. Если v – правый, а u – левый сингулярные векторы, соответствующие наибольшему сингулярному числу, то (см. (10.1.4)):

$$\|Bv\| = \|\sigma_{max}u\| = \sigma_{max} = \sigma_{max}\|v\|, \quad \text{и} \quad \frac{\|Bv\|}{\|v\|} = \sigma_{max}.$$

Геометрическая интерпретация неравенства (13.2.3) очевидна: при умножении ненулевого вектора на матрицу слева евклидова норма этого вектора изменяется. Норма вектора в \mathbb{R}^3 – это длина соответствующего ему направленного отрезка. Поэтому естественно назвать отношение

$\frac{\|Bx\|}{\|x\|}$ "коэффициентом растяжения" вектора этой матрицей. Тогда неравенство (13.2.3) показывает, что коэффициент растяжения для каждой матрицы не может быть больше, чем ее наибольшее сингулярное число.

Определение. *Нормой* матрицы называется число

$$\|B\| = \max_{x \neq \theta} \frac{\|Bx\|}{\|x\|} = \sigma_{max}. \quad (13.2.4)$$

Установим свойства матричной нормы:

1. $\|B\| \geq 0; \quad \|B\| = 0 \Leftrightarrow B = \Theta$ – нулевой оператор (матрица).
2. $\|\alpha B\| = |\alpha| \cdot \|B\|.$
3. $\|A + B\| \leq \|A\| + \|B\|.$
4. $\|AB\| \leq \|A\| \cdot \|B\|.$
5. $\|B^*\| = \|B\|.$
6. Если B – обратимая матрица, то $\|B^{-1}\| = \frac{1}{\sigma_{min}(B)}.$
7. Норма неотрицательно определенной матрицы равна ее наибольшему собственному числу.
8. Норма унитарной матрицы равна единице.

Доказательство. 1. Из определения матричной нормы следует, что $\|B\| \geq 0$. Далее, очевидно, $\|\Theta\| = 0$. С другой стороны, если $\|B\| = 0$, то для всех $x \in \mathbb{C}^n$ имеем $\|Bx\| = 0$, т.е. $Bx = \theta$. Таким образом, $B = \Theta$.

$$2. \|\alpha B\| = \max_{x \neq \theta} \frac{\|\alpha Bx\|}{\|x\|} = |\alpha| \cdot \max_{x \neq \theta} \frac{\|Bx\|}{\|x\|} = |\alpha| \cdot \|B\|.$$

3. Для любого $x \in \mathbb{C}^n$ имеем

$$\begin{aligned} \|(A + B)x\| &= \|Ax + Bx\| \leq \|Ax\| + \|Bx\| \leq \\ &\leq \|A\| \cdot \|x\| + \|B\| \cdot \|x\| = (\|A\| + \|B\|) \cdot \|x\|. \end{aligned}$$

Отсюда

$$\|A + B\| = \max_{x \neq \theta} \frac{\|(A + B)x\|}{\|x\|} \leq \max_{x \neq \theta} \frac{(\|A\| + \|B\|) \cdot \|x\|}{\|x\|} = \|A\| + \|B\|.$$

4. Для любого $x \in \mathbb{C}^n$

$$\|(AB)x\| = \|A(Bx)\| \leq \|A\| \cdot \|Bx\| \leq \|A\| \cdot \|B\| \cdot \|x\|.$$

Поэтому

$$\|AB\| = \max_{x \neq \theta} \frac{\|(AB)x\|}{\|x\|} \leq \max_{x \neq \theta} \frac{\|A\| \cdot \|B\| \cdot \|x\|}{\|x\|} = \|A\| \cdot \|B\|.$$

5. В п.10.1 доказано, что ненулевые собственные числа матриц B^*B и BB^* совпадают. Поэтому $\|B^*\| = \sigma_{\max}(B^*) = \sigma_{\max}(B) = \|B\|$.

6. Пусть σ – сингулярное число матрицы B . Тогда σ^2 – собственное число матриц B^*B и BB^* . Отсюда $\frac{1}{\sigma^2}$ – собственное число матрицы $(BB^*)^{-1} = (B^{-1})^*B^{-1}$, т.е. $\frac{1}{\sigma}$ – сингулярное число матрицы B^{-1} . Следовательно,

$$\sigma_{\max}(B^{-1}) = \frac{1}{\sigma_{\min}(B)}.$$

7. В п.10.1 доказано, что для неотрицательно определенной матрицы сингулярные числа совпадают с собственными числами.

8. Если U – унитарная матрица, то $U^*U = I$, и все сингулярные числа равны единице. ■

Отметим, что свойства **1 – 3** матричной нормы совпадают со свойствами нормы вектора (п.7.2).

Замечание. Мы использовали в определении (13.2.4) евклидову норму вектора. Поэтому полное наименование введенной матричной нормы – *норма, подчиненная евклидовой норме вектора*.

Мы уже отмечали ранее (п.7.2), что норму вектора в линейном пространстве можно вводить различными способами. Соответственно появятся различные нормы матрицы. Если в двух конечномерных линейных пространствах X и Y введены нормы $\|\cdot\|_X$ и $\|\cdot\|_Y$ соответственно, то любому линейному оператору $B : X \rightarrow Y$ можно сопоставить число

$$\|B\|_{X \rightarrow Y} = \max_{x \neq \theta} \frac{\|Bx\|_Y}{\|x\|_X}.$$

Это число называют *нормой оператора*. Любая операторная норма обладает свойствами **1 – 4**, доказанными для нормы (13.2.4).

13.3. Трансформированная погрешность решения системы линейных алгебраических уравнений. Число обусловленности матрицы

Решение любой вычислительной задачи можно записать формулой

$$y = F(x).$$

Здесь $x \in \mathbb{C}^n$ – вектор исходных данных, $y \in \mathbb{C}^m$ – вектор результатов, F – отображение, действующее из \mathbb{C}^n в \mathbb{C}^m .

В реальной ситуации вектор x содержит неопределенность, т.е. "на вход" подается не вектор x , а некоторый другой вектор \tilde{x} . Точно так же отображение реализуется неточно, т.е. фактически работает некоторое другое отображение \tilde{F} . Таким образом, вместо требуемого результата y мы получаем "на выходе" некоторый другой результат \tilde{y} . Обозначив $\tilde{y} = F(\tilde{x})$, можно записать

$$\Delta y = \tilde{y} - y = (\tilde{y} - \tilde{y}) + (\tilde{y} - y) = (\tilde{F}(\tilde{x}) - F(\tilde{x})) + (F(\tilde{x}) - F(x)).$$

Мы представили погрешность результата в виде суммы двух слагаемых. При этом $\Delta y_T = F(\tilde{x}) - F(x)$ – погрешность, возникающая при точной реализации вычислительного алгоритма из-за погрешности в исходных данных, а $\Delta y_M = \tilde{F}(\tilde{x}) - F(\tilde{x})$ – погрешность, возникающая из-за неточной реализации алгоритма.

Существуют различные наименования для этих составляющих. Мы будем называть Δy_T трансформированной погрешностью, а Δy_M – погрешностью метода.

Анализ погрешностей решения системы линейных алгебраических уравнений начнем с простейшего случая.

Пусть в системе

$$Ax = b \quad (13.3.1)$$

невырожденная $n \times n$ -матрица A известна точно, и все арифметические операции выполняются без округлений и усечений, а свободный член содержит погрешность Δb , т.е. фактически вместо системы (13.3.1) решается система

$$A\tilde{x} = b + \Delta b.$$

Вычитая из этого уравнения (13.3.1), получим

$$A \cdot (\tilde{x} - x) = \Delta b, \quad \text{или} \quad \Delta x = A^{-1} \cdot \Delta b, \quad (13.3.2)$$

где $\Delta x = \tilde{x} - x$ – трансформированная погрешность решения.

По свойствам матричной нормы из (13.3.1) и (13.3.2) имеем

$$\|b\| \leq \|A\| \cdot \|x\|; \quad \|\Delta x\| \leq \|A^{-1}\| \cdot \|\Delta b\|.$$

Перемножив эти неравенства, получим

$$\|b\| \cdot \|\Delta x\| \leq \|A\| \cdot \|A^{-1}\| \cdot \|\Delta b\| \cdot \|x\|,$$

т.е.

$$\frac{\|\Delta x\|}{\|x\|} \leq \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\Delta b\|}{\|b\|}.$$

Введя понятие относительной погрешности

$$\delta x = \frac{\|\Delta x\|}{\|x\|}, \quad \delta b = \frac{\|\Delta b\|}{\|b\|},$$

придем к основному неравенству

$$\delta x \leq \text{cond}(A) \cdot \delta(b), \quad (13.3.3)$$

где $\text{cond}(A) = \|A\| \cdot \|A^{-1}\|$ – так называемое *число обусловленности* матрицы A .

Из (13.3.3) видно, что относительная погрешность результата при точно известной матрице и точном выполнении арифметических операций может превысить относительную погрешность исходных данных не более, чем в $\text{cond}(A)$ раз.

Покажем, что в (13.3.3) может достигаться равенство. Если $x, \delta x$ – *нормированные* правые сингулярные векторы матрицы A , причем x соответствует наибольшему сингулярному числу, а Δx – наименьшему, то

$$\begin{aligned} \|x\| &= 1, \quad \|\Delta x\| = 1, \quad \delta x = 1; \\ \|b\| &= \|Ax\| = \sigma_{\max}, \quad \|\Delta b\| = \|A\Delta x\| = \sigma_{\min}, \quad \delta b = \frac{\sigma_{\min}}{\sigma_{\max}}; \\ \frac{\delta x}{\delta b} &= \frac{\sigma_{\max}}{\sigma_{\min}} = \|A\| \cdot \|A^{-1}\| = \text{cond}(A). \end{aligned}$$

Результат, очевидно, не изменится, если x и Δx не совпадают с сингулярными векторами, а лишь коллинеарны им.

Таким образом, число обусловленности матрицы коэффициентов системы линейных алгебраических уравнений – это наибольшее значение "коэффициента усиления" относительной погрешности в задании свободного члена. При решении "плохо обусловленной" системы, т.е. системы, матрица коэффициентов которой имеет большое число обусловленности, может происходить (даже при точной арифметике!) катастрофическая потеря точности.

Из сказанного видно, что число обусловленности матрицы может служить мерой близости ее к вырожденной. Формально можно считать, что для вырожденной матрицы $\text{cond}(A) = +\infty$.

Покажем, как можно конструировать плохо обусловленные системы.

Пример. Для матрицы $A = \begin{bmatrix} 1 & 0 \\ N & 1 \end{bmatrix}$ ($N > 0$) имеем

$$A^*A = \begin{bmatrix} 1 + N^2 & N \\ N & 1 \end{bmatrix}, \quad P_{A^*A}(\lambda) = \lambda^2 - (N^2 + 2)\lambda + 1,$$

$$\lambda_{\max}(A^*A) = \frac{N^2 + 2 + N \cdot \sqrt{N^2 + 4}}{2} > N^2 + 1, \quad \lambda_{\min}(A^*A) = \frac{1}{\lambda_{\max}(A^*A)},$$

$$\operatorname{cond}(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} = \left(\frac{\lambda_{\max}(A^*A)}{\lambda_{\min}(A^*A)} \right)^{1/2} = \lambda_{\max}(A^*A) > N^2 + 1.$$

Итак, при решении этой системы (всего-то два уравнения) относительная погрешность задания свободного члена может трансформироваться в относительную погрешность результата, усилившись более, чем в N^2 раз!

Анализ трансформированной погрешности усложняется, если ошибки имеются и в матрице коэффициентов. В предположении, что относительные погрешности исходных данных *малы*, можно показать, что

$$\delta x \leq \operatorname{cond}(A) \cdot \frac{\delta A + \delta b}{1 - \operatorname{cond}(A) \cdot \delta(A)}$$

(здесь δA – относительная погрешность задания матрицы A).

Рассмотрим некоторые свойства числа обусловленности:

1. $\operatorname{cond}(A^{-1}) = \operatorname{cond}(A)$.
2. $\operatorname{cond}(A^*) = \operatorname{cond}(A)$.
3. $\operatorname{cond}(\alpha A) = \operatorname{cond}(A)$ при $\alpha \neq 0$.
4. $\operatorname{cond}(AB) \leq \operatorname{cond}(A) \cdot \operatorname{cond}(B)$.
5. $\operatorname{cond}(A) \geq 1$.
6. Если U – унитарная матрица, то

$$\operatorname{cond}(U) = 1; \quad \operatorname{cond}(AU) = \operatorname{cond}(UA) = \operatorname{cond}(A).$$

Доказательство. 1. Следует из определения.

2. Следует из $\|A^*\| = \|A\|$ (свойство 5 матричной нормы).
3. $\operatorname{cond}(\alpha A) = \|\alpha A\| \cdot \|(\alpha A)^{-1}\| = |\alpha| \cdot \|A\| \cdot |\alpha^{-1}| \cdot \|A^{-1}\| = \operatorname{cond}(A)$.
4. По свойству 4 матричной нормы

$$\begin{aligned} \operatorname{cond}(AB) &= \|AB\| \cdot \|(AB)^{-1}\| \leq \\ &\leq \|A\| \cdot \|B\| \cdot \|A^{-1}\| \cdot \|B^{-1}\| = \operatorname{cond}(A) \cdot \operatorname{cond}(B). \end{aligned}$$

$$5. \ 1 = \text{cond}(I) = \text{cond}(A \cdot A^{-1}) \leq \text{cond}(A) \cdot \text{cond}(A^{-1}) = (\text{cond}(A))^2.$$

6. $\text{cond}(U) = \|U\| \cdot \|U^{-1}\| = 1 \cdot 1 = 1$. Далее, по свойству 4

$$\text{cond}(AU) \leq \text{cond}(A) \cdot \text{cond}(U) = \text{cond}(A).$$

Но $A = (AU)U^*$, поэтому

$$\text{cond}(A) \leq \text{cond}(AU) \cdot \text{cond}(U^*) = \text{cond}(AU).$$

Отсюда $\text{cond}(AU) = \text{cond}(A)$, и точно так же $\text{cond}(UA) = \text{cond}(A)$. ■

Замечание. Бытует суеверие (нелепое, как все суеверия), связывающее погрешности решения системы линейных алгебраических уравнений с величиной определителя ее матрицы коэффициентов. Из свойства 3 видно, что произвольно меняя величину этого определителя (умножая матрицу коэффициентов на различные числа), мы не меняем число обусловленности. Приведенный же выше пример показывает, что не меняя величину определителя ($\det \begin{bmatrix} 1 & 0 \\ N & 1 \end{bmatrix} = 1$ при любом N), мы можем получить сколь угодно большое число обусловленности. Этот пример также показывает, что большое число обусловленности можно получить и у матрицы малого размера.

Серьезное предупреждение. Подчеркнем, что погрешности исходных данных не могут быть известны по определению. Обычно удается получить лишь некоторую их оценку (например, оценку сверху относительной погрешности $\delta b = \|\Delta b\|/\|b\| \leq \varepsilon$). Соответственно, и для погрешности решения можно получить лишь оценку сверху (например, $\delta x = \|\Delta x\|/\|x\| \leq \text{cond}(A) \cdot \varepsilon$). Получение точного результата при "зашумленных" исходных данных невозможно. Поэтому трансформированную погрешность часто называют "неустранимой".

13.4. Факторизация матриц и число обусловленности

Как уже упоминалось, при решении задач линейной алгебры используются разложения матриц на множители специального вида. В связи с этим рассмотрим вопрос о влиянии такого разложения на трансформированные погрешности.

Итак, пусть система (13.3.1) решается с помощью факторизации матрицы A :

$$A = A_1 \cdot A_2,$$

т.е. вместо системы (13.3.1) решаются последовательно две системы:

$$A_1y = b, \quad A_2x = y.$$

Если вектор b имеет погрешность Δb , то вектор y будет получен с погрешностью $\Delta y = A_1^{-1} \cdot \Delta b$, а вектор x – с погрешностью

$$\Delta x = A_2^{-1} \cdot \Delta y = A_2^{-1} \cdot A_1^{-1} \cdot \Delta b = (A_1 A_2)^{-1} \cdot \Delta b = A^{-1} \Delta b.$$

Казалось бы, погрешность результата не зависит от способа факторизации. Однако в наших выкладках мы упустили из виду, что при решении первой системы к трансформированной погрешности прибавится погрешность метода $\widetilde{\Delta y}$, вызванная неточностью машинной арифметики. При решении второй системы эта погрешность играет роль погрешности в исходных данных и порождает дополнительную трансформированную погрешность $A_2^{-1} \cdot \widetilde{\Delta y}$. Поэтому большое значение имеют числа обусловленности матриц-сомножителей, и их произведение естественно считать критерием качества факторизации.

По свойству **4** числа обусловленности

$$\operatorname{cond}(A_1) \cdot \operatorname{cond}(A_2) \geq \operatorname{cond}(A_1 A_2) = \operatorname{cond}(A),$$

т.е. никакая факторизация не может улучшить "плохую" матрицу. Однако неудачная факторизация может "испортить" даже хорошо обусловленную матрицу.

Пример. Произведем LU -разложение унитарной матрицы (ее число обусловленности равно единице)

$$A = \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ \sin(\varphi) & \cos(\varphi) \end{bmatrix} = LU = \begin{bmatrix} 1 & 0 \\ \operatorname{tg}(\varphi) & 1 \end{bmatrix} \cdot \begin{bmatrix} \cos(\varphi) & -\sin(\varphi) \\ 0 & 1/\cos(\varphi) \end{bmatrix}.$$

При $0 < \varphi < \frac{\pi}{2}$ положим в примере из п.13.3 $N = \operatorname{tg}(\varphi)$. Тогда $\operatorname{cond}(L) > 1 + \operatorname{tg}^2(\varphi) = 1/\cos^2(\varphi)$. Аналогичные вычисления дают $\operatorname{cond}(U) > 1/\cos^2(\varphi)$. При значениях φ , близких к $\frac{\pi}{2}$, получаются очень плохо обусловленные матрицы. Этот пример показывает, в частности, что за исключением специальных случаев не следует строить LU -разложение без выбора ведущего элемента.

Хорошими свойствами обладает QR -разложение. Из свойства **6** числа обусловленности имеем

$$\operatorname{cond}(Q) = 1; \quad \operatorname{cond}(R) = \operatorname{cond}(Q^* A) = \operatorname{cond}(A).$$

Итак, $\operatorname{cond}(Q) \cdot \operatorname{cond}(R) = \operatorname{cond}(A)$.

Так же доказывается, что в сингулярном разложении $A = U\Sigma V^*$ имеет место равенство

$$\operatorname{cond}(U) \cdot \operatorname{cond}(\Sigma) \cdot \operatorname{cond}(V^*) = \operatorname{cond}(A).$$

Приведем еще пример, когда "хорошей" оказывается модификация LU -разложения. Пусть A – самосопряженная положительно определенная матрица. Тогда *можно показать*, что процесс LDU -разложения можно вести без перестановок строк и столбцов, причем это разложение имеет вид $A = U^*DU$. Если $x \neq \theta$, то

$$\langle Dx, x \rangle = \langle (U^*)^{-1}AU^{-1}x, x \rangle = \langle A \cdot (U^{-1}x), U^{-1}x \rangle > 0,$$

и, в частности, $d_{jj} = \langle De^{(j)}, e^{(j)} \rangle > 0$, $j = 1, \dots, n$.

Обозначив $D^{1/2} = \operatorname{diag} [\sqrt{d_{11}}, \dots, \sqrt{d_{nn}}]$, получим

$$A = U^*D^{1/2}D^{1/2}U = H^*H, \quad (13.4.1)$$

где $H = D^{1/2}U$ – верхняя треугольная матрица.

Имеем

$$\operatorname{cond}(H) = \frac{\sigma_{\max}(H)}{\sigma_{\min}(H)} = \left(\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \right)^{1/2} = (\operatorname{cond}(A))^{1/2}.$$

Формула (13.4.1) называется *разложением Холецкого*⁶⁸ *положительно определенной матрицы*, а основанный на ней метод решения линейных систем – *методом Холецкого* или *методом квадратного корня*.

⁶⁸Андре-Луи ХОЛЕЦКИЙ (A.-L. Cholesky, 1875-1918) – французский военный геодезист. Изобретенный им алгоритм широко применялся при решении задач геодезии, но опубликован был лишь после смерти автора, в 1924 г.

Глава 14. ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ СИСТЕМ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ

14.1. Метод простой итерации

Все до сих пор рассматривавшиеся нами методы решения систем относились к классу *прямых* методов: они приводили к решению за конечное число шагов. Здесь мы рассмотрим один из *итерационных* методов⁶⁹ – метод *простой итерации*.

Пусть система n линейных уравнений с n переменными записана в виде

$$x = Ax + b. \quad (14.1.1)$$

Начиная с произвольного вектора $x^{(0)}$, построим последовательность

$$x^{(1)} = Ax^{(0)} + b, \dots, x^{(k)} = Ax^{(k-1)} + b, \dots \quad (14.1.2)$$

Теорема. Если $\|A\| = q < 1$, то система (14.1.1) имеет единственное решение, причем для любого $x^{(0)}$ последовательность (14.1.2) сходится к этому решению.

Доказательство. Рассмотрим однородную систему $(I - A)x = \theta$. Если x – ее решение, то

$$x = Ax \implies \|x\| = \|Ax\| \leq q\|x\| \implies x = \theta.$$

Поэтому $\det(I - A) \neq 0$, и система (14.1.1) имеет единственное решение. Обозначим его \tilde{x} и вычтем из (14.1.2) равенство $\tilde{x} = A\tilde{x} + b$. Получим

$$x^{(k)} - \tilde{x} = Ax^{(k-1)} - A\tilde{x},$$

откуда

$$\|x^{(k)} - \tilde{x}\| = \|A(x^{(k)} - \tilde{x})\| \leq q \cdot \|x^{(k-1)} - \tilde{x}\|, \quad (14.1.3)$$

т.е.

$$\|x^{(k)} - \tilde{x}\| \leq q^k \cdot \|x^{(0)} - \tilde{x}\| \rightarrow 0 \quad \text{при } k \rightarrow +\infty. \quad \blacksquare$$

Теорема доказана и, кроме того, установлено, что последовательность (14.1.2) сходится к решению не медленнее, чем геометрическая прогрессия со знаменателем q .

⁶⁹Ранее мы уже рассмотрели один итерационный метод – метод Якоби решения полной проблемы собственных значений (см. п.8.2).

Из (14.1.3) и неравенства треугольника имеем

$$\begin{aligned}\|x^{(k-1)} - \tilde{x}\| &\leq \|x^{(k-1)} - x^{(k)}\| + \|x^{(k)} - \tilde{x}\| \leq \\ &\leq \|x^{(k-1)} - x^{(k)}\| + q \cdot \|x^{(k-1)} - \tilde{x}\|,\end{aligned}$$

откуда

$$\|x^{(k-1)} - \tilde{x}\| \leq \frac{\|x^{(k-1)} - x^{(k)}\|}{1 - q}.$$

Таким образом, прекратив итерации, когда норма разности двух соседних приближений станет меньше, чем $\varepsilon(1 - q)$, мы получим решение с погрешностью не большей, чем ε (по норме!).

Замечание. *Можно показать, что однозначная разрешимость системы (14.1.1) и сходимость процесса простых итераций к решению обеспечивается, если модули всех собственных чисел матрицы A меньше единицы. При этом, правда, не работает оценка (14.1.3).*

Существуют различные методы сведения общей системы линейных уравнений с квадратной матрицей к виду (14.1.1).

Пример. Пусть $B = B^*$ – положительно определенная матрица. Систему $Bx = b$ преобразуем так:

$$Bx = b \iff x = (I - \alpha B)x + \alpha b.$$

Здесь $\alpha > 0$ – пока что произвольное число. Обозначим $A = I - \alpha B$. Тогда $\lambda(A) = 1 - \alpha\lambda(B) < 1$ (поскольку B положительно определена).

Видно, что с ростом α собственные числа матрицы A "ползут" по числовой оси влево от единицы. Поскольку A эрмитова,

$$\|A\| = \sigma_{max}(A) = \max(|\lambda_{max}(A)|, |\lambda_{min}(A)|).$$

Наименьшего значения $\|A\|$ достигает, когда $\lambda_{min}(A) = -\lambda_{max}(A)$, т.е. при

$$1 - \alpha\lambda_{max}(B) = -(1 - \alpha\lambda_{min}(B)).$$

Это дает $\alpha = \frac{2}{\lambda_{max}(B) + \lambda_{min}(B)}$, и

$$\|A\| = \lambda_{max}(A) = 1 - \frac{2\lambda_{min}(B)}{\lambda_{max}(B) + \lambda_{min}(B)} = 1 - \frac{2}{1 + cond(B)}. \quad (14.1.4)$$

В последнем равенстве учтено, что для положительно определенной матрицы

$$cond(B) = \frac{\sigma_{max}(B)}{\sigma_{min}(B)} = \frac{\lambda_{max}(B)}{\lambda_{min}(B)}.$$

Анализ (14.1.4) показывает, что при большом числе обусловленности матрицы коэффициентов рассматриваемой системы знаменатель геометрической прогрессии в (14.1.3) близок к единице, т.е. итерации сходятся медленно, а за счет вычислительных погрешностей могут и расходиться.

Ни какими методами нельзя хорошо решить плохую систему!

На практике ситуация осложняется незнанием собственных чисел матрицы коэффициентов системы. Вместо них обычно используются некоторые их оценки. Подробное рассмотрение этой проблемы выходит за рамки нашего курса, как и рассмотрение других, более сложных итерационных методов.

14.2. Итерационное уточнение решений систем линейных алгебраических уравнений

В главе 1 был подробно рассмотрен один из прямых методов решения системы $Ax = b$ – метод Гаусса–Йордана. При этом предполагалось, что арифметические операции выполняются точно. В реальном компьютере вследствие конечности разрядной сетки это условие нарушается. Поэтому найденный *любым* прямым методом вектор $x^{(0)}$ не будет, вообще говоря, решением системы ($Ax^{(0)} \neq b$), и невязка окажется отличной от нуля:

$$d^{(0)} = b - Ax^{(0)} \neq \theta.$$

Попробуем подобрать такой вектор Δx ("добавку" к $x^{(0)}$), чтобы выполнялось равенство

$$A(x^{(0)} + \Delta x) = b.$$

Искомый вектор $\Delta x^{(0)}$ найдем, решая систему

$$A\Delta x = b - Ax^{(0)} = d^{(0)}$$

с той же матрицей A , тем же прямым методом.

Понятно, что из-за неточности машинной арифметики полученный вектор $x^{(1)} = x^{(0)} + \Delta x^{(0)}$ также не будет, вообще говоря, решением системы ($Ax^{(1)} \neq b$). Найдем новую невязку, и т.д.

Мы построили итерационный процесс

$$d^{(k)} = b - Ax^{(k)}, \quad x^{(k+1)} = x^{(k)} + \Delta x^{(k)}.$$

Здесь $\Delta x^{(k)}$ – полученное прямым методом "решение" системы $A\Delta x^{(k)} = d^{(k)}$.

Замечания. 1. Вычисление невязок должно выполняться в арифметике повышенной точности, иначе итерации не обеспечат уточнения.

2. Если прямой метод основан на какой-нибудь факторизации матрицы коэффициентов системы, то эту факторизацию – самую трудоемкую часть работы – следует выполнить один раз, полученные сомножители хранить и использовать на каждом шаге итерации.

3. Подробное исследование описанного процесса уточнения решения показало, что для не очень плохо обусловленных систем он сходится чрезвычайно быстро: три-четыре итерации обеспечивают получение решения с машинной точностью. Отсутствие же сходимости свидетельствует об очень большом числе обусловленности, что обычно означает, что эту систему решать не следует.

Мы настоятельно рекомендуем использовать
для решения систем линейных уравнений
только библиотечные программы
с итерационным уточнением.