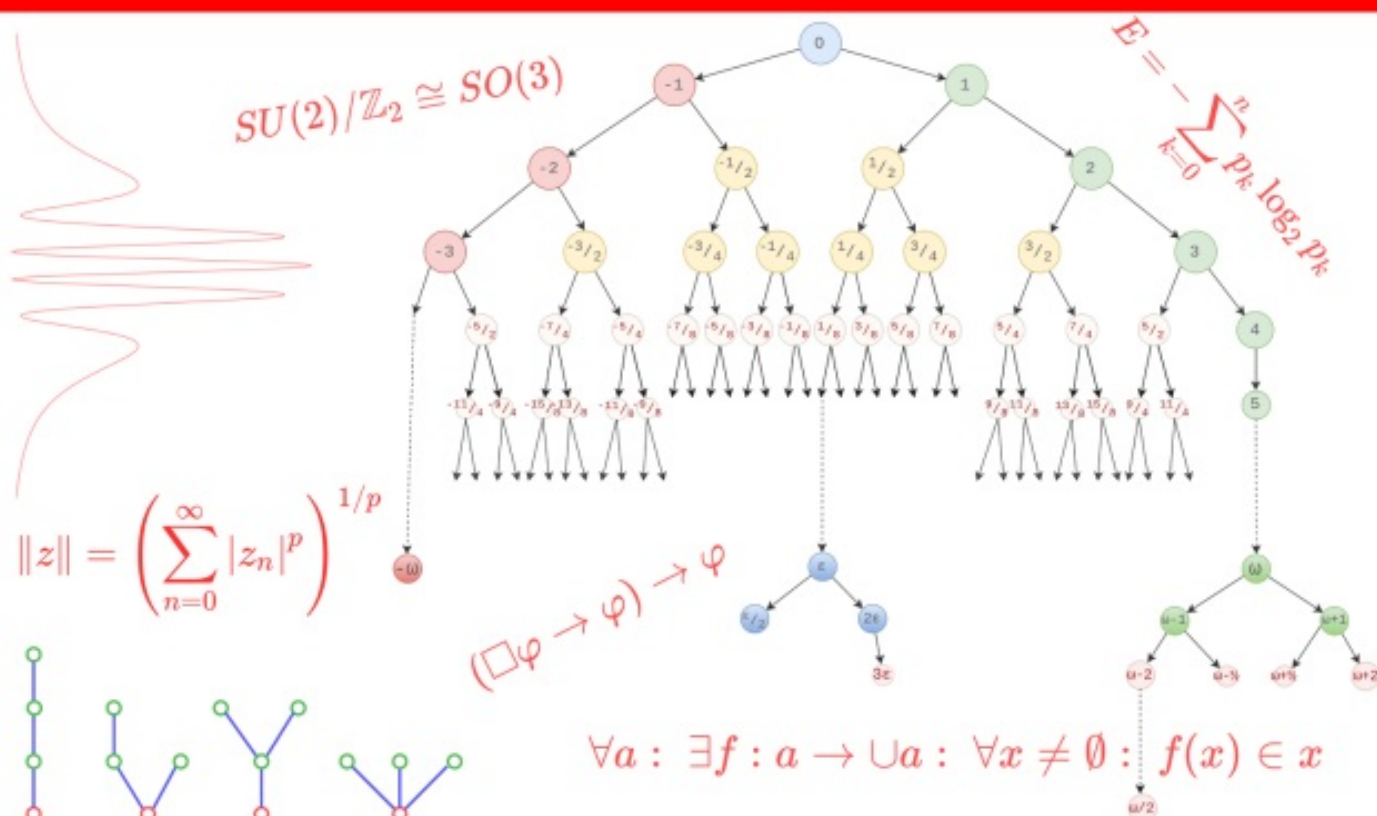


АРХЕТИПЫ МАТЕМАТИКИ

$$e^{i\pi} + 1 = 0$$



общие методы, приемы,
конструкции, идеи
математики и ее оснований

Н. И. Казимиров

АРХЕТИПЫ МАТЕМАТИКИ

общие методы, приемы,
конструкции, идеи
математики и ее оснований

Н. И. Казимиров

Москва, ЮСТИЦИНФОРМ, 2019

УДК 510.8
ББК 22.1
К14

Казимиров Н. И.

К14 Архетипы математики: общие методы, приемы, конструкции, идеи математики и ее оснований. — М.: Юстицинформ, 2019. — 612 с. : ил.

ISBN 978-5-7205-1599-7

В книге на примере изучения логики, теории множеств, алгебры, геометрии, анализа, теории графов рассматриваются некоторые общие характерные приемы и методы конструирования математических объектов и теорий, так называемые **архетипы**. Даются их наименования, описание и ссылки на них по ходу изложения материала. Сам материал представлен как на наивном, так и на современном строгом уровне, ряд теорем приводится с доказательством, подробно рассматривается построение конечных множеств и логики первого порядка с целью донести до широкого круга читателей возможность компьютерного моделирования математических доказательств. Вместе с тем, делается акцент на необходимости участия человека в этом процессе. В книге также дается ряд примеров из теории чисел, в частности, доказательство Великой Теоремы Ферма для случая $n = 3, 4$, подробно рассмотрена теорема Гудстейна. Кроме того, проводится детальное описание движений в геометрии, всесторонне рассматривается аксиома выбора и ее влияние на математику, приводится ряд примеров из теории случайных графов с анализом их свойств.

Официальный сайт книги: <https://mathem.at>.

Ключевые слова: теория множеств, основания математики, алгебра, теория групп, топология, теория вероятностей, теория графов, геометрия, теорема Гудстейна, теоремы Геделя, квантовые вероятности.

УДК 510.8
ББК 22.1

© Казимиров Н. И.

ISBN 978-5-7205-1599-7

© ООО «Юстицинформ», 2019

ОГЛАВЛЕНИЕ

Введение	9
Глава 1. Множества и мультимножества	13
Глава 2. Числа	107
Глава 3. Дальнейшие обобщения чисел	195
Глава 4. Матлогика. Исчисления I	335
Глава 5. Анализ. Исчисления II	413
Глава 6. Графы	531
Финал	583
Приложение А. Указатель архетипов	587
Приложение В. Схемы и таблицы	591
Приложение С. Листинги программ	597
Список литературы	603

Введение	9
Глава 1. Множества и мультимножества	13
1.1 Начальные множества	13
1.1.1 Грамматика теории множеств	13
1.1.2 Рекурсия и индукция	25
1.1.3 Первые архетипы	26
1.1.4 Финальные дополнения к языку	27
1.1.5 Равенство, принадлежность и объекты теории	29
1.1.6 Скобочная запись множеств	33
1.1.7 Числа грамматики	39
1.1.8 «Начальные» мультимножества	41
1.1.9 Универсальные множества и универсальный код	44
1.1.10 Теорема Гудстейна	50
1.2 Аксиоматика Цермело–Френкеля	55
1.2.1 Равенство и единственность	55
1.2.2 Ограничительные аксиомы	59
1.2.3 Первая бесконечность	65
1.3 Основные инструменты	71
1.3.1 Порядок отношений и отношения порядка	71
1.3.2 Трансфинитная рекурсия	84
1.3.3 Перечень инструментов	98
1.4 Универсумы и мультимножества	99
Глава 2. Числа	107
2.1 Арифметика порядковых чисел	108
2.1.1 Сложение	108
2.1.2 Умножение	111
2.1.3 Степень	115
2.1.4 Разложения ординалов	117
2.1.5 Доказательство теоремы Гудстейна	119
2.1.6 «Начальные» мультимножества	120
2.2 Кардинальная арифметика	121
2.3 Немного теории чисел	125
2.4 Числовые структуры	132
2.4.1 Группы, кольца, поля	134

2.4.2	Целые числа	139
2.4.3	Рациональные числа	144
2.4.4	Поле действительных чисел	149
2.4.5	Гипердействительные числа	157
2.4.6	Сюрреальные числа	168
2.4.7	Поле комплексных чисел	181
Глава 3.	Дальнейшие обобщения чисел	195
3.1	Окно в общую алгебру	195
3.1.1	Архетип переноса свойств на базовое множество	196
3.1.2	Алгебраические структуры	197
3.1.3	Алгебра множеств	199
3.1.4	Немного теории групп	204
3.1.5	Идеалы, модули, базис	218
3.2	Матричное представление чисел	232
3.3	Гауссовы целые числа	238
3.3.1	Делимость и простые числа	239
3.3.2	Некоторые приложения гауссовых чисел	242
3.4	Целые числа Эйзенштейна	244
3.5	Линейные пространства и операторы	249
3.6	Экскурс в геометрию	261
3.6.1	Движения в действительном пространстве	264
3.6.2	Движения в \mathbb{R}^3 , кватернионы	268
3.6.3	Подобия в действительных пространствах	280
3.6.4	Аффинные преобразования и однородные координаты	281
3.6.5	Проективная геометрия	283
3.6.6	Нестандартные геометрии	295
3.6.7	Геометрия на сфере	301
3.7	Многочлены	308
3.7.1	Конечные алгебраические расширения	310
3.7.2	Свойства многочленов	314
3.7.3	О построениях циркулем и линейкой	320
3.7.4	Нормальные расширения	322
3.7.5	Элементы классической теории Галуа	323
3.8	Группы: завершающий аккорд	329
3.9	Числовые архетипы	331
Глава 4.	Матлогика. Исчисления I	335
4.1	Исчисление высказываний и предикатов	335
4.1.1	Исчисление высказываний	335
4.1.2	Исчисление предикатов	340
4.1.3	Модели	349

4.1.4	Ультрастепени	353
4.1.5	Примеры формальных теорий	356
4.2	Аксиома выбора: полезная и странная	366
4.2.1	Полезные следствия аксиомы выбора	375
4.2.2	Странные следствия аксиомы выбора	377
4.2.3	О сверхбольших кардиналах	381
4.2.4	Конкурент АС: аксиома детерминированности	385
4.2.5	Промежуточный итог	389
4.3	Вычислимость и доказуемость	390
4.3.1	Нестрогое введение в теорему Гёделя о неполноте	390
4.3.2	О вычислимости, разрешимости, перечислимости	391
4.3.3	Про сигма-определимость	396
4.3.4	Доказательство теорем Гёделя	399
4.3.5	Несколько слов о рекурсиях	405
Глава 5.	Анализ. Исчисления II	413
5.1	Пространства и отображения	413
5.1.1	Связность и непрерывность	414
5.1.2	Непрерывность и группы	421
5.1.3	Фильтры и пределы	429
5.1.4	Компактность	430
5.1.5	Метрика	431
5.1.6	Норма и скалярное произведение	436
5.1.7	Ограниченные операторы	440
5.1.8	Псевдометрика	442
5.1.9	Мера и интеграл	445
5.2	Преобразования пространств	456
5.3	Конечные разности и вариации	460
5.3.1	Смещения и разности	460
5.3.2	Пространство смещений	461
5.3.3	Целочисленные разности	464
5.3.4	Вариации и дифференциалы	472
5.3.5	Вариационное исчисление	476
5.4	Вероятности	492
5.4.1	Классическая вероятность	492
5.4.2	Квантовая вероятность	508
5.4.3	Общий алгебраический подход	523
5.5	Несколько слов об общей картине	529
Глава 6.	Графы	531
6.1	Подходы к определению	531
6.2	Обычные графы	538

6.2.1	Разнообразие деревьев	548
6.3	Вероятности на графах	559
6.3.1	Модель Эрдёша–Реньи	562
6.3.2	Случайные деревья и леса	567
Финал		583
Приложение А. Указатель архетипов		587
Приложение В. Схемы и таблицы		591
Приложение С. Листинги программ		597
Список литературы		603

- A1. Данная книга является учебником.
- A2. Данная книга не является учебником.
- A3. Утверждения A1 и A2 не противоречат данной книге.
- A4. Невозможно доказать, что утверждение A3 невозможно опровергнуть.

Знатокам и любителям матлогики уже понятно, о чем идет речь в этой книге. Но это лишь часть предлагаемого к обзору материала. На самом деле, книга содержит широкий спектр тем, связанных с математическим знанием и отчасти с анализом самой математики как предмета изучения. Здесь предлагается широкими мазками ознакомиться с внутренней картиной математики, заглянув в такие ее уголки как теория множеств, алгебра и топология, а также некоторые аспекты компьютерной науки и прикладной математики.

Книга носит нестандартный для математической литературы формат и структуру, поскольку преследует нестандартные цели. Здесь мы постараемся посмотреть на математические конструкции через общие идеи и методы, с помощью которых можно объединить схожие друг с другом понятия из различных математических дисциплин. Это объединение фиксируется с помощью термина «архетип».

Автор не ставит перед собой задачу сделать академическое описание или длинное научное исследование по всем канонам философской или математической мысли. Задача данной книги — показать многогранность, единство и красоту математики всем, кто ею интересуется: от студентов до работающих математиков и приверженцев естественных наук.

В главе 1 мы занимаемся довольно тщательным построением языка теории множеств и построением самой теории. Сначала мы оперируем «наивными» понятиями множества и операций над множествами, вводим грамматику языка и изучаем полученные синтаксические конструкции. Особое внимание уделено понятию равенства как свойства языка. С помощью равенства мы отделяем понятие объекта теории от его формального написания и начинаем оперировать множествами в их обычном математическом понимании, а также вводим в обращение натуральные числа. Попутно рассматриваются мультимножества и некоторые их свойства. Раздел первой главы, посвященный теории «начальных» множеств, мы заканчиваем обсуждением теоремы Гудстейна, которая вынуждает нас ввести понятие актуально бесконечного числа.

Глава 1 примечательна тем, что в ней мы погружены в один-единственный формализм, который являет собой мир множеств Цермело–Френкеля. Позже мы увидим, что математику, как исследователю, вовсе не обязательно находиться в одном этом мире, что таких миров может быть много, и каждый волен выбирать себе математический мир по своему вкусу в зависимости от задач и целей исследования. Важно лишь то, что все эти миры удивительным образом согласованы друг с другом через язык и логику.

Глава 2 посвящена «линейным» числам, т. е. числам, упорядоченным отношением линейного порядка. Начиная от натуральных чисел и ординалов, мы продвигаемся через рациональные и действительные к сюрреальным числам. Единственное исключение в этой главе — комплексные числа, обойти которые было бы непростительно и контрпродуктивно, а выносить их в главу 3 — не совсем удобно, т. к. в этой главе комплексные числа появятся как часть метода построения новых чисел с помощью матриц. Поэтому их определение и ряд свойств освещаются также в главе 2.

В главе 3 мы уже довольно серьезно погружаемся в алгебру, рассматриваем ряд алгебраических конструкций и на их основе изучаем некоторые числовые структуры, которые на первый взгляд трудно отнести к числам (например, многочлены и матрицы). Тем не менее, у них примерно столько же общего с натуральными числами, сколько и у ординалов или сюрреальных чисел, а дополнительные измерения не являются геометрическим препятствием. Тем более что и внутри действительных чисел мы увидим конечные и бесконечномерные пространства над числами рациональными. Тесная взаимосвязь всех этих чисел через алгебраические свойства будет прослеживаться на протяжении всей третьей главы.

В дальнейшем мы выходим из теоретико-числовой канвы и принимаемся изучать сами методы изучения, в частности, формальные теории и связанные с ними языки и исчисления.

Глава 4 преимущественно посвящена основаниям математики и связанным с ними исчислениям. По сути, все это — набор методов работы с числами и их символикой, набор алгоритмов и задач, посвященных проблемам разрешимости, т. е. ситуациям, когда та или иная задача может быть детерминирована, а ее решение может быть получено за конечное число простых действий. По большому счету, мы наблюдаем здесь тесную взаимосвязь между основаниями математики и Computer Science, поскольку все определения строгости языка той или иной теории по сути своей сводятся к умению запрограммировать эту теорию на вычислительной машине и с ее помощью получать детерминированные ответы на задачи данной теории.

В главе 5 мы поднимаемся обратно в математический мир и рассматриваем тот аналитический аппарат, который в основном лежит в мире действительных чисел и непрерывных объектов. Здесь нас тоже интересуют исчисления, однако связанные более с физико-математическими задачами, чем с

формальной логикой. Но и в этом случае исчисления остаются строгим детерминированным аппаратом, служащим для извлечения пользы при решении строго описанных задач. Примечательно, что само открытие и построение такого аппарата — есть чисто математический, творческий процесс, но когда аппарат исчислений приобретает законченные черты и плавно переходит в область преподавания на 1–2 курсе естественно-научных и инженерных специальностей, математики начинают терять к нему интерес и переключаются на поиск новых пространств для исследования.

В главе 6 рассматриваются графы. Несмотря на то, что графы будут встречаться и раньше в качестве вспомогательных конструкций, было решено вынести их рассмотрение в отдельную главу, чтобы иметь возможность разобраться в самом понятии графа более подробно. Здесь мы вынуждены предупредить читателя о том, что некоторые данные нами определения, связанные с теорией графов, могут оказаться слишком надуманными и сложными. Тем не менее, такое развитие классических определений графов требует лейтмотив книги, поэтому было бы неправильно оставить их в стороне.

Помимо собственно графов, в этой главе мы рассматриваем несколько примеров случайных графов, востребованность которых в науке в последнее время только растет, прежде всего, в связи с изучением информационных и социальных сетей, методов машинного обучения.

Главу 6 изначально планировалось дополнить разделом, посвященным теории категорий, поскольку нам представляется, что между графами и категориями имеется некая глубинная архетипическая связь. Однако рамки книги и многочисленные отсылки в тексте к теории Гёделя–Бернайса с классами множеств, которые могут рассматриваться как предтеча категорий, заставили исключить этот материал. То же самое можно сказать и о теории типов.

На протяжении всей книги мы придерживаемся следующей парадигмы: чем больше номер страницы, тем менее подробными (в среднем) будут доказательства. То же самое относится к языку: постепенно мы будем отходить от языка, заданного грамматикой в первой главе, и проявлять вольности в обозначениях. Демонстрировать строгость и полноту доказательств полезно, но если это делать на протяжении сотен страниц — это сильно утомляет и автора, и читателя. К тому же за нагромождением формул можно не увидеть картину в полном объеме. Тем не менее, в каждой из главных тем книги (основания, алгебра, графы) найдутся теоремы и доказательства, по сложности превышающие средний университетский уровень.

| *За деревьями
леса не видно*

Одна из целей книги — показать ряд выдающихся математических результатов и подходы к их доказательствам. Детальное изложение этих доказательств занимает порой тома, превосходящие нашу книгу в несколько раз. Но общую методику, наметки идей мы показать можем. Тем более что использованные приемы зачастую сами по себе могут быть причислены к

архетипам.

Книга снабжена рекомендациями по самостоятельным упражнениям, которые отмечены на микроврезках (например, | *Упражнение 1.1*) и иногда сопровождаются пояснительным текстом. Внимание! В книге нет ответов к упражнениям! К некоторым упражнениям ответов нет и в каких-либо других книгах. Мы считаем, что такой подход поспособствует свободному поиску читателем необходимого материала и более глубокому его изучению.

Комментарий 1.

В книге присутствуют комментарии, в которых приводятся примеры, прямо или косвенно связанные с изучаемой темой. Часть из них посвящена задачам машинного обучения.

Жирным шрифтом выделяются определяемые или особо важные термины и высказывания, в том числе **архетипы**. Мы не выделяем определения так же явно, как теоремы, поскольку считаем читателя достаточно искушенным в математической терминологии и символике. Некоторые определения и термины выделены *курсивом*, поскольку не имеют прямого отношения к контексту.

В ряде мест имеются ссылки на видеолекции, находящиеся в открытом доступе. Эти ссылки выполнены гиперссылками в pdf-версии книги и, кроме того, после каждой из них для читателей бумажной версии в квадратных скобках помещен код видеоролика, по которому его легко найти на youtube. Например, [879xeBir-2I].

В конце книги приводится список определенных в ней архетипов со ссылками на страницы, где они впервые определяются. В то же время, мы не приводим по традиции список всех терминов и обозначений, использованных в книге, для этого достаточно взять, например, справочник Корна [58], в крайнем случае — открыть Wikipedia.

Список литературы включает те книги, которые, по мнению автора, лучше всего подходят к заявленным в книге темам. Этот список ни в коем случае не является исчерпывающим (его можно было бы увеличить раз в 10), но вполне репрезентативным.

Цветная PDF-версия книги находится на сайте <https://mathem.at>.

Автор выражает искреннюю благодарность своим друзьям, коллегам и участникам математического форума dxdy.ru за разъяснения некоторых вопросов и конструктивную критику, а также своей супруге Диане — за терпение и понимание.

Множества и мультимножества

1.1 Начальные множества

Подобно тому, как Г. Кантор начинал развивать теорию множеств с так называемой «наивной теории множеств», и лишь потом его последователями была построена аксиоматическая теория, мы начнем наши построения с теории «начальных» множеств. Для этого рассмотрим простую задачу: дать понятие множества наиболее примитивным и в то же время формальным способом, привлекая как можно меньше сущностей, так, чтобы *научить компьютер теории множеств*. | Задача —
проще не
бывает!

Как и в «наивной» теории множеств мы полагаем, что множество есть совокупность некоторых элементов. Но при этом мы оговариваемся сразу, что и элементами множества всегда будут множества, и совокупность множеств, построенная по определенным правилам, также всегда будет множеством, и все функции и отношения на множествах также будут множествами. Иначе говоря, в нашей теории *всё есть множество*! | Кроме самой
теории,
разумеется

1.1.1 Грамматика теории множеств

Первое, с чем нужно определиться, — это **язык**. Мы хотели бы, чтобы язык теории был как можно проще и строже одновременно, а значит, мы будем использовать ограниченный перечень символов (алфавит) для изложения нашей теории. Теория оперирует записями, которые представляют собой наборы этих символов, т. е. текстовые строки. Записи еще называют словами над алфавитом теории.

Для построения языка мы воспользуемся формальной грамматикой в форме Бэкуса–Наура. Грамматика языка — это набор правил, по которым формируются «правильные» записи на этом языке. При этом не имеет никакого значения, что ставится в соответствие этим записям как в самой математике, так и в других науках, за одним исключением: формулы имеют смысл именно формул, т. е. объектов изучения исчисления предикатов. Это требование

связано с тем, чтобы мы могли использовать всю мощь аппарата математической логики в нашей простейшей формальной теории.

Грамматика строится по правилам, в которых рекурсивным способом записывается построение понятий и утверждений, начиная от простейших грамматических единиц, т.е. символов алфавита языка. Например, правило «Формула \longrightarrow Формульная переменная» говорит нам о том, что всякая формульная переменная является формулой, а правило «Формула \longrightarrow (Формула)» — что всякая формула, взятая в круглые скобки, является также формулой. При этом исходные символы алфавита мы относим к терминальным, а словесные обозначения понятий — к нетерминальным символам.

Итак, определим следующую грамматику теории множеств.

Алфавит:

- A1 (терминальный алфавит):¹ $\in, =, \{, \}, (,), :, ', ', 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, \alpha, \beta, \gamma, \delta, \varepsilon, \zeta, \eta, \theta, \varpi, \lambda, \mu, \nu, \xi, \rho, \sigma, \tau, \varphi, \chi, \psi, \neg, \wedge, \vee, \longrightarrow, \longleftrightarrow, \forall, \exists$
- A2 (нетерминальный алфавит): Цифра, Число, Объектная переменная, Терм, Список, Формульная переменная, Логическая связка, Формула

Правила грамматики (символ | означает альтернативный выбор, позволяя сократить количество правил):

- G1 Цифра $\longrightarrow 0 \mid 1 \mid 2 \mid 3 \mid 4 \mid 5 \mid 6 \mid 7 \mid 8 \mid 9$
- G2 Число \longrightarrow Цифра | Число Цифра
- G3 Объектная переменная $\longrightarrow a \mid b \mid c \mid d \mid e \mid f \mid g \mid h \mid i \mid j \mid k \mid l \mid m \mid n \mid o \mid p \mid q \mid r \mid s \mid t \mid u \mid v \mid w \mid x \mid y \mid z \mid \alpha \mid \beta \mid \gamma \mid \delta \mid \varepsilon \mid \zeta \mid \eta \mid \theta \mid \varpi \mid \lambda \mid \mu \mid \nu \mid \xi \mid \rho \mid \sigma \mid \tau \mid \chi$
- G4 Список \longrightarrow Терм
- G5 Список \longrightarrow Список, Терм
- G6 Терм $\longrightarrow \{ \} \mid$ Число | Объектная переменная
- G7 Терм $\longrightarrow \{$ Список $\}$
- G8 Терм $\longrightarrow \{$ Переменная: Формула $\}$
- G9 Формульная переменная $\longrightarrow \varphi \mid \psi$
- G10 Формула \longrightarrow Формульная переменная

¹В этом списке есть запятая, для наглядности она взята в 'кавычки'

G11 Формула \longrightarrow Формульная переменная(Список)

G12 Формула \longrightarrow Терм \in Терм

G13 Формула \longrightarrow Терм = Терм

G14 Формула \longrightarrow (Формула)

G15 Логическая связка $\longrightarrow \wedge \mid \vee \mid \rightarrow \mid \leftrightarrow$

G16 Формула $\longrightarrow \neg$ Формула \mid Формула Логическая связка Формула

G17 Формула $\longrightarrow \forall$ Объектная переменная Формула $\mid \exists$ Объектная переменная Формула

Обсудим «работу» правил грамматики. Первые два правила являются правилами записи чисел, т.к. число есть любая цифра от 0 до 9, а также число есть два записанных подряд числа (пробелы в правилах грамматики проставлены только для удобства их чтения, они не участвуют в генерации грамматических конструкций). Таким образом, если мы сначала в качестве числа использовали только цифры, то правило G2 позволяет создать число из двух цифр, и у нас в запасе уже имеются числа из 1 и 2 цифр, что позволяет с помощью правила G2 сгенерировать числа из 3 и 4 цифр, составляя различные комбинации предыдущего этапа, и т.д. Одно и то же правило может применяться рекурсивно и многократно к результатам работы как самого себя, так и других правил. Таким образом, правила G1 и G2 позволяют сгенерировать за конечное число шагов любые цепочки цифр, например, 100132 или 0000012. Для простоты мы позволяем генерировать числа, начинающие свою запись с нулей. В то же время, мы не используем здесь обозначения дробных чисел. Для них пришлось бы ввести третье правило вида «Число \longrightarrow Число.Число». Однако у нас пока нет задачи описать грамматику математического анализа, поэтому записи ни десятичных, ни *Ну, слава Богу!* каких-либо других дробей мы не определяем.

Понятия Терм, Список и Формула тесно связаны между собой, т.к. каждое из них участвует в генерировании двух остальных. Но мы можем рассмотреть некоторую упрощенную часть всего объема термов, очень важную для нашего дальнейшего изложения.

Итак, Терм есть $\{\}$ (правило G6). Это краеугольный камень всей теории, т.к. данная запись олицетворяет такое понятие как «пустое множество». Далее, следуя правилам G4 и G5, мы понимаем, что список есть список термов, перечисленных через запятую. В частности, списком будет, например, такая запись: $\{\}, \{\}, \{\}$. Но списком также будет и запись пустого множества, т.е. $\{\}$, в силу правил G6 и G4.

Далее, правило G7 позволяет навешивать фигурные скобки на любой список термов, так что и $\{\{\}\}$, и $\{\{\}, \{\}, \{\}\}$ будут термами.

Из этих термов можно вновь образовать список (G4 и G5) и построить еще боле сложный терм (G7), например, такой: $\{\{\}, \{\{\}\}, \{\{\}, \{\}, \{\}\}\}$. Продолжая этот процесс генерации, мы получим всевозможные скобочные

Большинство
из них мы
потом
выбросим за
ненужностью

записи, которые, в общем-то, и будут записями «начальных» множеств.

Условимся в дальнейшем все грамматически верные записи, состоящие только из фигурных скобок и запятых, называть **объектными записями**. Стоит обратить внимание на то, что объектные записи, — это еще не сами «начальные» множества, т. е. объекты теории. К пониманию, что такое объект теории, мы сможем подойти лишь после подробного описания языка теории и введения понятия равенства.

Но, к сожалению, теория окажется очень невыразительной, если в ней будут отсутствовать **переменные**. Их мы вводим в грамматику с помощью правил G3 и G9. С точки зрения грамматики между числами, объектными записями и объектными переменными существует только одно отличие, которое дано в правиле G17 — только объектные переменные могут стоять под квантором! Но не все так просто...

На самом деле понятие переменной заслуживает того, чтобы остановиться на нем подробнее. Поскольку мы сейчас находимся в рамках задачи «научить компьютер теории множеств», то и понятие переменной следует определять «компьютерным» способом.

Часто в учебниках по программированию описывают переменную как некий ящик, в который можно положить все, что туда влезет. Под словом «влезет» понимается совместимость типов переменной и вкладываемого объекта (в целочисленную переменную можно положить только целое число и т.п.). Этот ящик, таким образом, становится вычисляемым объектом, и всякий раз, когда он где-то используется, из него извлекается объект и используется вместо ящика.

Но поскольку автору импонирует язык программирования Python с его ссылочной моделью, дадим определение понятия переменной как **именованной ссылки на терм** (мы сейчас пропустим формульные переменные, ссылающиеся на формулы, но по сути это примерно то же самое). Терм может быть либо объектной записью (скобки и запятые), либо более сложной грамматической конструкцией, включающей в себя другие переменные (а иногда и формулы). В любом случае терм — это грамматически верная текстовая строка, которая однажды сгенерирована по правилам грамматики и лежит где-то в памяти нашего компьютера.²

Когда мы вводим имя для переменной и присваиваем ей значение (терм), у нас создается именованная ссылка на это значение, и в дальнейшем имя

²Предполагается, что память не содержит дубликаты записей, хотя в результате каких-то операций над ними могут возникнуть дубли. Предполагается также, что в нашем компьютере памяти так много, что потребность в «чистильщике» отпадает.

этой ссылки используется как второе обозначение данного терма. Но лишь до того момента, когда ссылка будет переопределена в результате каких-то манипуляций с данной переменной. Две переменные–ссылки можно приравнять друг к другу (или доказать их равенство в некотором контексте), и в этом случае их использование внутри других термов и формул будет взаимозаменяемым. Но как только одну из них мы переопределим, взаимозаменяемость нарушится.

Объектные записи можно отнести к неизменяемым типам данных, а термы, содержащие переменные, — к изменяемым. Таким образом, запись $\{\{\}\}$, $\{\{\}\{\}\}$ — неизменяемая, а запись $\{a, b\}$ — изменяемая, она напоминает запись $\{a, b\}$ типа `set` в Python.³ Изменяемая запись сама содержит ссылки на другие объекты грамматики. При этом сама по себе такая запись не зависит от значений входящих в нее переменных. Эти значения начинают играть роль только в момент *вычисления* значения записи $\{a, b\}$. Иначе говоря, если у нас $c = \{a, b\}$, т.е. переменная c указывает на терм $\{a, b\}$, то нам неважно, на какие термы ссылаются переменные a и b , формально мы имеем дело только с конструкцией «фигурные скобки с двумя ссылками с именами a и b ».

Да-да, в Python есть множества!

Комментарий 1.

Кстати, если вспомнить квантовую физику, то параметры электрона тоже не определены, пока мы не произведем их измерение. В некотором смысле электрон — это терм с переменными, отвечающими за свойства электрона.

Таким образом, переменная не обозначает ни какой-то конкретный объект теории, ни какой-то конкретный класс объектов. Переменная временно хранит ссылку на некоторую произвольную запись на языке теории. Так, например, в цикле `For` любого языка программирования мы сталкиваемся с переменной, меняющей свои значения при смене итерации цикла. В этом смысле цикл `For` аналогичен квантору всеобщности. Когда мы говорим $\forall x(x \notin x)$, мы мысленно запускаем цикл по всем объектным записям и для каждой из них проверяем, что $x \notin x$. Но при этом внутри каждой итерации оба вхождения переменной x имеют одно и то же значение — не может получиться так, что в какой-то момент мы будем проверять истинность, например, такого высказывания: $(\{\} \notin \{\{\}\})$, где левое вхождение x приняло значение $\{\}$, а правое — $\{\{\}\}$.

Наряду с переменными мы используем **константы**. Это всего лишь пере-

³Строго говоря, сравнение не совсем верное. Дело в том, что Python во время присваивания $c = \{a, b\}$ вычислит значения переменных и положит в множество их значения, так что последующие изменения a и b не отразятся на значении c . В математике мы не вычисляем переменные до тех пор, пока это не потребуется, и поэтому значение переменных c следует судьбе переменных a и b .

обозначения некоторых объектных записей. Если объектная запись слишком длинная и сложная, а сам объект часто используется в теории, то для него *Например, \arcsctg :)* обычно вводится одно-двух-трехсимвольное обозначение. В каком-то смысле константы — это те же объектные переменные, только записанное в них значение запрещено менять. Для констант мы зарезервировали в грамматике Числа. Однако дальнейший список констант не будет ограничиваться только лишь числами. Но это будет уже за пределами теории «начальных» множеств.

В математической логике есть специальный символ $\|$, обозначающий замену переменных другими переменными или термами. Так, если в формуле или терме в списке аргументов мы встречаем запись вида $(a, b \| x, y)$, то это означает, что мы произвели замену всех свободных вхождений переменной a на переменную x , и всех свободных вхождений переменной b на переменную y . При этом под *свободным вхождением* переменной понимается тот факт, что она не стоит в этой формуле или терме под квантором или в записи вида $\{a : \varphi\}$, которая также связывает данную переменную.

Опять же, для упрощения нашей грамматики мы не будем определять символ $\|$ и «узаконивать» замену переменных. В то же время, легко доказать следующее утверждение (оно не относится к теории множеств, поэтому названо метатеоремой, подробнее об этом мы поговорим в главе 4):

Метатеорема 1.1 (о согласованности грамматики). *При замене свободных переменных в терме (формуле) на другие переменные или, в общем случае, термы, получается грамматически верный терм (формула).*

Действительно, если терм с переменной a построен грамматически верно, то можно предъявить последовательность итераций правил грамматики, приведших к его построению, отправляясь от элементарных конструкций языка⁴ — переменных, скобок и чисел. Если же теперь вместо переменной a подставить какой-либо произвольный терм, то это всего лишь удлинит процедуру построения итогового терма, но не выведет нас из поля правил грамматики. То же самое относится и к подстановке термов вместо объектных переменных в формулы языка.

Вернемся к обсуждению правил грамматики.

Аналогично построению объектных записей (из фигурных скобок) строятся простейшие **термы**. Действительно, если вместо $\{ \}$ мы начнем генерировать термы от переменной a , то по правилам G4–G7 легко построить термы: $\{a\}$, $\{a, a, a\}$, $\{a, \{a\}, \{a, a, a\}\}$. Продолжая пользоваться этими правилами, мы получим в точности такой же набор термов, как определенные выше объектные записи, только в них всюду вместо $\{ \}$ будет стоять переменная a . Точно так же можно построить «куст» грамматически верных записей, в ос-

⁴Это свойство грамматики называется *разрешимостью*.

нове которых будет лежать любая другая объектная переменная или число.

Например, для числа 10 мы можем строить записи:

$$\{10\}, \{10, \{10\}\}, \{10, \{10\}, \{10, \{10\}\}\}, \{10, 10, 10\}$$

и т.д. И до тех пор, пока мы не дали определение числам, мы вправе рассматривать их в качестве начальных (стартовых) записей, порождающих «куст» скобочных записей, полностью эквивалентный набору объектных записей. На этом, кстати, строится *теория множеств с атомами*, когда множеств без элементов (пустых) может быть сколь угодно много (как чисел в нашей грамматике), и ее обобщение — теория мультимножеств, где встречаются не только атомарные множества, но еще и кратности их вхождения в другие множества. Например, у мультимножества $\{10, 10, 10\}$ элемент 10 имеет кратность вхождения 3, а у мультимножества $\{1, 1, 1, 1, 1\}$ кратность вхождения элемента 1 равна 5.

На самом деле, все эти вещи относительно легко погружаются в обычную теорию множеств, которой мы здесь занимаемся. Достаточно вполне конкретные готовые множества обозначить константными символами — и вот вам модель теории множеств с атомами. Если провести дополнительные построения, то и мультимножества можно смоделировать в рамках обычной теории множеств. Позже мы это сделаем, когда будем строить универсумы множеств и мультимножеств.

При этом стоит отметить, что грамматика записей для этих трех видов теории множеств может использоваться одна и та же. Разницу между множествами-атомами, не имеющими элементов и искусственно введенными множествами-константами (которые имеют элементы) можно будет увидеть лишь при помощи корректно определенного равенства множеств и аксиоматики, связывающей отношения равенства и принадлежности.

Наконец, рассмотрим грамматическое понятие **Формула**.

Подобно символу пустого множества и числам у формул есть свои отправные точки для генерации грамматически верных конструкций. Это — формула принадлежности и формула равенства (правила G12 и G13). Для удобства мы можем сначала рассмотреть простейшие формулы, где в качестве термов используются переменные: $a \in b$ и $a = b$ (переменные a и b могут быть заменены любыми другими объектными переменными). Эти формулы называются **атомарными**.

Поскольку мы сразу объявляем такие конструкции формулами, т.е. объектами изучения математической логики, мы тем самым обязаны следовать соответствующему синтаксису формулостроения. И действительно, у нас появляется правило G15, которое вводит несколько символов-связок (\wedge — конъюнкция, \vee — дизъюнкция, \rightarrow — импликация, \leftrightarrow — эквиваленция), известных из логики, а правило G16 прямо постулирует, что из любых формул мож-

но собрать новую путем их соединения через логическую связку и отрицание (\neg). Таким образом, у нас появляются формулы вида $(a \in b) \wedge (a = b)$, $(a = b) \leftrightarrow (b = a)$ и т.п.

Кроме того, у нас появляются две переменные, зарезервированные для обозначения формул. Это φ и ψ (правило G9). С помощью правил G9–G11 мы можем записать конструкции вида φ , $\varphi(a, b)$, $\psi(a, b, c)$ и т.д., поскольку нам разрешается следом за формульной переменной приписывать Список термов. Глубокий смысл этих обозначений мы предполагаем известным читателю из мат.логики или мат.анализа. Здесь же отметим только, что данные грамматические конструкции позволяют для формулы указать список (не обязательно всех) свободных переменных, участвующих в формуле, а затем еще и воспользоваться заменой переменных и подстановкой вместо них любых грамматически верных термов, что позволяет записывать нам какие угодно высказывания о множествах.

Наконец, у нас остались три кванторных правила грамматики: G8, G16, G17. Два последних хорошо знакомы всем, они вводят в язык такие понятия, как квантор всеобщности и квантор существования. G8 — это аналог кванторной формулы, только для формирования термов. Терм $\{x : \varphi\}$ также связывает переменную x , как формулы $\forall x \varphi$ и $\exists x \varphi$. Как этот терм будет использоваться в самой теории, мы увидим ниже. Примечательно здесь то, что кванторными правилами мы очень тесно переплетаем термы и формулы: можно строить формулы при помощи термов, но можно строить и термы с помощью формул. Например, можно соорудить следующую конструкцию-терм:

$$\{\{\}, a, \{a, b\}, \{x : (x \in a) \wedge \neg(x = b) \rightarrow (b \in \{y : \neg\varphi(y) \wedge \psi\})\}\}$$

На этом можно не останавливаться и вырастить из данного терма «куст» скобочных записей, повторяя все процедуры, генерирующие объектные записи, только подставляя всюду данный терм вместо $\{\}$, а можно еще и вместо каждой свободной переменной (a и b) подставить не менее жуткие по своему виду термы, а вместо формульных переменных φ, ψ подставить какие-то формулы со свободной переменной y .

И все это великое разнообразие укладывается в правила G1–G17!

Итак, мы видим, что задача построения простейшей теории начальных множеств постоянно сводится к одному фундаментальному процессу — **генерации сущностей**. «Новые» объектные записи задаются из «старых» путем навешивания скобок, «новые» термы создаются из «старых» либо тем же способом, либо подстановкой «старых» термов вместо переменных, «новые» формулы создаются из «старых» формул либо подстановкой термов вместо переменных, либо соединением «старых» формул при помощи логических

Точнее — о
их
обозначениях

Зоопарк,
если уж не
то пошло...

связок. Процесс генерации, собственно, и является процессом изучения объектов теории. При этом не нужно думать, что «новые» сущности действительно являются новыми, т. е. появились позже появления «старых». «Новыми» они являются лишь для того, кто их впервые записывает, на деле же все формулы и термы — это способ конструктивного понимания сущностей теории. Построить сущность — не значит ее создать, это значит — указать, как она соотносится с другими сущностями, как ее можно воспроизвести алгоритмически.

Несмотря на все разнообразие терминологии и смыслов, стоящих за терминами, фундаментальный процесс генерации является неким единым комбинаторным принципом, порождающим из одних сущностей другие при помощи соединений и подстановок. Поэтому к любой сущности нашей теории можно приложить ее «родовое» дерево, наглядно демонстрирующее процесс построения. Например, на рисунке 1.1 представлены деревья для трех грамматических конструкций: объектной записи, терма с переменными и формулы.

И не только нашей теории

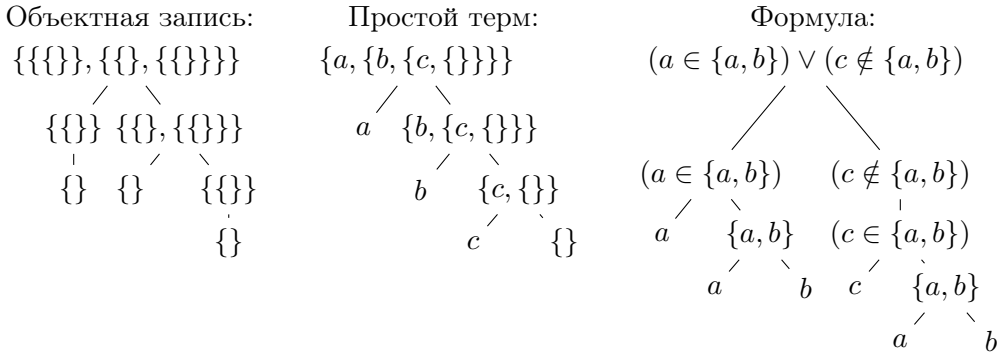


Рис. 1.1: Представление записи деревом.

Из приведенных примеров легко увидеть сформулированную выше теорему о согласованности грамматики. Достаточно представить, что мы вместо переменной a в простом терме подставили объектную запись (которая слева), как дерево терма вырастет и превратится в нечто большее (рисунок 1.2).

Дадим следующее определение: деревья, в которых узлами являются записи, соответствующие грамматике, а инцидентность (связь между узлами) устанавливается одним из правил грамматики G7, G8, G11, G16, G17, назовем **Гамма-деревьями** (гамма от слова «грамматика»).

Перечисленные в определении правила грамматики осуществляют переход от одной или многих записей к новой путем навешивания скобок или их соединения через логические символы. В случае если мы рассматриваем гамма-дерево объектной записи, то инцидентность устанавливается только при навешивании скобок на список термов или один терм. Дерево считается

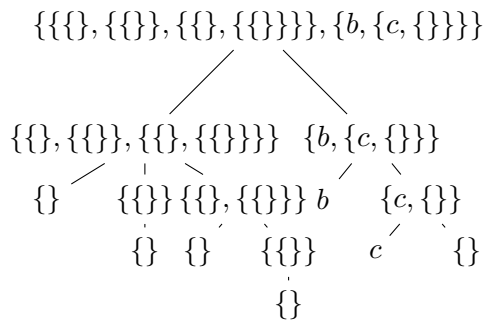


Рис. 1.2: В простой терм вместо a подставили объектную запись.

ориентированным от более длинных записей в сторону более коротких, так что в случае, например, гамма-дерева объектной записи прямыми потомками записи являются элементы списка, определяющего эту запись.

На рис.1.1 ориентация, очевидно, направлена сверху вниз. Листья гамма-дерева соответствуют элементарным термам (числам, объектным переменным, $\{\}$) и формульным переменным, а корень гамма-дерева — это полученная по правилам грамматики запись, соответствующая всему гамма-дереву.

Объектные записи и соответствующие им гамма-деревья можно сравнивать.

Для двух объектных записей a и b положим $a \prec b$, если запись a является некорневым узлом в гамма-дереве записи b . Можно показать, что $a \prec b$ тогда и только тогда, когда a входит (как подстрока) в запись b и не совпадает с b (т. е. имеет меньшее число символов). Пусть также $a \preceq b$, если $a \prec b$ или a и b посимвольно совпадают⁵. Для обозначения тождества a и b как строк будем использовать выражение $a \equiv b$.

Легко проверить также, что (1) $(a \preceq a)$, (2) $(a \preceq b) \wedge (b \preceq a) \rightarrow (a \equiv b)$, (3) $(a \preceq b) \wedge (b \preceq c) \rightarrow (a \preceq c)$, т. е. такое отношение на объектных записях является частичным упорядочением, которое мы подробнее рассмотрим в соответствующем разделе. Для любой объектной записи a либо $a \equiv \{\}$, либо $\{\} \prec a$.

Трактовка грамматических конструкций через гамма-деревья и связанное с ним отношение сравнения записей очень полезны при анализе грамматики и написании алгоритмов, поскольку они демонстрируют процесс «появления» записей путем рекурсивного применения правил грамматики. Так что если какой-либо записи соответствует гамма-дерево, то мы можем быть уверены в ее соответствии правилам грамматики.

⁵Мы пока опасаемся использовать понятие равенства, поскольку оно будет зависеть от вкладываемого в него смысла, к обсуждению равенства мы вернемся чуть позже.

Но как быть, если нам дали некоторую запись и попросили проверить — соответствует ли она грамматике? На самом деле, идея простая: нужно постараться построить гамма-дерево этой записи, применяя правила грамматики в обратную сторону. Если это сделать удалось, значит, запись грамматически верная. Эту задачу можно вполне решать на компьютере, пользуясь нехитрым алгоритмом. Рекурсивная процедура проверки терма без формул (т. е. без вхождений термов вида $\{x : \varphi\}$) на корректность схематически представлена на рисунке 1.3.

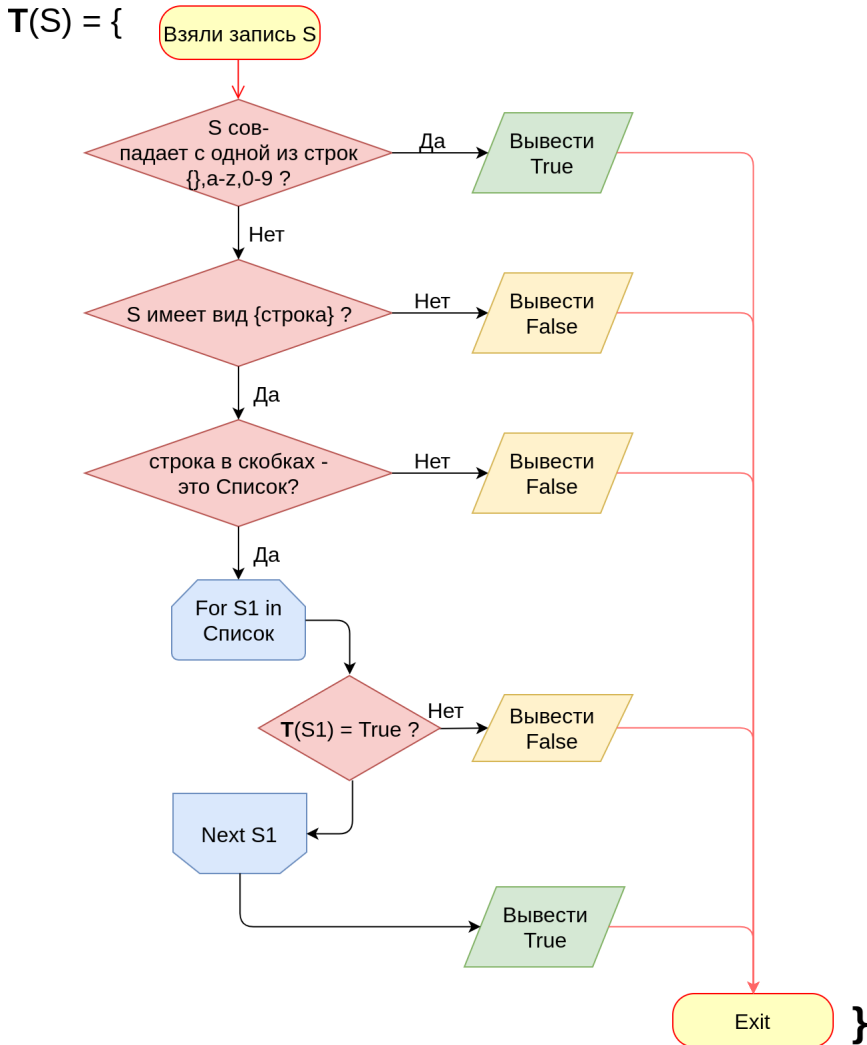


Рис. 1.3: Рекурсивная процедура проверки записи на соответствие грамматике.

Мы намеренно не стали здесь рассматривать все возможные конструкции грамматики, ограничившись лишь проверкой на то, что запись является термом без вхождений кванторных символов, чтобы алгоритм выглядел обзорным и понятным. Но на основе этого примера уже нетрудно понять,

Упражнение
1.1.
Проделайте
это само-
стоятельно.

как его усовершенствовать до получения полного алгоритма проверки соответствия записи грамматическим правилам G1–G17. Суть алгоритма заключается в том, что любой терм можно редуцировать до списка термов, входящего в него по правилу G7, а

затем применить эту же процедуру проверки ко всем элементам списка, и действовать так до тех пор, пока не дойдем до элементарных термов. Любое несоответствие на любом этапе проверки приводит к отрицательному ответу процедуры, а конечная длина исходной записи гарантирует конечное время работы алгоритма.

В общем-то, вся математика может быть сведена к такого рода процессам: построению термов и формул из первоначальных объектных записей и переменных/констант через их соединение различного рода операторами, и проверке произвольных записей на соответствие правилам грамматики. Здесь можно вспомнить и дерево вывода в формальной теории доказательств, и алгоритмы вычислительных машин, и все математические конструкции в любой теории любой сложности, а также все проверки корректности определений. Эта схема работает везде, т. к. является неотъемлемой частью нашего мышления.

Ценой
неимоверных
усилий

На первый взгляд может показаться, что возникает порочный круг, когда мы допускаем, что переменные могут указывать на термы, а термы состоять из переменных. На самом деле, тут нет противоречия. Если переменная a указывает на конкретный терм, то это означает, что всюду вместо вхождения a можно подставить данный терм. Если при этом оказывается, что терм содержит в себе переменную a , то такая связь терма и переменной является разновидностью уравнения, например, $a = \{a, b\}$. Но не стоит забывать, что дерево формулы (а уравнение — это формула) строится всегда без учета его содержательной части. Поэтому для формулы $a = \{a, b\}$ будет одно дерево, а для формулы $a = \{\{a, b\}, b\}$ — другое. Разные формулы — разные деревья. Несмотря на то, что второе уравнение является частным случаем первого. Такого рода «неудобные» конструкции могут возникать только в области термов и переменных, но не в области объектных записей — записей «начальных» множеств. Как будет видно из дальнейшего, в теории множеств есть даже специальная аксиома, запрещающая «зацикленность» множеств по отношению принадлежности — именно с той целью, чтобы убрать все возможные парадоксы, связанные с порочным кругом в конструкции. Иначе говоря, логическая зацикленность в области термов допускается (уравнения), а конструктивная зацикленность в области объектов — нет.

1.1.2 Рекурсия и индукция

Обсуждение правил грамматики и построение алгоритма проверки записи на соответствие этим правилам вывело нас на один фундаментальный процесс — **рекурсивное определение сущностей**. Данный процесс предполагает использование в определении одних и тех же понятий как с определяемой, так и с определяющей стороны. Достаточно посмотреть на правило G5 (Список есть Список, Терм), где понятие список определяется само через себя, позволяя строить цепочки термов неограниченной длины.

В дальнейшем мы увидим, что и в аксиоматической теории множеств объекты будут строиться по рекурсивным правилам, точнее, будут заданы правила, по которым из известных множеств можно построить новые.

И раз уж нам сама теория дает рекурсивно построенные объекты, то почему бы не пользоваться их устройством для построения новых рекурсивных определений?

Так, имея дерево объектной записи, мы можем определить некоторое понятие сначала для листьев (база рекурсии), а затем для всех вышестоящих узлов по правилу, в котором будут использоваться уже определенные для нижних узлов объекты (шаг рекурсии). В итоге мы получим построение нового понятия для всего дерева объектной записи.

Например, таким способом можно ввести понятие *высоты* гамма-дерева. Для листьев определим высоту равной нулю. Для любой ветви гамма-дерева определим ее высоту как максимум высот дочерних ветвей, увеличенный на 1. Нетрудно видеть, что тогда высота записи совпадет с максимальной длиной пути от вершины дерева записи к листьям. Так, высота гамма-дерева на рисунке 1.1 равна 3. В дальнейшем мы увидим и более сложные рекурсивные определения.

Рекурсию можно рассматривать не только как инструмент для построения новых сущностей в теории, но и как инструмент проверки истинности формул. Такая рекурсивная проверка обычно называется индукцией. Например, всем хорошо известная математическая индукция по натуральным числам: если известно что формула $\varphi(n)$ истинна при $n = 0$, и всякий раз, когда истинна $\varphi(n)$, истинна также и $\varphi(n + 1)$, то для всех n формула $\varphi(n)$ истинна. Мы еще вернемся к этому принципу, когда введем в нашу теорию числа и арифметические операции. Обобщением принципа математической индукции является трансфинитная индукция, для формулировки которой нам потребуются бесконечные числа — ординалы.

Вообще, всегда, когда у нас имеется объект, построенный рекурсивно (например, запись грамматики и ее гамма-дерево), можно аналогичным образом построить и индуктивное рассуждение, например, для проверки правильности построения такого объекта или доказательства его свойств.

Метатеорема 1.2 (Индукция по гамма-дереву объектной записи). Пусть c — объектная запись и $\varphi(x)$ — некоторая формула. Пусть $\varphi(\{\})$ истинно и для любого $b \preceq c$ если $\varphi(a)$ истинно для всех $a \prec b$, то $\varphi(b)$ истинно. Тогда $\varphi(c)$ истинно.

Доказательство. Предположим, что $\varphi(c)$ ложно. Тогда существует $a_1 \prec c$ такое, что $\varphi(a_1)$ ложно. Аналогично, существует $a_2 \prec a_1$ такое, что $\varphi(a_2)$ ложно. Продолжая строить таким способом путь в гамма-дереве от корня вниз $c \rightarrow a_1 \rightarrow a_2 \rightarrow \dots$, мы должны за конечное число шагов прийти либо до листа $\{\}$, либо до такого a_k , что для всех $a \prec a_k$ $\varphi(a)$ истинно. В первом случае имеем противоречие с базой индукции, во втором — с шагом индукции. \square

Во всех случаях принцип индукции устанавливается методом «от противного», а кроме того, используется тот факт, что на рассматриваемой структуре (числах или деревьях) можно найти минимальный элемент в произвольном ограниченном наборе объектов (вершин). Индукция по дереву в паре с рекурсией по дереву позволяют устанавливать факт единственности рекурсивно построенного на дереве понятия, чем мы тоже будем активно пользоваться в дальнейшем.

Более строго мы рассмотрим рекурсию и индукцию после освоения необходимой теоретико-множественной техники, а пока примем к сведению этот полезный инструментальный только в отношении гамма-деревьев.

1.1.3 Первые архетипы

После всех манипуляций с грамматикой, построением термов и формул, а также проверки записи на соответствие грамматике легко обратить внимание на то, что формулы и термы — очень похожие по своему поведению сущности. Те и другие строятся по рекурсивным правилам грамматики, отправляясь от переменных, чисел и скобок, и могут рассматриваться как многоместные функции от объектных переменных. Разница лишь в том, что значениями термов-функций являются объектные записи, а значениями формул-функций являются «истина» и «ложь». Такое сходство позволяет, например, изучать высказывания (формулы) как объекты формальной теории, в которой формулы сами по сути становятся термами от булевских переменных, т. е. функциями.

Более того, все функции, как мы увидим в дальнейшем, можно представить как множества специального вида, что позволяет нам все формулы погрузить во множества и, тем самым, свести язык только лишь к оперированию термами, обозначающими множества. Поэтому базовым **архетипом** математики является **множество** и все возможные его обобщения — **классы**, **объекты**, **мультимножества** и т. п.

Собственно, на этом можно было бы и закончить книгу, но проблема заключается в том, что чисто психологически нам удобнее отделять объекты и операции над ними (их преобразования) — как логические, так и синтаксические, — наполняя эти операции самыми разными прикладными смыслами.

И здесь мы неявно подходим к первому и, пожалуй, главному (не путать с базовым!) архетипу математики — **функции** (сюда же можно отнести **высказывания, предикаты, отображения, функционалы, операции, операторы, функторы и т. д. и т. п.**). Функция чаще всего воспринимается как некий процесс преобразования одних объектов в другие, например, одних множеств в другие, вследствие чего при обозначении функции принято пользоваться символом «стрелка». На таком особом статусе и обозначении функции основывается понятийный аппарат *теории категорий*, где мы оперируем только объектами и стрелками (морфизмами) между ними.

Наконец, еще один фундаментальный архетип математики — это **рекурсия** (сюда же можно отнести *правила грамматики, алгоритмы, конечные автоматы, вычислимые функции и т. п.*). Как мы только что видели, начиная с грамматики рекурсия (и основанная на ней **индукция**) плотно вживается в жизнь математика и преследует его во всех сферах математической деятельности. При этом рекурсия чаще всего связана еще и с функцией, поскольку шаг рекурсии определяется некоторой функцией, вычисляющей значение для объекта теории через уже вычисленные значения для предыдущих (дочерних) объектов теории.

Архетипич-
ный
процесс

Таким образом, если на данном этапе нашей теории охватить взглядом целевые объекты книги, то мы заметим некоторое обширное поле, состоящее из множеств, или, точнее, из их записей, построенных по рекурсивным правилам грамматики, над этим полем работают во всех направлениях функции и формулы, а с их помощью вдоль всего поля прокладываются ветвистые дороги рекурсивных процедур и алгоритмов, защищенные индукцией.

Наша задача теперь — изучить эту картину детально.

1.1.4 Финальные дополнения к языку

Часто вместо термина $\{x : \varphi\}$ (правило G8) используется терм $\{x | \varphi\}$. Мы не вводили его в формальной грамматике, чтобы избежать путаницы со смыслом символа $|$, однако в дальнейшем условимся по определению считать оба этих термина тождественными по определению.

Помимо записей из формальной грамматики мы будем использовать некоторые дополнительные обозначения, вводя их по определению через уже известные в грамматике записи. Например, полезными символами для формул являются \subseteq , \subset , а также симметричные им символы \supseteq , \supset и их отрицания, для термов полезными символами являются \cap , \cup и \setminus . Для того, чтобы вводить определения, мы будем использовать символ \Leftrightarrow , помещая слева определяемое

выражение, а справа — определяющее.

Итак, для формул положим:

$$\begin{array}{ll}
a \ni b \quad \Rightarrow \quad b \in a & a \notin b \quad \Rightarrow \quad \neg(a \in b) \\
a \neq b \quad \Rightarrow \quad \neg(a = b) & \\
a \subseteq b \quad \Rightarrow \quad \forall x(x \in a \rightarrow x \in b) & a \subset b \quad \Rightarrow \quad (a \subseteq b) \wedge (a \neq b) \\
a \supseteq b \quad \Rightarrow \quad b \subseteq a & a \supset b \quad \Rightarrow \quad b \subset a \\
a \not\subseteq b \quad \Rightarrow \quad \neg(a \subseteq b) & a \not\subset b \quad \Rightarrow \quad \neg(a \subset b) \\
a \not\supseteq b \quad \Rightarrow \quad \neg(a \supseteq b) & a \not\supset b \quad \Rightarrow \quad \neg(a \supset b) \\
\forall a \in b : \varphi \quad \Rightarrow \quad \forall a (a \in b) \rightarrow \varphi & \exists a \in b : \varphi \quad \Rightarrow \quad \exists a (a \in b) \wedge \varphi \\
\exists a = b \quad \Rightarrow \quad \exists a a = b & \exists! a \varphi(a) \quad \Rightarrow \quad (\exists a \varphi(a)) \wedge \\
& \quad \forall a \forall b (\varphi(a) \wedge \varphi(b) \rightarrow a = b)
\end{array}$$

Для термов положим:

$$\begin{array}{ll}
a \cap b \quad \Rightarrow \quad \{x \mid (x \in a) \wedge (x \in b)\} & \text{(пересечение)} \\
a \cup b \quad \Rightarrow \quad \{x \mid (x \in a) \vee (x \in b)\} & \text{(объединение)} \\
a \setminus b \quad \Rightarrow \quad \{x \mid (x \in a) \wedge (x \notin b)\} & \text{(разность)} \\
a \triangle b \quad \Rightarrow \quad (a \setminus b) \cup (b \setminus a) & \text{(симметрическая разность)} \\
\{a\} \quad \Rightarrow \quad \{x \mid x = a\} & \text{(синглетон)} \\
\{a, b\} \quad \Rightarrow \quad \{x \mid (x = a) \vee (x = b)\} & \text{(пара)} \\
\{a, b, \dots, c\} \quad \Rightarrow \quad \{x \mid (x = a) \vee (x = b) \vee \dots \vee (x = c)\} & \text{(конечный набор)}
\end{array}$$

здесь многоточие слева соответствует любому Списку термов (в смысле определений грамматики), а многоточие справа — соответствующий перечень атомарных формул равенства, связанных дизъюнкцией.

В данных определениях переменные a, b могут быть заменены любыми термами грамматики. Отметим, что введенные по определению новые обозначения всюду в дальнейших выводах и построениях можно без потерь заменять их определениями, сводя всю теорию исключительно к построенной грамматике. Проблема будет лишь одна — текст невозможно будет читать человеку.

*В отличие
от того
самого ком-
пьютера?*

Мы также присоединяем к нашей теории **исчисление высказываний**, позволяющее строить выводы над формулами грамматики, заменяя формулы эквивалентными и т.д., т.е. будем полностью использовать аппарат математической логики. Поэтому наша грамматика теории множеств уже не является «подвешенной в воздухе» абстрактной конструкцией, она опирается на хорошо проработанный аппарат формальной теории, которая к тому же обладает сильным свойством — полнотой. В связи с этим,

даже не имея еще никаких аксиом, можно получить первое простейшее утверждение теории множеств: *для любого a верно $a \subseteq a$* , т. к. утверждение $(x \in a \rightarrow x \in a)$ истинно независимо от своих аргументов.

Часто в утверждениях, далеких от теории множеств, для обозначения термов мы будем использовать символы не из грамматики (например, $\mathbb{R}, \mathbb{Q}, \mathbb{C}$, прописные латинские буквы, различные акценты и индексы, и т.п.), их определение мы также будем вводить с помощью символа \Rightarrow и терма, соответствующего нашей грамматике. Иначе говоря, *расширение языка теории мы допускаем только в виде переобозначений записей, соответствующих грамматике*. После введения такого определения мы считаем, что левая и правая части определения а) для термов — равны (в смысле равенства, определенного в теории), б) для формул — эквивалентны в смысле исчисления высказываний, и это правило действует до тех пор, пока определяемый символ не будет (пере)определен снова.

В данном случае это хорошее свойство

1.1.5 Равенство, принадлежность и объекты теории

В грамматике языка теории множеств мы уже ввели формулы вида $(a = b)$, но грамматика не дает никакой информации о свойствах равенства.

Мы можем пойти простым путем и считать все объектные записи различными множествами. То есть различать, например, $\{\{\}\}$ и $\{\{\}, \{\}\}$, а также $\{\{\}, \{\{\}\}\}$ и $\{\{\{\}\}, \{\}\}$. В первом случае записи отличаются количеством вхождения $\{\}$ в список, образующий терм, а во втором — порядком элементов списка, образующего терм.

К вопросу о мультимножествах

Вроде бы разумный подход, однако вспомним теперь об отношении принадлежности, которое также введено грамматикой, но пока не обладает никакими свойствами. Вопрос: *как его связать с отношением равенства?*

Обозначим $a \Leftarrow \{\{\}\}$, $b \Leftarrow \{\{\}, \{\}\}$, $c \Leftarrow \{\{\}, \{\{\}\}\}$, $d \Leftarrow \{\{\{\}\}, \{\}\}$. Поскольку a, b, c, d — объектные переменные, то мы можем не только записать такие формулы, но и считать, что временно положили в ячейки памяти a, b, c, d соответствующие термы. Таким образом, мы предполагаем, что $a \neq b$ и $c \neq d$.

Строго говоря, запись $a \in b$ вообще ничего не значит без аксиом. Например, можно считать, что $a \in b$, если в скобочной записи a скобок меньше, чем в скобочной записи b , если a и b суть объектные записи. Но такое определение никак не вписывается в наше представление о множестве как о «коробке с предметами» или выделенном наборе точек на прямой. Поэтому нужно искать некое представление о «коробке» среди грамматических конструкций языка. И такое представление есть! Это построение терма в виде {Список}, где внешние фигурные скобки соответствуют стенкам коробки, а элементы списка — предметам в этой коробке.

Но в таком случае для наших термов a, b, c, d мы получаем следующие

утверждения:

$$\{\} \in a, \quad \{\} \in b, \quad \{\} \in c, \quad \{\} \in d, \quad \{\{\}\} \in c, \quad \{\{\}\} \in d$$

Отношение принадлежности в данном случае никак не отражает ни кратность вхождения термов, ни их порядковый номер вхождения. Для этой информации в отношении \in просто не предусмотрено никакого параметра (например, индекса кратности).

Но тогда мы получаем, что множества a и b состоят из одних и тех же элементов, и множества c и d состоят из одних и тех же элементов, но, как мы условились выше, $a \neq b$ и $c \neq d$. То есть не равны две коробки, состоящие из одних и тех же предметов. Стало быть, введенное нами отношение равенства слишком «низкоуровневое» — оно работает на уровне самих записей и различает предметы там, где отношение принадлежности их различать не умеет. Такое рассогласование «лечится» прямым определением равенства через принадлежность, а именно:

$$a = b \leftrightarrow \forall x(x \in a \leftrightarrow x \in b). \quad (1.1)$$

Здесь мы полагаем множества равными в том случае, если они состоят из одних и тех же элементов. Данный постулат принято называть **аксиомой объемности**.

Теперь у нас будут выполняться равенства $a = b$ и $c = d$. То есть *аксиома объемности отождествляет те термы, в которых списки отличаются только порядком и/или кратностью элементов, либо не отличаются вовсе (как строки)*.

Рассмотрим теперь четыре новых терма:

$$\{\{\{\}\}\}, \quad \{\{\{\}, \{\}\}\}, \quad \{\{\{\}, \{\{\}\}\}\}, \quad \{\{\{\{\}\}, \{\}\}\},$$

которые, как легко видеть, можно переписать в виде $\{a\}$, $\{b\}$, $\{c\}$, $\{d\}$. Очевидно, что $a \in \{a\}$, $b \in \{b\}$, $c \in \{c\}$, $d \in \{d\}$. Но верно ли, что $a \in \{b\}$? Ведь мы все еще считаем, что $a \in \{b\}$, если запись множества a встречается в списке, формирующем терм $\{b\}$. Но список терма $\{b\}$ состоит из единственного элемента — b , а его запись не является записью a . Стало быть, неверно, что $a \in \{b\}$ и неверно, что $b \in \{a\}$. То же самое относится к паре c и d , ведь их записи отличаются.

Это обстоятельство говорит нам о том, что и отношение принадлежности мы не должны настолько «плотно» привязывать к грамматическим записям. Его нужно определить более общо. И это нам позволяет сделать **аксиома экстенциональности**⁶, которая обычно сопровождает отношение равенства

⁶**Экстенциональность** (от лат. *extensio* — протяжение, расширение) — заменяемость равных (синонимичных) выражений применительно к контекстам и языкам.

во всех формальных теориях. Данная аксиома утверждает, что какова бы ни была грамматически верно построенная формула $\varphi(x)$ со свободной переменной x , если $a = b$, то формулы $\varphi(a)$ и $\varphi(b)$ эквивалентны. Иначе говоря, на равных термах формула принимает одно и то же логическое значение.

В частности, если в качестве формулы φ взять атомарную формулу (отношение принадлежности), то справедливо следующее:

$$a = b \rightarrow \forall x(a \in x) \leftrightarrow (b \in x) \quad (1.2)$$

Таким образом, если мы уже установили факт, что $a = b$ и $c = d$ (например, на основе аксиомы объемности), то по аксиоме экстенциональности мы автоматически получаем, что $a \in \{b\}$ и $b \in \{a\}$, а также что $c \in \{d\}$ и $d \in \{c\}$. Перепишем полученные выводы в исходных символах для объектных записей:

$$\begin{aligned} \{\{\}\} &= \{\{\}, \{\}\}, & \{\{\}, \{\{\}\}\} &= \{\{\{\}\}, \{\}\}, \\ \{\{\}\} &\in \{\{\{\}, \{\}\}\}, & \{\{\}, \{\}\} &\in \{\{\{\}\}\}, \\ \{\{\}, \{\{\}\}\} &\in \{\{\{\{\}\}, \{\}\}\}, & \{\{\{\}\}, \{\}\} &\in \{\{\{\}, \{\{\}\}\}\} \end{aligned}$$

Из приведенных примеров видно, что введенное нами **равенство**, во-первых, **отождествляет термы, списки которых отличаются только порядком и кратностью вхождения элементов-записей**, а во-вторых, **при проверке принадлежности одного терма другому достаточно, чтобы любой равный первому терм входил в список второго терма**.

Говоря о принадлежности, необходимо учесть еще один вариант терма: $\{x \mid \varphi\}$. Дело в том, что он использует в своем определении формулу, а значит, его использование в атомарной формуле необходимо согласовать. Пусть x свободно входит в φ , тогда мы полагаем по определению, что

$$(a \in \{x \mid \varphi(x)\}) \leftrightarrow \varphi(a). \quad (1.3)$$

Если же x не входит в φ свободно, то $(x \in \{x \mid \varphi\}) \leftrightarrow \varphi$.

Таким образом, терм $\{x \mid \varphi\}$ перечисляет в своих элементах область истинности формулы φ относительно переменной x . Например, терм $\{a, b\}$ равен терму $\{x \mid \varphi(x)\}$, где $\varphi(x)$ эквивалентно $(x = a) \vee (x = b)$.

Без определения данного терма-квантора в виде (1.3) невозможно никакими средствами теории установить истинность или ложность формулы $a \in \{x \mid \varphi(x)\}$, а также установить его равенство с каким-либо другим термом, даже если мы имеем дело только с «начальными» множествами. Поэтому в строгих формальных теориях (языках первого порядка) вообще не используются термы-кванторы.

Еще одно свойство терма-квантора, обосновывающее его «кванторность», заключается в том, что в нем можно заменять связанную переменную на любую другую, если это не приведет к коллизии переменных внутри формулы,

определяющей терм-квантор:

$$\{x \mid \varphi(x)\} = \{y \mid \varphi(y)\}.$$

Это прямо вытекает из (1.3) и того, что в этой формуле переменную a можно заменять на любую другую.

Итак, мы теперь полностью согласовали грамматику термов с логикой отношений принадлежности и равенства. Зафиксируем эти согласования следующими **Правилами согласования**:

C1 если $a = b$ и b является элементом списка термов в терме c (то есть $c \equiv \{\dots, b, \dots\}$), то $a \in c$;

C2 для всех c , для которых верно $c \in a$, верно также и $c \in b$, и обратно, для всех $d \in b$ верно $d \in a$ (в смысле согласования C1), тогда и только тогда, когда $a = b$;

C3 $a \in \{x \mid \varphi(x)\}$ тогда и только тогда, когда $\varphi(a)$.

В указанных правилах все переменные можно заменять на любые другие объектные переменные и термы.

Теперь мы можем корректно говорить об **объектах теории**. Для нас объектами теории, т. е. «начальными» множествами, будут такие сущности, которые записываются равными объектными записями. Иначе говоря, если две записи равны, то это записи одного и того же объекта-множества. Если бы мы уже владели развитым аппаратом теории множеств, то мы сказали бы, что объектами теории являются классы эквивалентности объектных записей, где под эквивалентностью записей понимается их равенство в определенном выше смысле.

Данное обстоятельство позволяет нам выбросить из рассмотрения все объектные записи множеств, в которых дублируются элементы списка. Так, нам не нужна запись $\{\{\}, \{\}\}$, обозначающая то же самое множество, что и более короткая запись $\{\{\}\}$.

Заметим, что мы фактически определили равенство и принадлежность, а не вводили их аксиоматическим путем. Это связано с тем, что пока мы рассматриваем вполне конкретную модель теории множеств, а не саму теорию как объект исследования. Тем не менее, уже сейчас мы можем зафиксировать еще один очень важный, можно сказать, системообразующий архетип математики: **архетип равенства**. Чуть позже мы вернемся к точным логическим формулировкам аксиом, связывающих равенство и принадлежность, а пока более детально изучим объектные записи нашей теории «начальных» множеств.

1.1.6 Скобочная запись множеств

Итак, мы видим, что из пустого множества путем применения нескольких правил грамматики можно построить очень богатый набор множеств. Точнее, богатый набор объектных записей. Мы научились отождествлять и различать эти записи путем введения согласований атомарных формул и записей. Тем самым, *мы отделили понятие объекта теории от понятия грамматической записи такого объекта и выяснили, что объект может быть записан по-разному.*

При этом, если забыть о переменных, то все «начальные» множества можно записать комбинациями только трех символов: фигурных скобок и запятых. Именно из этих символов строятся объектные записи. Понятно, что не любой набор этих символов будет обозначать множество, построенное в соответствии с этими правилами. Мы даже можем немного модифицировать алгоритм **T**, проверяющий произвольную запись на соответствие грамматике, так, чтобы он проверял произвольную запись на ее соответствие объектным записям. Однако более математический (и более красивый) подход заключается в том, что мы можем указать правила, которым должны удовлетворять все скобочные записи множеств, и только они. Причем, эти правила никак не связаны с рекурсивным построением записей, чем в корне отличаются от алгоритмической верификации.

Эти правила таковы.

- (b1) скобочная запись начинается с открытой скобки, заканчивается закрытой;
- (b2) запятая может быть вхождением только в последовательность $\},\{$;
- (b3) последовательность $\}\{$ запрещена;
- (b4) открытых и закрытых скобок одинаковое количество;
- (b5) слева от любой (не первой) скобки открытых скобок больше, чем закрытых.

Платой за отказ от явного использования рекурсии здесь является неявное использование понятия числа (определение которого рекурсивно).

Действительно, правила (b4) и (b5) предписывают нам сравнить количество открытых и закрытых скобок (в том числе, на отрезке записи — слева от произвольно выбранного в записи символа). На самом деле, сравнение количества здесь не предполагает прямого подсчета скобок — мы можем просто зачеркивать в скобочной записи скобки парами (открытую и закрытую), пока не останется ни одной скобки определенного вида: если во всей записи таким

*Упражнение 1.2.
Как нужно модифицировать алгоритм?*

А вот зачеркивание возвращает нас к явному использованию рекурсии

образом не останется ни одной скобки, то выполнено (b4), а если на отрезке записи останутся только открытые скобки, то выполнено (b5).

Теорема 1.1. *Скобочная запись является объектной записью тогда и только тогда, когда она удовлетворяет правилам (b1)–(b5).*

Это утверждение можно проверить непосредственно, пользуясь правилами грамматики G4–G7, либо используя модифицированный под объектные записи алгоритм **T**. Но мы, немного забегая вперед, все-таки воспользуемся помощью арифметики и построим доказательство теоремы с привлечением чисел. К сожалению, книга линейна, и мы не можем представить ее в виде

Напоминает
обход дерева
в прямом
порядке
(сверху
вниз) ;-)

дерева, как множество, поэтому то и дело нам придется забегать вперед и возвращаться назад.

Рассмотрим следующую конструкцию. Пусть имеется некоторая объектная запись, в которой использовано n скобок и какое-то количество запятых. Перемещаясь слева направо по всем выписанным скобкам мы будем подсчитывать «баланс» скобок слева от курсора — количество открытых скобок минус количество закрытых. Так, на первой скобке этот баланс будет равен 1, т. к. запись множества начинается открытой скобкой, а на последней — нулю, т. к. в скобочной записи множеств одинаковое количество открытых и закрытых скобок.

В итоге мы получим некоторый набор из n чисел $(1, k, m, \dots, 0)$, который назовем *характеристическим набором* для исходной скобочной записи и соответствующего ей множества. Например, для пустого множества $\{\}$ этот набор равен $(1, 0)$. Для множества $\{\{\}\}$ он равен $(1, 2, 1, 0)$, для множества $\{\{\}, \{\{\}\}\}$ — $(1, 2, 1, 2, 3, 2, 1, 0)$, а для множества $\{\{\}\} \cup \{\{\{\}\}\}$ он равен $(1, 2, 1, 0) \cup (1, 2, 3, 2, 1, 0) = (1, 2, 1, 2, 3, 2, 1, 0)$.

Нетрудно проверить, что для любой объектной записи характеристический набор обладает следующими свойствами:

- (a1) начинается с 1, заканчивается 0;
- (a2) соседние числа отличаются ровно на 1;
- (a3) все числа, кроме последнего, больше 0.

Более того, любой набор чисел, удовлетворяющий требованиям (a1)–(a3), порождает корректную скобочную запись, т. е. такую запись, которая будет скобочной записью множества. Для этого достаточно записывать скобку $\{$ всякий раз, когда число в наборе увеличивается, и $\}$ — когда уменьшается, а затем все вхождения $\{\}$ заменить на $\},\{$

Докажем это. Пусть имеется произвольный набор f , удовлетворяющий свойствам (a1)–(a3). Произведем его редукцию $\text{red}(f)$ следующим образом: отбросим начальную единицу и завершающий ноль, затем все числа в наборе уменьшим на 1. Набор $\text{red}(f)$ будет удовлетворять свойствам (a1) и (a2).

Далее возможны два случая: (а) набор $\text{red}(f)$ не содержит внутренних нулей и (б) — содержит внутренние нули.

В случае (а) набор $\text{red}(f)$ удовлетворяет свойствам (а1)–(а3), и к нему можно снова применить редукцию.

В случае (б) справа от каждого внутреннего нуля стоит 1 (по свойству (а2) исходного набора f). Тогда представим набор $\text{red}(f)$ как склейку наборов $f_1 \dots f_k$, начинающихся с 1 и заканчивающихся 0, и не содержащих внутренних нулей. Наборы f_1, \dots, f_k удовлетворяют свойствам (а1)–(а3), и к ним тоже можно применить редукцию.

Таким образом, путем применения редукции мы в конечном итоге дойдем до таких наборов, которые удовлетворяют свойствам (а1)–(а3), но к которым редукция неприменима, а такими наборами могут быть только наборы $(1, 0)$.

Поскольку набор $(1, 0)$ однозначно определяет скобочную запись $\{\}$, исходный набор f был приведен путем редукции к набору корректных скобочных записей. Теперь по индукции нетрудно показать, что исходный набор также соответствует корректной скобочной записи.

Действительно, переход от $\text{red}(f)$ к f означает навешивание скобок, и если запись, соответствующая $\text{red}(f)$, была корректной, то такой же будет и f . Переход от $f_1 \dots f_k$ к f означает построение записи $\{s_1, \dots, s_k\}$, где s_1, \dots, s_k — скобочные записи, соответствующие наборам $f_1 \dots f_k$, и если все записи s_1, \dots, s_k были корректными, то такой же будет и f . Но процесс, обратный редуцированию, начинается, как мы показали, с записи пустого множества $\{\}$. Следовательно, запись, соответствующая исходному набору f , будет построена корректно.

Итак, любой набор f , обладающий свойствами (а1)–(а3), порождает корректную скобочную запись множества, и, наоборот, любая скобочная запись множества порождает набор f , удовлетворяющий свойствам (а1)–(а3). При этом нужно помнить, что одному множеству может соответствовать бесконечно много корректных записей, поскольку множества не зависят от порядка и дублирования элементов, а записи — зависят.

Свойства (а1)–(а3) эквивалентны свойствам (b1)–(b5), следовательно, теорема 1.1 верна. Больше того, если мы «выбросим» из объектных записей запятые (их несложно однозначно восстановить по комбинации скобок $\{\}$), то правила (b1)–(b5) можно заменить на правила (b1), (b4), (b5). То есть корректность объектных записей верифицируется достаточно просто и доступно для интуиции.

С точки зрения арифметики здесь будет интересен следующий вопрос: если нам дано $n + 1$ открытых скобок и $n + 1$ закрытых, то сколько корректных записей множеств можно из них построить? Ответом будут т.н. **числа Каталана** $C_n = \binom{2n}{n} \frac{1}{n+1}$ (подробнее см. [2]). Эти числа встречаются во многих

Упражнение 1.3.
Докажите это!

Упражнение 1.4.
Проверить эквивалентность.

задачах, природа которых отвечает рекуррентному соотношению:

$$C_n = C_0 C_{n-1} + C_1 C_{n-2} + \dots + C_{n-1} C_0,$$

где $n > 0$ и $C_0 = 1$. Как видим, уже на этапе «проектирования» множеств мы сталкиваемся с арифметическими объектами, имеющими собственное название. Обычно это свидетельствует о глубине и красоте изучаемой структуры

*А еще это
говорит о
сложности
предстоя-
щих
упражнений.*

и, говоря языком данной книги, ее «архетипичности».

Однако мы все еще находимся в затруднительном положении, поскольку мы не предъявили никакого эффективного (т. е. дающего детерминированный ответ за конечное число шагов) алгоритма, позволяющего для произвольных двух объектных записей вычислить логическое значение атомарных формул принадлежности и равенства.

Основным препятствием для такого алгоритма является то, что проверка равенства определяется через проверку принадлежности, а проверка принадлежности — через проверку равенства. Так, если мы хотим вычислить формулу $a \in b$, где a и b — некоторые объектные записи, то сначала мы должны представить b в виде {Список}, после чего проверить, не является ли a элементом списка, и если нет, то проверить, не выполняется ли равенство $a = c$ для некоторого c , являющегося элементом этого списка. Только после получения ответа на данный вопрос мы сможем предъявить ответ и на вопрос об истинности формулы $a \in b$.

Аналогично проверяется формула $a = b$. Сначала нужно представить a в виде {Список1}, b — в виде {Список2}, затем осуществить перекрестную проверку вхождения элементов списков друг в друга в смысле отношения принадлежности. И только если списки совпадут (в смысле равенства элементов), то можно будет дать ответ $a = b$.

Таким образом, нам требуется сразу два параллельных рекурсивно работающих алгоритма, один из которых проверяет отношение принадлежности, другой — равенства.

Представленные на схемах 1.4 и 1.5 алгоритмы **IN** и **EQ**, проверяющие, соответственно, принадлежность и равенство двух записей, работают совместно. Читателю предлагается самостоятельно проверить, что

1.5.

данные алгоритмы заканчивают работу за конечное число шагов.

Для этого нужно обратить внимание на то, что каждый раз, когда алгоритм переходит к элементам Списков, он спускается по дереву множества на 1 уровень. А это значит, что рано или поздно он достигнет листьев этого дерева, т. е. записей пустого множества.

В предъявленных алгоритмах проверки атомарных формул теории множеств основной проблемой является скорость их работы. Действительно, если множества и их элементы записаны различными способами (отличаются кратностью и перестановкой элементов), то нам приходится «смотреть» дерево этих множеств все глубже и глубже, пока не дойдем до пустых множеств,

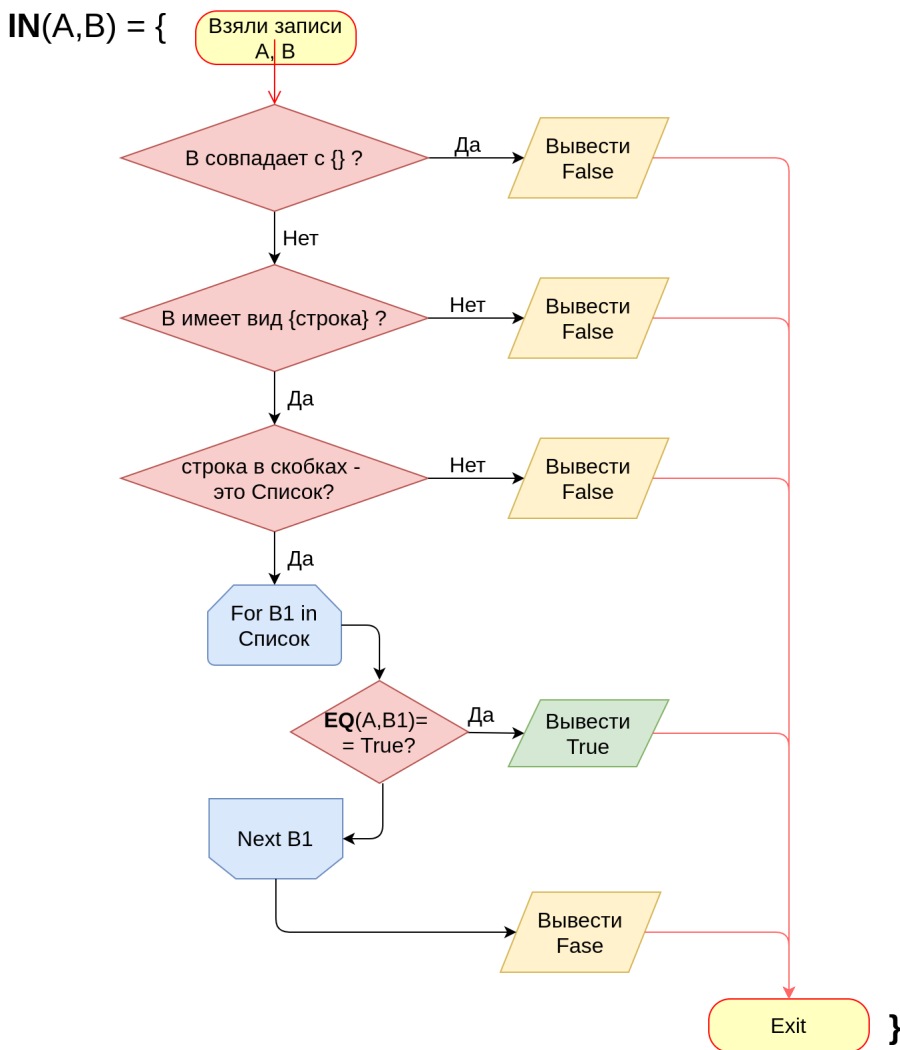


Рис. 1.4: Алгоритм проверки принадлежности.

запись которых уникальна (пустое множество в скобочном представлении не записать никак иначе, чем $\{\}$). И всякий раз мы вынуждены запускать перебор по спискам двух множеств, равенство которых нам нужно установить.

Программисты закидали бы нас тухлыми яйцами за такое расточительство памяти, времени и стека компьютера, ведь рекурсивное исполнение (в отличие от динамического программирования) для каждого узла гамма-дерева инициирует запуск подпрограммы, т. е. выделяет ей ссылку в стеке, замораживает исполнение вышестоящей программы, а главное — проход по всему

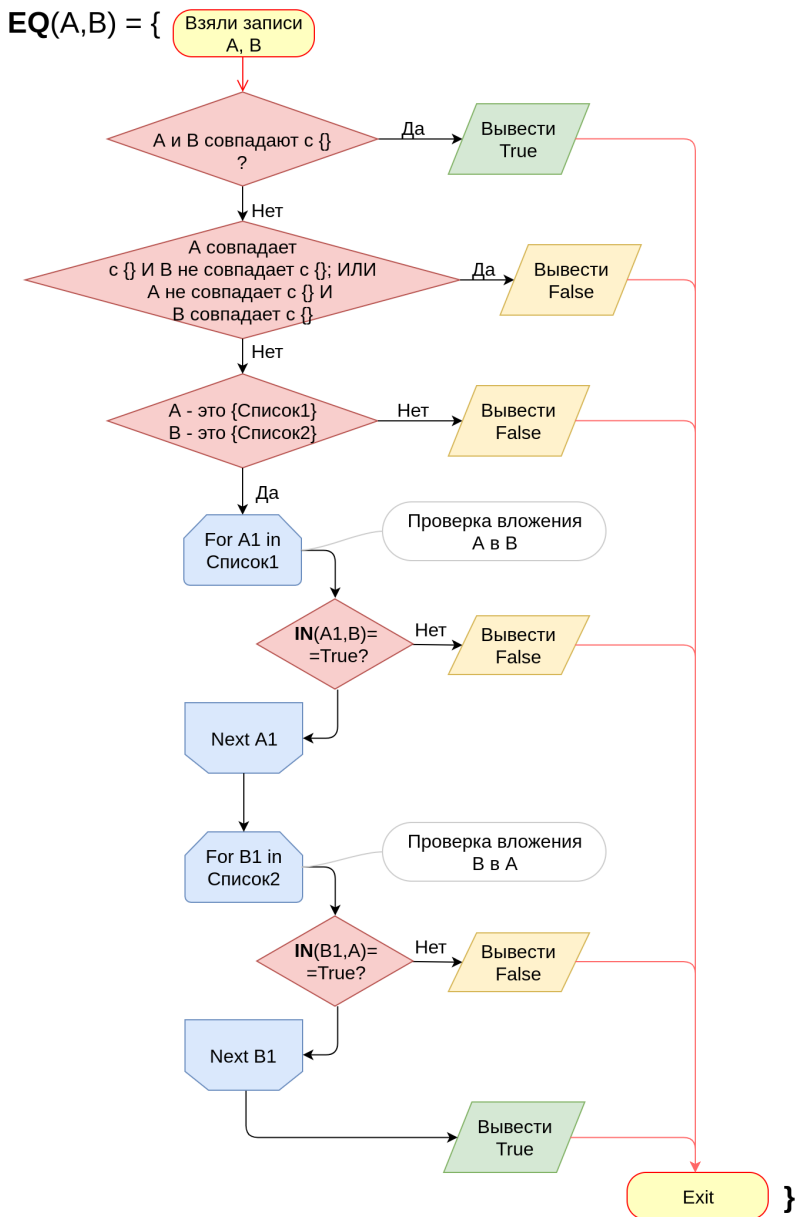


Рис. 1.5: Алгоритм проверки равенства.

дереву столь же неэкономичен, как вычисление чисел Фибоначчи рекурсией. Но наша задача сейчас состоит в том, чтобы наглядно показать возможность за конечное (пусть и большое) время разобрать произвольные строки и ответить на вопрос об их корректности, равенстве или принадлежности друг

другу, т. е. о разрешимости произвольной строки в нашей грамматике.

О том, как можно было бы их ускорить, не привлекая дополнительных знаний об обрабатываемых строках, мы предлагаем подумать читателю самостоятельно.

Заметим также, что алгоритмы можно было бы существенно ускорить, если бы мы смогли договориться о некоторой *канонической* («правильной») записи множеств, устраняя все кратности элементов и располагая элементы в определенном порядке. Тогда мы смогли бы сначала привести произвольную запись множества к канонической (что тоже является довольно затратной по времени задачей), а затем выполнить ее посимвольное сравнение с другой канонической записью для установки равенства множеств (например, алгоритмом Кнута–Морриса–Пратта). Совпадение канонических записей означало бы совпадение множеств, а несовпадение канонических записей — их различие.

И вот здесь у нас, пожалуй, впервые возникает неустраняемая потребность в **числах**. Дело в том, что числа — это некоторый инструмент упорядочения и сравнения. Мы можем еще ничего не знать об арифметике и операциях над числами, но, имея шкалу чисел, мы можем их сортировать и сравнивать. Следовательно, если бы нам удалось всем «начальным» множествам приписать некоторый уникальный числовой код, то мы смогли бы а) упорядочивать элементы множеств и б) легко проверять атомарные высказывания ($a \in b$) и ($a = b$), просто сверяя коды множеств a и b . Иначе говоря, мы бы просто свели работу к алгоритмам с числами, а это всегда приятнее, чем возиться с произвольными строками.

В связи с этим мы теперь воспользуемся заложенными в грамматике символами-константами для чисел (правила G1–G2).

1.1.7 Числа грамматики

Нетрудно заметить, что грамматика вводит в язык десятичное представление чисел, поскольку использует цифры от 0 до 9. Но до тех пор, пока записи чисел никак не связаны с объектами теории, мы не можем их использовать в полную силу (разве что обозначать ими избранные множества). Поэтому здесь мы дадим рекурсивное определение операции '+1' и затем сопоставим числа вполне конкретным «начальным» множествам.

Введем следующие обозначения:

Число0 + 1 \Rightarrow Число1	Число3 + 1 \Rightarrow Число4	Число6 + 1 \Rightarrow Число7
Число1 + 1 \Rightarrow Число2	Число4 + 1 \Rightarrow Число5	Число7 + 1 \Rightarrow Число8
Число2 + 1 \Rightarrow Число3	Число5 + 1 \Rightarrow Число6	Число8 + 1 \Rightarrow Число9
Число9 + 1 \Rightarrow (Число + 1) 0		

Здесь под нетерминальным символом Число следует понимать любой объект грамматики, соответствующий определению данного нетерминального символа. Проще говоря, Число — это произвольная конечная последовательность десятичных цифр. Кроме того, мы неявно предполагаем, что Число может быть заменено пустой строкой, чтобы иметь возможность применять операцию '+1' к цифрам. Это требование несущественное, поскольку мы вообще всегда можем оперировать числами, записанными двумя и более цифрами (просто в начале будут стоять незначимые нули). Последнее правило означает, что сначала следует к Числу прибавить 1, пользуясь этими правилами, а затем приписать справа 0.

Например, $123+1$ следует представить по правилу грамматики G2 как Число3+1, где Число — это 12. Тогда в силу приведенных выше соглашений Число3+1 обозначает Число4, т.е. 124. Аналогично, $0999+1$ следует представить как Число9+1, что приводит к Число+10, т.е. $099+10$, далее по тому же правилу получим $09+100$, наконец, получим $0+1000$, и, в итоге, 1000.

Таким образом, мы можем к любому числу добавлять 1, т.е. увеличивать его на 1. Осталось рекурсивно связать определение Числа со вполне конкретными множествами. Дадим следующее определение:

$$0 \Rightarrow \{\}, \quad a + 1 \Rightarrow a \cup \{a\},$$

где переменная a , в соответствии с грамматикой, является термом, т.е. обозначает произвольное множество.

Заметим, что a в данном определении не обязано быть числом, т.е. обозначать множество, которому приписано некоторое числовое обозначение, но в случае совпадения a с числом мы получаем следующую закономерность:

$$\begin{aligned} 0 &= \{\} \\ 1 &= 0 + 1 = \{\} \cup \{\{\}\} = \{0\} \\ 2 &= 1 + 1 = 1 \cup \{1\} = \{0, 1\} \\ 3 &= 2 + 1 = 2 \cup \{2\} = \{0, 1, 2\} \end{aligned}$$

В общем случае имеем: $n + 1 = \{0, 1, \dots, n\}$. Так в теории множеств принято определять натуральные числа (по фон Нейману). Для двух натуральных чисел n, m положим $n < m$, если путем конечного применения операции '+1' из n можно получить m ($m = n + 1 + 1 + \dots + 1$).

Отметим ряд свойств натуральных чисел:

- N1 натуральное число есть множество;
- N2 если n, m — натуральные числа, то $n \leq m$ тогда и только тогда, когда $n \subseteq m$;
- N3 для любых натуральных чисел n, m : $(n \leq m) \vee (m \leq n)$;

Упражнение
1.6.
Докажите
свойства.

N4 из $(n \leq m) \wedge (m \leq n)$ следует $n = m$.

Иначе говоря, любые два натуральных числа сравнимы по вложению множеств, а отношение $<$ для них эквивалентно отношению \subset (собственного вложения).

Обычно для обозначения натуральных чисел используются буквы i, j, k, l, m, n .

Понятно, что, имея операцию '+1', можно построить полноценные операции сложения, умножения и возведения в степень для натуральных чисел, а также всю теорию делимости. Забегая вперед, мы будем считать, что нам уже знакомы данные операции, и вновь вернемся к идее построения специального числового кода всех «начальных» множеств, который облегчил бы нам их сравнение.

| В очередной раз!

1.1.8 «Начальные» мультимножества

Когда мы вводили понятие объекта теории — «начального» множества, — мы решили согласовать атомарные высказывания «принадлежность» и «равенство» так, чтобы не различать множества с кратными и отличающимися сортировкой элементами. Таков классический подход к понятию множества. Однако в окружающем нас мире чаще всего мы сталкиваемся с тем, что одни и те же элементы повторяются многократно. Например, электроны в атоме, молекулы в газе. Даже на бытовом уровне мы склонны отождествлять между собой предметы, различие между которыми для нас несущественно или практически незаметно. Например, товары одной партии выпуска, имеющие одинаковый артикул.

С этой точки зрения довольно трудно себе представить множество без кратных элементов, т.е. множества вида $\{0,0,0,1,1,3,3,3,3\}$ и т.п. Например, это может быть запись разложения числа по степеням простых: $1800 = 2^3 3^2 5^2$ соответствует запись $\{2,2,2,3,3,5,5\}$.

Поэтому математики и, в особенности, программисты, вынуждены использовать понятие мультимножества, т.е. множества с кратными элементами. Обычно оно вводится искусственно на базе стандартных множеств и функций, но у нас есть уникальная возможность определить мультимножество сразу в грамматике.

Понятно, что определенные нами равенство и принадлежность не годятся для мультимножеств, поэтому нам потребуются дополнительные символы, обозначающие равенство и принадлежность мультимножеств. Введем их в нашу грамматику с помощью следующих правил:

G18 : Формула \rightarrow Терм $\in^{\text{Число}}$ Терм

G19 : Формула \rightarrow Терм \equiv Терм

Кроме того, положим $(a \in^0 b) \equiv \forall n (n > 0) \rightarrow \neg(a \in^n b)$, т. е. нулевой индекс принадлежности отвечает за отсутствие элемента a в мультимножестве b с какой угодно ненулевой кратностью.

Здесь атомарное отношение \in^n будет отвечать за указание кратности вхождения элемента в мультимножество, а атомарное отношение \equiv — за равенство мультимножеств. Равенство мультимножеств более «разборчивое», чем равенство множеств, т. к. оно различает кратные вхождения записей в Список, но менее «разборчивое», чем тождество строк, обозначенное ранее как \equiv , которое умеет различать еще и порядок элементов Списка.

Мультимножества записываются ровно такими же объектными записями, что и множества, поэтому для мультимножества точно так же определяется характеристический набор чисел и гамма-дерево. Вообще, следует особо подчеркнуть тот факт, что и множества, и мультимножества записываются совершенно одинаково с точки зрения синтаксиса, хотя семантически это разные объекты. Понять их отличие «на бумаге» можно только тогда, когда они участвуют в каких-либо формулах.

Тем не менее, есть ряд объективных отличий в свойствах множеств и мультимножеств.

Во-первых, требуется иначе согласовывать отношения принадлежности и равенства с грамматикой:

СМ1 : если $a \in^n b$ и $a \in^m b$, то $n = m$;

СМ2 : $a \in^n b$, если $b \equiv \{\text{Список}\}$ и в Списке есть ровно n элементов a_1, \dots, a_n , для которых верны равенства $a \equiv a_1 \equiv \dots \equiv a_n$;

СМ3 : $a \equiv b$, если $(c \in^n a) \rightarrow (c \in^n b)$ и $(d \in^m b) \rightarrow (d \in^m a)$ для любых c, d, n, m .

Заметим, что мы не стали вводить специальный кванторный терм для создания мультимножеств, полагая, что таким образом мы можем определять только обычные множества, даже если формула $\varphi(x)$ содержит атомарные формулы, связанные с мультимножествами. Например, $\{x | x \in^k a\}$ сформирует множество(!) всех таких x , которые входят в мультимножество a с кратностью k .

Во-вторых, в случае мультимножеств можно по-разному определять отношение вложения. Более правильным с точки зрения оперирования скобочными записями мы считаем такое [46]:

$$a \subseteq b \equiv \forall c \forall n > 0 \exists m \geq n : (c \in^n a) \rightarrow (c \in^m b),$$

то есть мультимножество b включает в себя те же элементы (и, возможно, другие), что и a , с той же или большей кратностью.

В-третьих, определим ряд операций над мультимножествами:

$$\begin{aligned} x \in^k (a \uplus b) &\leftrightarrow k = \max(n, m) \wedge (x \in^n a) \wedge (x \in^m b) \\ x \in^k (a \sqcap b) &\leftrightarrow k = \min(n, m) \wedge (x \in^n a) \wedge (x \in^m b) \\ x \in^k (a \oplus b) &\leftrightarrow k = n + m \wedge (x \in^n a) \wedge (x \in^m b) \\ x \in^k (a \ominus b) &\leftrightarrow k = \max(0, n - m) \wedge (x \in^n a) \wedge (x \in^m b) \end{aligned}$$

Заметим, что в этих формулах k — число, а значит, множество, поэтому к нему применяется обычная атомарная формула равенства со знаком $=$.

Терм $a \uplus b$ обозначает операцию объединения мультимножеств так, как будто все кратные вхождения их элементов являются разными элементами обычных множеств, например, $\{a, a, a, b, b\} \uplus \{a, a, b, b, b\} \doteq \{a, a, a, b, b, b\}$.

Аналогично $a \sqcap b$. Например, $\{a, a, a, b, b\} \sqcap \{a, a, b, b, b\} \doteq \{a, a, b, b\}$.

Новыми для нас являются операции сложения и вычитания. Поясним на примерах:

$$\begin{aligned} \{a, a, a, b, b\} \oplus \{a, a, b, b, b\} &\doteq \{a, a, a, a, a, b, b, b, b, b\}, \\ \{a, a, a, b, b\} \ominus \{a, a, b, b, b\} &\doteq \{a\}. \end{aligned}$$

Поскольку отрицательных кратностей не бывает, от элемента a просто ничего не осталось.

В-четвертых, самая сложная операция — это мультимножество всех подмультимножеств данного мультимножества.

$$x \in^1 \mathcal{P}(a) \leftrightarrow x \subseteq a.$$

Заметим, что мы здесь использовали символ \in^1 , поскольку рассматриваем $\mathcal{P}(a)$ и его элементы как мультимножества, с обычным символом \in пришлось бы считать $\mathcal{P}(a)$ обычным булеаном множества a . При этом стоит отметить, что в мультимножестве $\mathcal{P}(a)$ все элементы входят с единичной кратностью, в то время как элементы мультимножества $x \subseteq a$ могут иметь не только единичную кратность.

Булеан —
множество
всех подмно-
жеств.

Таким образом, описанная нами грамматика с дополнительными правилами G18 и G19 представляет собой язык записи как множеств, так и мультимножеств, а отличие двух моделей состоит лишь в том, какими атомарными формулами мы пользуемся. Больше того, мы можем по умолчанию принять за правило, что мы работаем с моделью мультимножеств, но как только мы встречаем атомарную формулу с символом \in или $=$, то ее истинность устанавливаем так, как будто мы находимся в модели «начальных» множеств.

Например, мы можем записать утверждение: $(a \subseteq b) \rightarrow (a \subseteq b)$, поскольку из вложения мультимножеств с учетом кратности следует вложение соответствующих множеств без учета кратности.

Такой чисто синтаксический прием позволяет уйти от сложных конструкций вроде классов эквивалентности и сопряжения отношений классов с отношениями их элементов (с чем мы непременно столкнемся в аксиоматической теории), поскольку тип объектов однозначно устанавливается применяемой к ним операцией (как в языках программирования высокого уровня), что избавляет от необходимости типизировать термы грамматики.

Для дальнейших применений мы можем расширить наш язык следующим обозначением: если a входит в Список записи $b \equiv \{\text{Список}\}$ с кратностью k , то вместо k вхождений a в Список будем записывать $k \bullet a$. Например, $\{a, a, a, b, b\} \equiv \{3 \bullet a, 2 \bullet b\}$. Кроме того, положим по определению, что $a \times k$ для числа k и мультимножества a означает мультимножество, в котором все кратности элементов увеличены в k раз. Например,

$$\{a, a, a, b, b\} \times 2 \equiv \{3 \bullet a, 2 \bullet b\} \times 2 \equiv \{6 \bullet a, 4 \bullet b\} \equiv \{a, a, a, a, a, a, b, b, b, b\}.$$

Для того, чтобы подчеркнуть различие двух операций умножения, умножение мультимножества на число мы записываем справа, причем используем другой символ для того, чтобы запись $k \times n$ понимать как умножение множества-числа на n , а запись $k \bullet n$ — как кратность элемента-числа n . Например,

$$\{2 \bullet \{a\}\} \times 2 \equiv \{4 \bullet \{a\}\} \not\equiv \{\{4 \bullet a\}\} \not\equiv \{\{a\}\} \times 4.$$

Итак, число перед мультимножеством означает его кратность, а число после мультимножества — умножение кратностей его элементов.

Упражнение 1.7. Количество элементов $\mathcal{P}(a)$ для всякого мультимножества $a \equiv \{k_1 \bullet x_1, \dots, k_s \bullet x_s\}$ вычисляется, как легко видеть, по формуле $(k_1 + 1) \dots (k_s + 1)$.

1.1.9 Универсальные множества и универсальный код

В данном разделе мы будем пользоваться грамматикой, включающей мультимножества, и все определения давать с учетом того, что элементы могут иметь кратность.

Для дальнейшего анализа нам потребуется понятие ранга (мульти)множества. Пока мы находимся в рамках модели «начальных» множеств это сделать очень просто: **рангом** (мульти)множества называется высота соответствующего ему гамма-дерева. Легко убедиться в том, что ранг (мульти)множества не зависит от того, для какой конкретно объектной записи построено гамма-дерево. Действительно, ни сортировка, ни дублирование ветвей этого дерева не меняют его высоту. Ранг (мульти)множества a обозначается $\text{rank}(a)$.

Из определения ранга «начального» множества нетрудно получить и его основные свойства:

R1 $\text{rank}(\{\}) = 0$;

R2 если $a \prec b$, то $\text{rank}(a) < \text{rank}(b)$;

R3 если $a \in^k b$, то $\text{rank}(b) = \text{rank}(a) + 1$;

R4 $\text{rank}(n) = n$ для числа n .

Если обратиться к введенному в разделе 1.1.6 понятию характеристического набора скобочной записи, то мы увидим также, что ранг множества на 1 меньше максимального числа в этом наборе.

Определим теперь универсальные (мульти)множества. Ранее мы допускали в грамматике только термы нефункционального вида, оставляя это свойство формулам. Пришло время усложнить себе жизнь и ввести в грамматику следующее правило:

G20 : Терм \longrightarrow Объектная переменная(Список) | Объектная переменная_{Список} |
Объектная переменная^{Список}

Это правило позволяет записывать конструкции вида $t(u, v)$, a^k , q_n , т. е. записывать привычные для математиков зависимости значения терма от переменных привычным способом. На самом деле, это примерно то же самое, что пара множеств $\{a, b\}$, только в других символах и со своим алгоритмом вычисления значения. При этом у нас появляется функциональный символ (переменная t в примере), который служит для краткого обозначения алгоритма вычисления (или определения) терма при замене переменных другими термами, он обладает свойствами, очень похожими на свойства формульных переменных φ, ψ .

Правило G20 позволяет ввести в язык теории множеств понятие «башни степеней»⁷ [2]:

$$n \uparrow\uparrow k \equiv \underbrace{n^{\overbrace{\dots}^n}}_{k \text{ раз}}; \quad n \uparrow\uparrow 0 \equiv 1.$$

Кроме того, присоединим прописные буквы латинского алфавита A, \dots, Z к списку объектных переменных (правило G3), чтобы иметь возможность синтаксически отличать более сложные конструкции от более простых.

Итак, терм V_r^n назовем r -ым универсумом до- n («доэнной») кратности ($n \geq 2$) и определим его рекурсивно:

V1 : $V_0^n \equiv \{\}$;

V2 : $V_{r+1}^n \equiv \mathcal{P}(V_r^n \times (n - 1))$,

⁷Мы пользуемся стрелочной нотацией Д. Кнута для обозначения башни степеней.

Упражнение 1.8.
Докажите это утверждение индукцией по дереву

здесь берется мультимножество всех подмультимножеств мультимножества $V_r^n \times (n - 1)$, т.е. в предыдущем универсуме кратности всех элементов увеличены до числа $n - 1$, и только потом берутся все подмультимножества.

Теорема 1.2 (об универсумах).

- (1) $a \in^1 V_r^n$ ($r \geq 1, n \geq 2$) тогда и только тогда, когда a — мультимножество ранга $< r$, все узлы гамма-дерева которого имеют кратность $< n$.
(2) Количество элементов (с учетом кратностей) V_r^n равно $n \uparrow\uparrow (r - 1)$.

Доказательство. Часть (1) доказывается индукцией по рангу r при фиксированном n . Для $r = 1$ утверждение очевидно, поскольку $V_1^n \doteq \mathcal{P}(\{\}) \doteq \{\{\}\}$. Предположим, что для r утверждение (1) справедливо и рассмотрим $a \in^1 V_{r+1}^n$. Это равносильно $a \in V_r^n \times (n - 1)$. Отсюда следует, что все элементы a — это такие x , что $x \in^1 V_r^n$ и, кроме того, $x \in^k a$, где $k < n$. Тогда $\text{rank}(x) < r$, а значит, $\text{rank}(a) < r + 1$. Таким образом, если $a \in^1 V_{r+1}^n$, то его ранг $< r + 1$ и кратность узлов $< n$.

Обратно, если $\text{rank}(a) < r + 1$ и кратность узлов гамма-дерева $< n$, то для любого $x \in^k a$ ранг x меньше r и кратность узлов в гамма-дереве, соответствующем x , меньше n . Тогда по предположению $x \in^1 V_r^n$, откуда следует, что $a \in V_r^n \times (n - 1)$.

Доказательство части (2) также проводится индукцией по рангу r . При $r = 1$ имеем $V_1^n \doteq \{\{\}\}$ — количество элементов равно 1, что согласуется с определением $n \uparrow\uparrow 0$.

Пусть V_r^n содержит $n \uparrow\uparrow (r - 1)$ элементов. Мультимножество $\mathcal{P}(V_r^n \times (n - 1))$ строится следующим образом: для каждого элемента V_r^n выбирается кратность $0, 1, \dots, n - 1$, получается кортеж (k_1, \dots, k_s) , где $s = n \uparrow\uparrow (r - 1)$, который взаимно-однозначно определяет подмультимножество $a \in V_r^n \times (n - 1)$. Всего таких наборов, очевидно, $n^s = n \uparrow\uparrow r$. \square

В частности, если положить $n = 2$, то мы получим универсумы для обычных множеств, причем $V_r^2 = \mathcal{P}^r(\{\})$, т.е. операция взятия булеана, проведенная r раз. И количество элементов r -го универсума (состоящего из множеств ранга $< r$) равно $2 \uparrow\uparrow (r - 1)$.

Элементы универсумов V_r^n будем называть мультимножествами до- n («до-энной») кратности.

Наконец, мы вплотную приблизились к заявленной ранее цели — присвоить всем множествам (а заодно и мультимножествам) некий универсальный код, который бы однозначно их идентифицировал и дал возможность упорядочивать (сортировать) любые «начальные» множества.

Символ $\text{Code}^n(a)$ назовем n -ым **универсальным кодом** (или, короче, n -кодом) мультимножества a до- n кратности и определим его рекурсивно:

$$\text{UC1} : \text{Code}^n(\{\}) = 0$$

$$\text{UC2} : \text{Code}^n(a) = \sum_{0 < k < n} \sum_{x \in^k a} kn^{\text{Code}^n(x)}$$

Теорема 1.3 (об универсальных кодах).

- (1) *мультимножества, являющиеся элементами V_r^n имеют коды от 0 до $n \uparrow\uparrow (r-1) - 1$.*
(2) *каждому мультимножеству до- n кратности соответствует единственный n -код, каждому n -коду соответствует единственное мультимножества до- n кратности.*

Доказательство. Утверждение (1) следует из неравенства $\text{Code}^n(a) < n \uparrow\uparrow (r-1)$, если $a \in^1 V_r^n$, которое доказывается индукцией по r .

Действительно, при $r = 1$ если $a \in^1 V_1^n$, то $a \equiv \{\}$ и $\text{Code}^n(a) = 0 = n \uparrow\uparrow (1-1) - 1$. Пусть (1) верно для r и $a \in^1 V_{r+1}^n$.

Тогда для $x \in^k a$ имеем: $x \in^k V_r^n \times (n-1)$, $0 < k < n$. По предположению $\text{Code}^n(x) \leq n \uparrow\uparrow (r-1)$. Следовательно,

$$\begin{aligned} \text{Code}^n(a) &= \sum_{0 < k < n} \sum_{x \in^k a} kn^{\text{Code}^n(x)} \leq (n-1) \sum_{x \in^1 V_r^n} n^{\text{Code}^n(x)} = \\ &= (n-1) \sum_{j=0}^{n \uparrow\uparrow (r-1) - 1} n^j = (n-1) \frac{n^{n \uparrow\uparrow (r-1)} - 1}{n-1} = (n \uparrow\uparrow r) - 1. \end{aligned}$$

Утверждение (1) доказано.

Первая часть утверждения (2) следует из единственности представления числа в виде суммы степеней по основанию n . Вторая часть (2) следует из утверждения (1) данной теоремы и утверждения (2) теоремы 1.2. \square

Утверждение (2) теоремы 1.3 основано на том арифметическом факте, что любое число единственным способом раскладывается по заданному супероснованию $n > 1$. Например, при $n = 2$ имеем разложение: $5 = 2^{2^1} + 1$, $11 = 2^{2^1+1} + 2^1 + 1$,

$$2019 = 2^{2^{2^1+1}+2^1} + 2^{2^{2^1+1}+1} + 2^{2^{2^1+1}} + 2^{2^{2^1}+2^1+1} + 2^{2^{2^1}+2^1} + 2^{2^{2^1}+1} + 2^1 + 1. \quad (1.4)$$

Поскольку мы уже демонстрировали погружение натуральных чисел в теорию «начальных» множеств (а значит, и мультимножеств), мы вправе пользоваться всеми плодами арифметики в своих рассуждениях, в том числе теоремами о единственности разложения, и всеми способами записи арифметических выражений.

Универсальные коды мультимножеств обладают следующими свойствами:

UC3 : если существует код $\text{Code}^n(a)$, то существует код $\text{Code}^{n+1}(a)$, который получается из кода Code^n заменой основания n на $n + 1$ в записи Code^n по супероснованию n ;

UC4 : если обычное множество рассматривать как мультимножество, то для него существуют коды любой кратности: $\text{Code}^2, \text{Code}^3, \dots$;

UC5 : если $\text{Code}^n(a) \leq \text{Code}^n(b)$, то $\text{rank}(a) \leq \text{rank}(b)$;

UC6 : $\text{Code}^n(a) = \text{Code}^n(b)$ тогда и только тогда, когда $a \equiv b$ (если a и b — мультимножества до- n кратности);

UC7 : если $\text{Code}^n(a)$ является одним из показателей степеней в разложении $\text{Code}^n(b)$ в сумму степеней по основанию n , то a есть элемент b , причем с кратностью, равной коэффициенту при $n^{\text{Code}^n(a)}$, в противном случае верно $a \in^0 b$;

UC8 : пусть $a = \mu$, где a — множество, μ — мультимножество до- n кратности, тогда $\text{Code}^2(a)$ получается из кода $\text{Code}^n(\mu)$ заменой всех коэффициентов на 1, заменой n на 2 и удалением дубликатов слагаемых.

Упражнение 1.9. Докажите свойство. Таким образом, если мы рассматриваем мультимножества с кратностями элементов $< n$, то все такие мультимножества взаимно-однозначно кодируются n -кодом, причем элементы универсумов кодируются начальным отрезком ряда натуральных чисел.

В частности, все «начальные» множества кодируются 2-кодом, что позволяет взаимно-однозначно погрузить их в натуральный ряд, так что, зная коды множеств, можно легко отвечать на вопрос об их равенстве: $\text{Code}^2(a) = \text{Code}^2(b)$ тогда и только тогда, когда $a = b$.

Приведем несколько универсальных 2-кодов «начальных» множеств для примера:

$$\begin{aligned} \text{Code}^2(\{\{\}\}) &= 2^{\text{Code}^2(\emptyset)} = 2^0 = 1, & \text{Code}^2(\{\{\}, \{\{\}\}\}) &= 2^0 + 2^1 = 3, \\ \text{Code}^2(\{\{\}, \{\{\}\}, \{\{\}, \{\{\}\}\}) &= 2^0 + 2^1 + 2^3 = 11. \end{aligned}$$

Как видим, в примерах мы нашли универсальный код для множеств-чисел 1, 2 и 3. Выпишем несколько первых кодов и соответствующих им множеств, максимально используя обозначения натуральных чисел при обозначении

элементов множеств:⁸

$$\begin{array}{lll}
 0 = \text{Code}^2(0) & 4 = \text{Code}^2(\{\{1\}\}) & 8 = \text{Code}^2(\{2\}) \\
 1 = \text{Code}^2(1) & 5 = \text{Code}^2(\{\{1\}, 0\}) & 9 = \text{Code}^2(\{2, 0\}) \\
 2 = \text{Code}^2(\{1\}) & 6 = \text{Code}^2(\{\{1\}, 1\}) & 10 = \text{Code}^2(\{2, 1\}) \\
 3 = \text{Code}^2(2) & 7 = \text{Code}^2(\{\{1\}, 1, 0\}) & 11 = \text{Code}^2(3)
 \end{array}$$

$$\text{Code}^2 \left\{ \{2, 1\}, \{2, 0\}, \{2\}, \{\{1\}, 1, 0\}, \{\{1\}, 1\}, \{\{1\}, 0\}, 1, 0 \right\} = 2019.$$

Сравните данную запись с выражением (1.4). Далее,

$$\text{Code}^2(4) = 2059, \quad \text{Code}^2(5) = 2^{2059} + 2059.$$

В общем случае для натуральных чисел имеем следующее рекуррентное выражение:

$$\text{Code}^2(n+1) = 2^{\text{Code}^2(n)} + \text{Code}^2(n), \quad n \geq 0,$$

которое легко доказывается по индукции.

Отсюда же легко видеть, что $2^{\text{Code}^2(n)} > 2 \uparrow\uparrow n$.

Свойства UC6 и UC7, на первый взгляд, избавляют нас от длительной процедуры проверки высказываний $a \in b$ и $a = b$ по алгоритмам **IN** и **EQ** для «начальных» множеств, поскольку для проверки $a \in b$ достаточно обнаружить $\text{Code}^2(a)$ среди степеней числа 2 в представлении кода b в виде суммы степеней по основанию 2, а для проверки равенства $a = b$ необходимо и достаточно просто сравнить коды множеств.

Проблема заключается в том, что эти коды уже для небольших множеств становятся огромными числами (см. примеры выше), а значит, как их вычисление, так и хранение самих множеств в виде кодов в памяти компьютера практически неосуществимо.

Мы еще о нем не забыли?

Выходом может быть промежуточный вариант: каждое множество представлять в виде гамма-дерева, в вершинах которого стоят двойки, ребра обозначают возведение в степень, а параллельные ветки складываются. В этом случае сравнение универсальных кодов множеств будет сводиться к умению машины оперировать символьными арифметическими выражениями, а запись множества в памяти потребует не больше места, чем обычная скобочная запись.

Рассмотрение аналогичной задачи для мультимножеств упирается в принудительное ограничение кратностей их элементов некоторым фиксированным числом n , которое позволит все такие мультимножества закодировать

⁸ Попробуйте сделать то же самое, используя для обозначения множеств только скобочную запись.

взаимно-однозначно n -кодом. Кодирование произвольных мультимножеств было бы возможно, если бы существовала некая условная бесконечная кратность, которую можно было бы использовать в вычислении универсальных кодов в качестве супероснования.

Это обстоятельство ставит перед нами запрос на введение в теорию множеств бесконечных чисел — ординалов, и специальных арифметических правил для них. Правда, полноценная реализация такой теории на компьютере осложняется отсутствием в физическом мире актуальной бесконечности, но на уровне «игры в символы» мы можем оперировать бесконечными множествами, иногда пренебрегая некоторыми их возможностями.

Заметим, что определением универсального кода мы успешно погрузили теорию начальных множеств в арифметику. Далее мы увидим, что и арифметика может быть погружена в теорию множеств. И на самом деле не так уж важно, что считать основой математики — арифметику или теорию множеств. Но только до тех пор, пока мы не выйдем за пределы конечных множеств и не начнем изучать свойства различных бесконечностей.

1.1.10 Теорема Гудстейна

Рассмотрим последовательность чисел:

- (g1) g_2 — произвольное целое число > 0 ;
- (g2) g_{n+1} образуется из g_n следующим способом: число g_n записывается по супероснованию n , затем все вхождения n заменяются на $n + 1$, затем из полученного числа вычитается 1.

Теорема 1.4 (Гудстейн). *Последовательность $\{g_n\}$ за конечное число шагов достигает нуля.*

Проделаем предлагаемые шаги для небольшого начального числа 8:

$$\begin{array}{llll}
g_2 = 8 & = 2^{2^1+1} & \rightarrow 3^{3^1+1} - 1 \\
g_3 = 80 & = 2 \cdot 3^{3^1} + 2 \cdot 3^2 + 2 \cdot 3 + 2 & \rightarrow (2 \cdot 4^{4^1} + 2 \cdot 4^2 + 2 \cdot 4 + 2) - 1 \\
g_4 = 553 & = 2 \cdot 4^{4^1} + 2 \cdot 4^2 + 2 \cdot 4 + 1 & \rightarrow (2 \cdot 5^{5^1} + 2 \cdot 5^2 + 2 \cdot 5 + 1) - 1 \\
g_5 = 6310 & = 2 \cdot 5^{5^1} + 2 \cdot 5^2 + 2 \cdot 5 & \rightarrow (2 \cdot 6^{6^1} + 2 \cdot 6^2 + 2 \cdot 6) - 1 \\
g_6 = 93395 & = 2 \cdot 6^{6^1} + 2 \cdot 6^2 + 1 \cdot 6 + 5 & \dots
\end{array}$$

[illegible]

Каким же образом в итоге мы получим 0?

Рассмотрим последовательность g_n в свете вышеизложенных свойств мультимножеств. Берем произвольное число $g_2 > 0$. При построении g_3 мы

Таблица 1.1: Пример соответствия множеств и мультимножеств в последовательности Гудстейна.

$\{g_n\}$	Мультимножество	Множество
g_3		
g_4		
g_5		
g_6		

сначала строим гамма-дерево мультимножества a , соответствующего 2-коду числа g_2 путем представления g_2 по супероснованию 2 (см. пример (1.4)), затем в этом представлении заменяем 2 на 3 и, тем самым, получаем 3-код того же самого мультимножества. От полученного 3-кода отнимаем 1.

Ясно, что при замене 2 на 3 ни гамма-дерево, ни соответствующее мультимножество

тимножество a не изменятся, т. е. после такой замены мы просто получим код $\text{Code}^3(a)$ для того же самого мультимножества. После вычитания 1 мы получим новый код $\text{Code}^3(a) - 1$, соответствующий, вообще говоря, другому мультимножеству, но при этом ранг нового мультимножества не будет превышать ранг исходного множества (свойство UC5).

Таким образом, итерации последовательности Гудстейна не приводят к увеличению ранга мультимножеств, соответствующих n -кодам g_n .

Далее. На каждом шаге мы имеем код некоторого мультимножества, которому соответствует единственное множество (получаемое путем устранения кратностей). Пример множеств и мультимножеств для последовательности с начальным числом 8 приведен в таблице 1.1. Очевидно, что последовательность g_n описывает множества ранга не выше заданного изначально, а их существует конечный набор.

Но тогда, если исключить попадание последовательности g_n в ноль, то последовательность g_n бесконечно часто попадает в такие коды мультимножеств, что соответствующее им множество получается одно и то же.

К сожалению, этот факт хоть и чрезвычайно интересен, но теорему Гудстейна не доказывает. Он всего лишь демонстрирует некоторые странности поведения последовательности g_n .

Предположим теперь, что стартовое число g_2 маленькое, а N — некоторое очень большое супероснование, которое, как мы полагаем, последовательность g_n «не сможет» достичь.

Элемент g_n по построению является n -кодом некоторого мультимножества μ_n . Положим теперь $f_n = \text{Code}^N(\mu_n)$ (Внимание! Это возможно только в том случае, когда $n \leq N$), т. е. f_n получается представлением g_n по супероснованию n с последующей заменой n на N .

Формально, определим рекурсивно следующую функцию от натурального числа:

$$\begin{aligned} \mathbf{S}_n^m(0) &\equiv 0; \\ \mathbf{S}_n^m(n^s \cdot k) &\equiv m \mathbf{S}_n^{(s)} \cdot k; \\ \mathbf{S}_n^m(n^{s_0} \cdot k_0 + \dots + n^{s_j} \cdot k_j + k) &\equiv \mathbf{S}_n^m(n^{s_0} \cdot k_0) + \dots + \mathbf{S}_n^m(n^{s_j} \cdot k_j) + k, \end{aligned} \quad (1.5)$$

где $1 < n < m$; $0 < k_0, \dots, k_j < n$; $0 < s_j < \dots < s_0$ и $0 \leq k < n$, $0 \leq s$.

Лемма 1.1. Функция $\mathbf{S}_n^m(t)$, $1 < n < m$, строго возрастает по t .

Доказательство. Докажем неравенство

$$\mathbf{S}_n^m(t) < \mathbf{S}_n^m(t+1) \quad (1.6)$$

Очевидно, оно выполняется для $t = 0$, т. к. $\mathbf{S}_n^m(0) = 0$ и $\mathbf{S}_n^m(1) = 1$. Предположим, что оно верно для всех $t < T$ и докажем его для T .

Пусть $T+1 = n^{s_0} \cdot k_0 + \dots + n^{s_j} \cdot k_j + k$, где $s_0 > \dots > s_j > 0$, $n > k_0, \dots, k_j > 0$, $n > k \geq 0$.

Случай первый: $k > 0$. Тогда $T = n^{s_0} \cdot k_0 + \dots + n^{s_j} \cdot k_j + (k-1)$. Тогда из (1.5) имеем:

$$\mathbf{S}_n^m(T) = m_n^{s_0} \cdot k_0 + \dots + m_n^{s_j} \cdot k_j + (k-1) = \mathbf{S}_n^m(T+1) - 1,$$

поэтому в данном случае требуемое неравенство верно.

Случай второй: $k = 0$. Тогда представим $T+1$ следующим образом:

$$T+1 = n^{s_0} \cdot k_0 + \dots + n^{s_j} \cdot (k_j - 1) + n^{s_j},$$

Откуда следует, что

$$\begin{aligned} T &= n^{s_0} \cdot k_0 + \dots + n^{s_j} \cdot (k_j - 1) + n^{s_j} - 1 = \\ &= n^{s_0} \cdot k_0 + \dots + n^{s_j} \cdot (k_j - 1) + (n^{s_j-1} + \dots + n^0)(n-1). \end{aligned}$$

Тогда

$$\begin{aligned} \mathbf{S}_n^m(T+1) &= m_n^{s_0} \cdot k_0 + \dots + m_n^{s_j} \cdot (k_j - 1) + m_n^{s_j} \\ \mathbf{S}_n^m(T) &= m_n^{s_0} \cdot k_0 + \dots + m_n^{s_j} \cdot (k_j - 1) + \left(m_n^{s_j-1} + \dots + m_n^{s_0} \right) (n-1) \end{aligned}$$

Осталось заметить, что $s_j - 1 < T$ (т. к. $T > n^{s_j-1}$), поэтому для последовательности $0 < 1 < \dots < s_j - 1$ можно воспользоваться предположением индукции: $\mathbf{S}_n^m(0) < \mathbf{S}_n^m(1) < \dots < \mathbf{S}_n^m(s_j - 1)$. Отсюда следует, что

$$\left(m_n^{s_j-1} + \dots + m_n^{s_0} \right) (n-1) < m_n^{s_j}, \quad (1.7)$$

поскольку

$$\begin{aligned} \left(m_n^{s_j-1} + \dots + m_n^{s_0} \right) (n-1) &\leq \left(\sum_{i=0}^{s_j-1} m_n^i \right) (n-1) = \\ &= m_n^{s_j-1+1} \cdot \frac{n-1}{m-1} \leq m_n^{s_j} \cdot \frac{n-1}{m-1} < m_n^{s_j}. \end{aligned}$$

Из (1.7) следует $\mathbf{S}_n^m(T) < \mathbf{S}_n^m(T+1)$. Индукция завершена. \square

Теперь с помощью функции \mathbf{S}_n^m определим величины g_n и f_n , о которых шла речь выше:

$$g_{n+1} = \mathbf{S}_n^{n+1}(g_n) - 1, \quad n \geq 2;$$

$$f_n = \mathbf{S}_n^N(g_n), \quad n \geq 2$$

Лемма 1.2. f_n строго убывает, пока $n < N$.

Доказательство. Действительно,

$$f_{n+1} = \mathbf{S}_{n+1}^N(g_{n+1}) = \mathbf{S}_{n+1}^N(\mathbf{S}_n^{n+1}(g_n) - 1) < \mathbf{S}_{n+1}^N(\mathbf{S}_n^{n+1}(g_n)) = \mathbf{S}_n^N(g_n) = f_n,$$

где мы воспользовались монотонностью \mathbf{S}_{n+1}^N и равенством $\mathbf{S}_{n+1}^N(\mathbf{S}_n^{n+1}(t)) = \mathbf{S}_n^N(t)$, которое можно легко проверить индукцией по t . \square

Итак, мы видим, что последовательность f_n строго убывает с ростом n , и если бы могли продолжать ее сколь угодно долго, то рано или поздно достигли бы 0, а это означало бы что соответствующее g_n также равно 0.

Проблема заключается в том, что число N можно найти, только уже зная результат теоремы Гудстейна, а значит, и доказать ее таким способом невозможно.

Выход из этой ситуации может быть найден только тогда, когда мы число N заменим на «число», заведомо большее всех натуральных чисел. А это можно сделать, только имея аксиому бесконечности и арифметику бесконечных чисел — ординалов. В 1982 г. Керби и Парис показали, что в теореме Гудстейна используется предположение о совместимости арифметики, и поэтому в силу второй теоремы Гёделя о неполноте теорема Гудстейна не может быть доказана в рамках обычной арифметики.

Задачей следующей части главы 1 будет как раз построение строгой аксиоматической теории множеств с бесконечными ординалами и изучение некоторых ее особенностей. После чего мы построим модель мультимножеств в данной аксиоматике, значительно расширив объем этого понятия.

Таким образом, теорема Гудстейна утверждает очень глубокий факт, связывающий теорию чисел, логику и теорию множеств, т. е. основания математики.

В приложении (см. листинг С.2) приведен листинг программы на языке Python, которая вычисляет последовательность Гудстейна до некоторых пределов (числа не более 1000 десятичных цифр и шагов не более 100). Конечно, программа работает, что называется, «в лоб», сначала раскидывая число по

супероснованию, фактически создавая дерево соответствующего мультимножества из массивов и ссылок, затем идет по этому дереву в обратном порядке, заменяя основание и вычисляя результат. Это тяжелое по времени и памяти, но наглядное решение. Можно было бы не вычислять промежуточные числа последовательности, а вместо этого работать с деревьями мультимножеств, редуцируя их по определенным правилам. Как мы уже отмечали выше, высота дерева при этом не растет, так что количество вершин в дереве числа g_n пропорционально логарифму этого числа (причем основание логарифма все время растет), что намного экономичнее в плане вычислительных ресурсов компьютера.

В сказочной формулировке это редуцирование называется отсечением голов дракона.

1.2 Аксиоматика Цермело–Френкеля

После того, как мы определили взаимосвязь между объектами-множествами и грамматическими конструкциями — объектными записями, мы в принципе можем обойтись без аксиоматики, т. к. грамматических правил G1–G17 вполне достаточно, чтобы конструировать любые «начальные» множества. Однако, мы хотим теперь перейти к общей (аксиоматической) теории множеств, частью которой является теория «начальных» множеств. И чтобы этот переход был наиболее комфортным для читателя, мы рассмотрим здесь последовательно ряд аксиом, увидим их связь с грамматическими конструкциями и поймем, что объектные записи грамматики являются удачной моделью для этих аксиом.

В данном разделе мы возвращаемся к грамматике, состоящей только из правил G1–G17.

1.2.1 Равенство и единственность

Аксиомы объемности (AV) и экстенциональности (AE) в разных источниках формулируются по-разному. Приведем три наиболее распространенных варианта.

Var1 (AV1) $a = b \leftrightarrow \forall x (x \in a \leftrightarrow x \in b)$

(AE1) $a = b \rightarrow \forall x (a \in x \rightarrow b \in x)$

Var2 (AV2) $\forall x (x \in a \leftrightarrow x \in b) \rightarrow a = b$

(AE2[φ]) $a = b \rightarrow (\varphi(a) \rightarrow \varphi(b))$

Var3 (AV3) $a = b \leftrightarrow \forall x (x \in a \leftrightarrow x \in b)$

(AE3[φ]) $a = b \rightarrow (\varphi(a) \rightarrow \varphi(b))$

Заметим, что во втором и третьем вариантах используется схема аксиом — аксиома экстенциональности задается для каждой формулы φ .

Теорема 1.5. Вар1, Вар2 и Вар3 попарно эквивалентны.

Доказательство. Докажем, что Вар1 \rightarrow Вар2, Вар2 \rightarrow Вар3 и Вар3 \rightarrow Вар1.

1) Вар1 \rightarrow Вар2. Очевидно, что AV2 является следствием AV1, поскольку из соотношения $\varphi \leftrightarrow \psi$ следует $\psi \rightarrow \varphi$.

Утверждение

$$(a = b \leftrightarrow \forall x (x \in a \leftrightarrow x \in b)) \wedge (a = b \rightarrow \forall x (a \in x \rightarrow b \in x)) \rightarrow \\ \rightarrow (a = b \rightarrow (\varphi(a) \rightarrow \varphi(b))) \quad (1.8)$$

доказывается индукцией по сложности формулы φ и входящих в нее термов.

Пусть $a = b$. Как мы помним из раздела о грамматике, любую формулу можно представить в виде дерева, листьями которого будут объектные переменные и термы-константы. Формулы $\varphi(a)$ и $\varphi(b)$ отличаются только тем, что все вхождения a в первой заменены на b во второй, а вся остальная структура, термы и логические связки посимвольно и позиционно совпадают.

Сначала докажем (1.8) в предположении, что формула φ не содержит кванторного терма $\{x | \psi\}$. Одновременно это означает, что и все термы, входящие в φ , имеют простой вид, т.е. построены только по правилам G1–G7 грамматики. В этом случае легко показать, что если $a = b$, то замена переменной a на переменную b в простом терме приводит к равному терму. Это следует из AE1, а также из AV1. Действительно, экстенциональность и посылка $a = b$ позволяют нам доказать, что отношение $a \in t$ инвариантно относительно замены переменных $a || b$, а объемность показывает, что равны термы, списки которых отличаются только заменой переменных $a || b$. Далее, пользуясь индукцией по сложности терма, приходим к выводу, что простые термы, отличающиеся только данной заменой переменных, равны.

Утверждение для формул доказывается аналогично, индукцией по сложности формулы. Если формула $\varphi(a)$ атомарная, т.е. имеет вид $(a \in x)$, то AE2 для нее следует прямо из AE1. Предположим теперь, что AE2 выполняется для $\varphi(a)$ и $\psi(a)$, и покажем, что она будет верна для $\neg\varphi(a)$, $\varphi(a) \wedge \psi(a)$ и $\forall x \varphi(a, x)$, где x свободно входит в φ . Проще всего доказательство для $\neg\varphi(a)$ увидеть на коммутативной диаграмме:

$$\begin{array}{ccc} (b = a) & \xleftarrow{\text{AV1}} & (a = b) \\ \text{предположение} \downarrow & \swarrow & \downarrow \text{ЧТД} \\ (\varphi(b) \rightarrow \varphi(a)) & \xrightarrow{\text{контрапозиция}} & (\neg\varphi(a) \rightarrow \neg\varphi(b)) \end{array}$$

Доказательство для $\varphi(a) \wedge \psi(a)$. По предположению имеем $(a = b) \rightarrow (\varphi(a) \rightarrow \varphi(b)) \wedge (\psi(a) \rightarrow \psi(b))$. Далее нужно получить импликацию

$$(\varphi(a) \rightarrow \varphi(b)) \wedge (\psi(a) \rightarrow \psi(b)) \rightarrow ((\varphi(a) \wedge \psi(a)) \rightarrow (\varphi(b) \wedge \psi(b))). \quad (1.9)$$

Заметим, что если $\varphi(a)$ или $\psi(a)$ ложно, то импликация $(\varphi(a) \wedge \psi(a)) \rightarrow (\varphi(b) \wedge \psi(b))$ истинна, а значит, истинна и импликация (1.9). Пусть $\varphi(a)$ и $\psi(a)$ одновременно истинны. Тогда $(\varphi(a) \rightarrow \varphi(b)) \wedge (\psi(a) \rightarrow \psi(b))$ равносильно $\varphi(b) \wedge \psi(b)$, а формула $(\varphi(a) \wedge \psi(a)) \rightarrow (\varphi(b) \wedge \psi(b))$ равносильна также $\varphi(b) \wedge \psi(b)$, значит, импликация (1.9) равносильна в этом случае $\varphi(b) \wedge \psi(b) \rightarrow \varphi(b) \wedge \psi(b)$, т. е. тоже истинна. Таким образом, для конъюнкции шаг индукции тоже доказан.

Квантор всеобщности вводится в формулу φ по правилу Бернайса:

$$\frac{a = b \rightarrow \varphi(a, x)}{a = b \rightarrow \forall x \varphi(a, x)}$$

Как мы знаем из матлогики, все остальные логические связки и квантор существования могут быть выражены через уже рассмотренные:

$$\begin{aligned}\varphi \vee \psi &\leftrightarrow \neg(\neg\varphi \wedge \neg\psi), \\ \varphi \rightarrow \psi &\leftrightarrow \neg\varphi \vee \psi, \\ \varphi \leftrightarrow \psi &\leftrightarrow (\varphi \rightarrow \psi) \wedge (\psi \rightarrow \varphi) \\ \exists x \varphi &\leftrightarrow \neg(\forall x \neg\varphi).\end{aligned}$$

Итак, индукцией по сложности термов и формул мы показали, что из аксиом AV1 и AE1 следуют аксиомы AV2 и AE2 для простых формул, т. е. формул без кванторного терма $\{x \mid \varphi(x)\}$. Отсюда, пользуясь свойством (1.3), нетрудно показать, что произвольный терм $\{x \mid \varphi(x)\}$, содержащий простую формулу φ , инвариантен относительно равенства при замене переменных $a \parallel b$, поскольку истинность формулы φ при этом сохранится.

Наконец, для любой формулы φ можно показать, что из аксиом AV1 и AE1 следуют аксиомы AV2 и AE2. Для этого нужно снова воспользоваться индукцией по сложности формулы и тем, что все атомарные формулы вида $(x \in y)$ с любыми термами x, y (в том числе содержащими кванторные термы), удовлетворяют AE2, как показано выше.

На этом доказательство Вар1 \rightarrow Вар2 закончено.

2) Вар2 \rightarrow Вар3. Очевидно, что из AE2 следует AE3. Покажем, что из AV2 и AE2 следует AV3. Для этого достаточно показать, что

$$a = b \rightarrow \forall x(x \in a \leftrightarrow x \in b).$$

Положим $\varphi(a) \equiv (x \in a)$. Тогда из AE2 следует $a = b \rightarrow (x \in a \rightarrow x \in b)$.

Положим $\varphi(a) \equiv (x \in b \rightarrow x \in a)$. Тогда из AE2 следует, что $a = b \rightarrow ((x \in b \rightarrow x \in a) \rightarrow (x \in b \rightarrow x \in b))$. Последняя формула эквивалентна следующей $a = b \rightarrow (x \in b \rightarrow x \in a)$.

Таким образом, из AE2 следует $a = b \rightarrow (x \in a \leftrightarrow x \in b)$, и по правилу введения квантора всеобщности, из AE2 следует $a = b \rightarrow \forall x(x \in a \leftrightarrow x \in b)$.

3) $\text{Var}3 \rightarrow \text{Var}1$. Очевидно, что $\text{AV}3$ выводит $\text{AV}1$, т. к. они тождественны.

Далее, $\text{AE}3$ выводит формулу $a = b \rightarrow (a \in x \rightarrow b \in x)$, если подставить $\varphi(a) \Leftrightarrow (a \in x)$. Следовательно, по правилу введения квантора всеобщности $\text{AE}3$ выводит $\text{AE}1$.

Теорема доказана. \square

В дальнейшем мы воспользуемся первым вариантом аксиом объемности и экстенциональности, чтобы не вводить без надобности схему аксиом (т. е. набор аксиом, зависящих от выбора формулы φ), но, учитывая доказанную теорему, пользоваться будем самым сильным — третьим — вариантом, причем в силу симметричности равенства $\text{AE}3$ можно усилить до $\text{AE}4$: $a = b \rightarrow (\varphi(a) \leftrightarrow \varphi(b))$.

Отметим также, что ослабить Аксиому объемности до варианта $\text{AV}0$ $a = b \rightarrow \forall x (x \in a \leftrightarrow x \in b)$ не получится, т. к. это приведет к потере симметричности равенства и, вообще говоря, равенство двух множеств не будет выводиться из аксиом.

В разделе 1.1.5 мы подробно обсуждали, как равенство и принадлежность связаны с грамматическими конструкциями «начальных» множеств, и ввели соглашения $\text{C}1$ и $\text{C}2$, по смыслу прямо соответствующие аксиомам AE и AV в варианте 1. Таким образом, соглашения о равенстве и принадлежности в грамматике моделируют формальные аксиомы $\text{AE}+\text{AV}$ аксиоматической теории.

Равенство множеств обладает следующими свойствами:

$\text{E}1$ (рефлексивность): $x = x$ для любого множества x ,

$\text{E}2$ (симметричность): если $a = b$, то $b = a$,

$\text{E}3$ (транзитивность): если $a = b$ и $b = c$, то $a = c$.

Упражнение
1.10.
Докажите
свойства
равенства

Понятие равенства позволяет определить понятие **единственности**. Мы говорим, что x , удовлетворяющий формуле $\varphi(x)$, единственный, если, во-первых, истинно $\varphi(x)$, а во-вторых, для любого y из $\varphi(y)$ следует $x = y$. Выше существование единственного множества, удовлетворяющего φ мы записывали как $\exists!x \varphi(x)$. Этим обозначением мы будем часто пользоваться в дальнейшем.

Теорема 1.6. Если существует множество $t = \{x \mid \psi(x)\}$, то оно единственное.

Доказательство. По определению кванторного термина t удовлетворяет формуле:

$$\varphi(t) \Leftrightarrow \forall a (a \in t \leftrightarrow \psi(a)).$$

Здесь мы просто в определении (1.3) заменили кванторный терм равным ему термом t , пользуясь аксиомой АЕ.

Пусть $\exists t \varphi(t)$. Предположим также, что для некоторого y тоже $\varphi(y)$. Тогда

$$\forall a (a \in y) \leftrightarrow \psi(a) \leftrightarrow (a \in t),$$

откуда следует $\forall a (a \in y) \leftrightarrow (a \in t)$, откуда по аксиоме АV получаем $y = t$.

Следовательно, $\exists t \varphi(t) \rightarrow \exists! t \varphi(t)$. \square

1.2.2 Ограничительные аксиомы

Добавим три аксиомы и одну схему к двум уже имеющимся аксиомам (АV и АЕ):

$$\mathbf{AU} : \forall x \neg(\forall y y \notin x) \rightarrow \exists u (z \in u \leftrightarrow \exists w z \in w \in x)$$

$$\mathbf{AP} : \forall x \exists p \forall y (y \in p \leftrightarrow (\forall z z \in y \rightarrow z \in x))$$

$$\mathbf{AF}[\varphi] : \forall x ((\forall y \forall w \forall z (y \in x \wedge \varphi(y, w) \wedge \varphi(y, z) \rightarrow w = z) \rightarrow \rightarrow \exists t \forall z (z \in t \leftrightarrow \exists y (y \in x \wedge \varphi(y, z))))$$

$$\mathbf{AR} : \forall x (\exists y y \in x) \rightarrow \exists y (y \in x \wedge (\forall z \neg(z \in x \wedge z \in y)))$$

Впечатление, что здесь кто-то вставил код на Ассемблере...

Чтобы привести аксиомы к более понятному виду, воспользуемся ранее введенными сокращениями для записей языка, а кроме того, наконец-то, введем обозначение пустого множества:

$\emptyset \equiv \{x \mid (x \in x) \wedge (x \notin x)\}$. Из (1.3) нетрудно видеть, что для

любого a утверждение $a \in \emptyset$ ложно. Существование и единственность этого множества мы докажем чуть позже, а пока переформулируем аксиомы на более понятном языке:

$$\mathbf{AU} : \forall x (x \neq \emptyset) \rightarrow \exists u = \{z \mid \exists w (z \in w \in x)\}$$

$$\mathbf{AP} : \forall x \exists p = \{y \mid y \subseteq x\}$$

$$\mathbf{AF} : \forall x ((\forall y \forall w \forall z (y \in x \wedge \varphi(y, w) \wedge \varphi(y, z) \rightarrow w = z) \rightarrow \exists t = \{z \mid \exists y \in x \varphi(y, z)\})$$

$$\mathbf{AR} : \forall x (x \neq \emptyset) \rightarrow \exists y \in x (x \cap y = \emptyset)$$

Аксиома объединения (или суммы) AU говорит о том, что для любого непустого множества x существует множество, являющееся объединением всех его элементов. Такое множество обозначается $\cup x$, т. е.

$$\cup x \equiv \{z \mid \exists w (z \in w \in x)\}.$$

В частности, для пары $\{a, b\}$ имеем $\cup\{a, b\} = a \cup b$, а для синглтона получим: $\cup\{a\} = a$. *Упражнение 1.11.*

Нетрудно видеть, что если $\cup x$ существует, то для каждого x оно единственно, поскольку определяется кванторным термом.

В грамматике «начальных» множеств аксиома суммы не требуется, т. к. $\cup x$ легко построить по правилам грамматики. Действительно, пусть $a = \{a_1, a_2\}$, $b = \{b_1, b_2\}$. Это значит, что на некотором этапе построения множеств a и b будут построены Список1, Список2, Список3, Список4, соответственно, для множеств a_1, a_2, b_1, b_2 . Тогда для получения $a \cup b$ достаточно построить Список вида 'Список1, Список2, Список3, Список4' и заключить его в фигурные скобки. Элементами Списка будут элементы a или b , и наоборот. Аналогично рассуждая можно показать построение любого $\cup x$ для $x \neq \{\}$.

В терминах гамма-деревьев построение $\cup x$ объяснить еще проще. Если множеству x соответствует некоторое гамма-дерево, то нужно в нем «выбросить» весь первый слой вершин, а корень напрямую соединить с вершинами второго слоя (если они есть) — получится гамма-дерево для $\cup x$.

Аксиома степени AP говорит о существовании множества всех подмножеств (булеана) любого множества x . По определению положим:

$$\mathcal{P}(x) = \{y \mid y \subseteq x\}.$$

Для каждого x это множество единственное, опять-таки, по построению через кванторное обозначение. Иногда можно встретить следующее обозначение степенного множества: $\exp(x)$.

В свете теории «начальных множеств» булеан строится по правилам грамматики следующим образом. Если $x = \{\text{Список}\}$, то существует не только алгоритм построения Списка, но и любой его части (подсписка), состоящей из произвольно выбранных элементов Списка. Следовательно, существует Спiсок', содержащий в себе все такие подспски Списка, заключенные в фигурные скобки, и тогда $\{\text{Список}'\}$ соответствует булеану x .

Упражнение 1.12. Построение гамма-дерева булеана мы предлагаем провести читателю самостоятельно.

В дальнейшем мы увидим, что две аксиомы AU и AP вводят две симметричные операции на множествах: одна (AU) понижает ранг исходного множества на 1,⁹ а другая (AP) повышает его на 1.

Итак, двум конструктивным аксиомам аксиоматической теории множеств в теории «начальных» множеств соответствует наличие алгоритма построения суммы и булеана.

Разберем подробнее **аксиому подстановки** AF. Пусть у нас имеется некая формула $\varphi(y, z)$ и непустое множество x , и для каждого $y \in x$ либо $\varphi(y, z)$ ложно, либо такое z единственное. В этом случае говорят, что φ **определяет отображение** на множестве x . Аксиома AF в этом случае гарантирует, что существует множество всех значений такого отображения $\{z \mid \exists y \in x \varphi(y, z)\}$.

⁹Это справедливо для конечных множеств, а для бесконечных ранг может остаться тем же, например, $\cup \omega = \omega$. Множество ω мы определим чуть позже.

Иначе говоря, если формула $\varphi(y, z)$ каждому элементу $y \in x$ сопоставляет не более одного множества z (неважно, где оно находится и что из себя представляет), то существует множество, элементами которого являются такие и только такие z . Забегая вперед, можно сказать, что $\varphi(y, z)$ задает функцию из x в $\{z \mid \exists y \in x \varphi(y, z)\}$, и область значений этой функции является множеством. Заметим, что если множество x пустое, то таково же и множество $\{z \mid \exists y \in x \varphi(y, z)\}$.

Упражнение
1.13.

Если посмотреть на реализацию АФ в теории «начальных» множеств, то нужно для начала понять, что формула $\varphi(y, z)$ сама по себе уже является алгоритмом, на вход которого подаются произвольные объектные записи y, z , а на выходе за конечное число шагов получается истина или ложь. При этом, если y пробегает некоторое множество x , то этот цикл также конечен, ибо множество x конечно (если, конечно, мы пользуемся некоторой его хорошей объектной записью без кратностей элементов). Но возникает проблема с поиском z таких, что $\varphi(y, z)$ при некотором заданном y . Потому что если такого z не существует, то нам нужно, вообще говоря, перебрать все имеющиеся множества (которых бесконечно много), чтобы получить в ответ ложь.

Поэтому мы не можем утверждать, что аксиома АФ является конструктивной и реализуется полностью в теории «начальных» множеств. Но дело в том, что подобного рода аксиомы доставляют неудобство не только компьютерам и программистам. Математики тоже стремятся все свои формулы сводить к некоторому замкнутому виду так, чтобы были четко определены как шаги алгоритма, так и критерии его остановки при вычислении значения формулы или терма. Отсюда выросла теория вычислимых и рекурсивных функций, центральное место в которой занимают **машина Тьюринга** и **тезис Чёрча** [21].

Недаром мы тут с самого начала рассуждали о том, как научить компьютер теории множеств.

С этой точки зрения, если мы условимся считать, что формула φ в аксиоме АФ конструктивна (например, реализует рекурсивную функцию, позволяющую за конечное число шагов вычислить z , зная y), то построить за *конечное число шагов* множество, являющееся областью значений отображения, заданного формулой φ , уже не составит никакого труда: мы сведем рекурсию, «защитую» в φ , к рекурсиям грамматики, получим список значений z для заданных $y \in x$, и построим итоговое множество.

Конечно, если рассуждать совсем строго, то нам потребуется сначала описать, какими правилами грамматики и в какой последовательности мы должны задавать формулу φ , затем — какими правилами грамматики и в какой последовательности мы должны по заданному y строить z , используя формулу φ (подобно тому, как происходит построение ДНК по химической формуле, записанной в РНК), и, наконец, как из полученного собрать множество всех таких z .

Но для понимания картины в целом достаточно знать, что реализация АФ в компьютерном мире «начальных» множеств требует некоторых конструктивных ограничений. И в этом смысле аксиоматическая теория множеств уже выглядит более сильной, чем теория «начальных» множеств, основанная на грамматике, хотя это превосходство у многих математиков и вызывает сомнения.

Тем не менее, АФ — это мощная аксиома, позволяющая строить достаточно большие совокупности множеств, избегая известных парадоксов. Мощь аксиомы демонстрирует следующая

Парадокс
Расселла,
парадокс
Бурали-
Форти и
т.п.

Теорема 1.7 (следствие аксиомы подстановки). Пусть $\psi(y)$ — произвольная формула со свободной переменной y , x — произвольное множество. Тогда терм $\{y \mid (y \in x) \wedge \psi(y)\}$ обозначает множество, однозначно определяемое формулой ψ и множеством x .

Доказательство. Рассмотрим формулу

$$\varphi(y, z) \equiv \psi(y) \wedge (y = z).$$

Нетрудно видеть, что $\varphi(y, z)$ определяет (тождественное) отображение на своей области истинности. Действительно, если $\varphi(y, z) \wedge \varphi(y, w)$, то, очевидно, $\psi(y) \wedge (y = z) \wedge (y = w)$, откуда следует $z = w$. Следовательно, по аксиоме АФ существует область значений $\varphi(y, z)$ при $y \in x$, которая задается термом

$$\{z \mid \exists y \in x \varphi(y, z)\} = \{z \mid \exists y (y \in x) \wedge \psi(y) \wedge (y = z)\} = \{z \mid (z \in x) \wedge \psi(z)\},$$

а это искомым терм с точностью до замены связанной переменной.

Однозначность определения такого множества следует из свойств терма-квантора. \square

Примем следующее определение:

$$\{x \in y \mid \varphi(x)\} \equiv \{x \mid (x \in y) \wedge \varphi(x)\}.$$

Этот терм обозначает **определяемую** формулой φ **часть множества** y и является множеством в силу доказанной теоремы. В дальнейшем мы часто будем пользоваться данным обозначением, не задумываясь о существовании и единственности такого множества.

Интересно заметить, что уже *определяемая часть* множества вполне себе реализуется в «начальных» множествах, поскольку для ее построения нужно всего лишь произвести конечный перебор элементов исходного множества и для каждого из них за конечное же время проверить выполнение формулы φ . Теорема 1.7 иногда включается в список аксиом теории множеств вместо аксиомы АФ, но нужно иметь ввиду, что она объективно слабее этой аксиомы.

Следствие 1.1. *Пустое множество существует и единственно.*

Доказательство. Данное утверждение вытекает из предыдущей теоремы, если в качестве ψ выбрать любое заведомо ложное высказывание¹⁰, например, $y \in y \wedge y \notin y$. \square

Пустое множество в скобочной записи выражается минимально возможной объектной записью $\{\}$. Действительно, по принятым нами согласованиям грамматики с отношениями равенства и принадлежности (C1 и C2) запись $\{\}$ не имеет списка, а значит, множество $\{\}$ не имеет элементов, т. е. оно пустое. С другой стороны, никакая другая объектная запись не может обозначать пустое множество, поскольку при большем количестве скобок в объектной записи Список внутри внешних фигурных скобок обязательно будет не пустой, а значит, не может обозначать пустое множество. Учитывая, что определенное ранее число 0 (см. раздел 1.1.7) как второе обозначение для $\{\}$, теперь также обозначает пустое множество, получаем, что

$$\emptyset = \{\} = 0$$

— обозначения пустого множества, первое принято в самой теории множеств и в программировании, второе — в теории «начальных» множеств, а третье нас отсылает к алгебраическому смыслу пустого множества.

Аксиома степени AP позволяет однозначно определить следующее множество $\mathcal{P}(\{\})$. Нетрудно показать, что оно состоит из единственного элемента — пустого множества, т. е. $a \in \mathcal{P}(\{\}) \leftrightarrow a = \{\}$, а значит, $\mathcal{P}(\{\}) = \{x \mid x = \{\}\}$, что по определению синглтона записывается как $\{\{\}\}$. То есть $\mathcal{P}(\{\}) = \{\{\}\}$.

Аналогично, легко получить равенство $\mathcal{P}(\mathcal{P}(\{\})) = \{\{\}, \{\{\}\}\}$. *Упражнение 1.14.*
Здесь мы вспомним еще одну числовую константу из раздела 1.1.7 и обозначим $1 \equiv \{\{\}\}$. Таким образом, $\mathcal{P}(\{\}) = \{0\}$ и $\mathcal{P}(\mathcal{P}(\{\})) = \{0, 1\}$.

Ранее мы вводили обозначение для синглтона и пары произвольных множеств. Настало время их «узаконить».

*А мы уж
засидались*

Следствие 1.2. *Для любых множеств a и b синглетон $\{a\}$ и пара $\{a, b\}$ существуют и единственны.*

Доказательство. Поскольку у нас имеется одна готовая пара $\{0, 1\}$, проще всего построить определяемое отображение этой пары на пару $\{a, b\}$, и тогда по аксиоме подстановки AF пара получит право на существование. Пусть

$$\varphi(y, z) \equiv (y = 0 \rightarrow z = a) \wedge (y = 1 \rightarrow z = b).$$

¹⁰На самом деле, нам еще нужно выбрать какое-то произвольное множество x , частью которого и будет пустое множество. Но любая формальная теория, если она непротиворечива, предполагает, что существует хотя бы один объект этой теории. В нашем случае мы, предполагая совместность теории множеств, считаем, что хотя бы одно множество существует, и уже на основании этого доказываем существование пустого множества.

Легко видеть, что такая формула определяет отображение на множестве $\{0, 1\}$, следовательно, его область значений $\{z \mid \exists y \in \{0, 1\} \varphi(y, z)\}$ является множеством и по свойствам терма-квантора определяется единственным образом.

Нетрудно показать, что $\exists y \in \{0, 1\} \varphi(y, z)$ эквивалентно $(z = a) \vee (z = b)$. Действительно, если $a \neq z \neq b$, то обе импликации в формуле $\varphi(y, z)$ истинны тогда и только тогда, когда $y \notin \{0, 1\}$, а это значит, что $\neg(\exists y \in \{0, 1\} \varphi(y, z))$. Тогда по правилу контрапозиции получаем, что

$$(\exists y \in \{0, 1\} \varphi(y, z)) \rightarrow ((z = a) \vee (z = b)).$$

Обратная импликация еще более очевидна. Это означает, что

$$\{z \mid \exists y \in \{0, 1\} \varphi(y, z)\} = \{z \mid (z = a) \vee (z = b)\} = \{a, b\},$$

где последнее равенство дано в силу определения пары.

Итак, пара $\{a, b\}$ существует и единственна для данных множеств a, b .

Если теперь положить $a = b$, то пара $\{a, b\}$ превратится в синглетон $\{a\}$ в силу аксиомы объемности АЕ, следовательно, он также существует и единственен. \square

Из «законности» пары сразу же следует «законность» объединения $a \cup b$, т. к. $a \cup b = \bigcup \{a, b\}$, как мы уже видели выше.

Далее, по теореме 1.7 для произвольных множеств a и b существует и единственно множество $\{x \in a \mid x \in b\}$, так что введенное нами ранее пересечение $a \cap b$ задает множество как определяемую часть множества a (или b).

Аналогично, разность $a \setminus b = \{x \in a \mid x \notin b\}$ существует и единственна для любых множеств a и b .

Если вспомнить алгебраическую терминологию и считать символы \cup, \cap, \setminus операциями на множествах, то можно заметить, что первые две операции (объединение и пересечение) коммутативны, а третья — нет. То есть $a \cup b = b \cup a$, $a \cap b = b \cap a$, но $a \setminus b$ не всегда равно $b \setminus a$.

Осталось рассмотреть **Аксиому регулярности** AR. Основная роль данной аксиомы — запрещать цепочки вида $x \in x$, $x \in y \in x$, $x \in y \in z \in x$ и т.д. Действительно, для синглтона $\{x\}$ по аксиоме регулярности существует его элемент y такой, что $y \cap \{x\} = \emptyset$. Но $y = x$ по определению синглтона, следовательно, $x \cap \{x\} = \emptyset$, откуда следует, что $x \notin x$, иначе данное пересечение было бы непустое.

Предположим теперь, что $x \in y \in x$ и применим AR к множеству $\{x, y\}$. Ясно, что $y \cap \{x, y\} = \{x\}$, т. к. $x \in y$ и $y \notin y$. Тогда в силу AR остается единственный элемент пары $\{x, y\}$, не имеющий с ней общих элементов: $x \cap \{x, y\} = \emptyset$. Но тогда $y \notin x$. Аналогично доказывается невозможность любых конечных циклов по отношению принадлежности.

В «начальных» множествах аксиома регулярности выполняется естественным образом — просто в силу того, что любая объектная запись конечна, и, производя процедуру редукции записи, мы за конечное число шагов дойдем до «дна» множества, т. е. спускаясь по цепочке принадлежности (она же является путем в гамма-дереве), рано или поздно мы придем к пустому множеству (листу гамма-деревя). В аксиоматической теории множеств это свойство постулируется с помощью AR.

Аксиомы $AE+AV+AU+AP+AF+AR$ являются аксиомами конечных множеств, а их моделью (с учетом вычислительных ограничений относительно AF) является теория «начальных» множеств, изложенная выше. Мы предлагаем читателю самостоятельно еще раз убедиться в том, что все эти аксиомы выполняются для объектных записей, а значит, реализуются в компьютере (с которого все и началось в нашей книге). | Упражнение 1.16.

Настало время проделать с «начальными» множествами примерно то же самое, что алгебраисты проделывают с полем рациональных чисел — выйти за его пределы с помощью некоего нового понятия, которое станет порождающим элементом для целого класса разнообразных конструкций. В случае поля рациональных чисел мы можем записать уравнение $q^2 = 2$ в языке обычной арифметики, но не можем его разрешить не только в целых, но и в рациональных числах. Для теории множеств существует аналогичный красивый способ поставить вопрос о расширении ее поля объектов, а именно — теорема Гудстейна, которая прекрасно формулируется в арифметике, но для доказательства требует наличия в теории сверхбольших чисел.

1.2.3 Первая бесконечность

Итак, мы добавляем **аксиому бесконечности**:

AI: $\exists u : (\emptyset \in u) \wedge (\forall x : x \in u \rightarrow x \cup \{x\} \in u)$.

Введем обозначение для формулы:

$$\text{Prog}(u) \equiv (\emptyset \in u) \wedge (x \in u \rightarrow x \cup \{x\} \in u).$$

Множество u , удовлетворяющее $\text{Prog}(u)$, будем называть **прогрессивным**. Аксиома бесконечности утверждает, что существует хотя бы одно прогрессивное множество: $\exists u \text{ Prog}(u)$.

В дальнейшем мы по-прежнему будем использовать обозначение $' + 1 '$ для операции $x \cup \{x\}$, т. е. $x + 1 \equiv x \cup \{x\}$.

Очевидно, что прогрессивное множество больше любого «начального», потому что любое начальное множество строится за конечное число шагов из пустого, а прогрессивное множество является результатом неограниченно продолжающегося процесса наращивания (всякий раз, когда в u добавлен элемент x , за ним тут же добавляется элемент $x + 1$, и этот процесс ничем не прерывается).

Таким образом, прогрессивное множество, несмотря на простоту определения и тесную связь с таким хорошо известным нам понятием как рекурсия, принципиально отличается от ранее известных нам «начальных» множеств.

Очевидно также, что прогрессивное множество не имеет объектной записи, т. к. все объектные записи конечны, а прогрессивное множество должно содержать в своей записи как минимум записи всех натуральных чисел, поскольку $0, 0+1, 1+1$ и т.д. являются элементами прогрессивного множества.

Ого, сколько
ж они
тратят
электроэнер-
гии в день? в
час? в
секунду?

Проиллюстрировать прогрессивное множество можно на примере «бесконечной гостиницы Гильберта». Пусть в некоторой гостинице есть номера, соответствующие всем натуральным числам: $0, 1, 2, 3$, и т.д. И пусть все эти номера заняты гостями. Приезжает новый гость и требует заселения. Что делать администрации такой гостиницы? Ответ прост: нужно отправить всем гостям предписание переселиться из номера n в номер $n + 1$. Поскольку для каждого n существует номер $n + 1$, требование выполнимо. При этом освободится номер 0 , в который и заселится новопривывший гость.

Картинка, безусловно, утопическая, однако она хорошо подчеркивает главное отличие бесконечного множества от конечного: бесконечное мно-

Очень
напоминает
фрактал!

жество можно сжать само в себя без потери нумерации элементов. Иначе говоря, бесконечное множество подобно своей собственной части.

Несмотря на то, что прогрессивное множество не имеет своей объектной записи, грамматика не запрещает нам пользоваться переменными для его обозначения и изучать различные его логические свойства.

Заметим, что *потенциальная бесконечность* (т.е. неограниченность модели) существует и в теории «начальных» множеств, поскольку мы можем неограниченно производить все более длинные объектные записи и формулы. Аксиома бесконечности вводит именно *актуальную бесконечность*, т.е. объект теории, являющийся бесконечным, к которому можно применять такие же точно построения и формулы, как и к обычным конечным объектам. Например, мы можем взять некое прогрессивное множество и навешивать на него фигурные скобки так же, как навешиваем их на пустое множество. И получим бесконечно много новых бесконечных множеств.

Теорема 1.8 (Свойства первого бесконечного множества).

- (1) Существует и единственно множество ω такое, что $0 \in \omega, n \in \omega \rightarrow n + 1 \in \omega, n \in \omega \setminus \{0\} \rightarrow \exists k \in \omega : n = k + 1$;
- (2) Если $\varphi(0)$ и $\varphi(n) \rightarrow \varphi(n+1)$, тогда $\varphi(n)$ для всех $n \in \omega$ (арифметическая индукция);
- (3) Если $n \in \omega$, то $n \subset \omega$, и если $n \in t$, то $n \subset t$ (транзитивность);
- (4) $n \in \omega$ тогда и только тогда, когда n — число в смысле грамматики G1–G17 и определения чисел (раздел 1.1.7).

Доказательство. Часть (1). Пусть u — некоторое прогрессивное множество (оно существует по аксиоме AI). Положим:

$$\omega = \{z \in u \mid \forall v : \text{Prog}(v) \rightarrow z \in v\},$$

т. е. ω — пересечение всех прогрессивных множеств. Поскольку ω — часть множества u , то по теореме 1.7 оно является множеством. Нетрудно видеть также, что такое ω единственное, т. к. его определение не зависит от выбранного u .

Кроме того, $\text{Prog}(\omega)$ истинно, поскольку \emptyset принадлежит всем прогрессивным множествам, и если n — элемент всех прогрессивных множеств, то таков же и $n + 1$.

Пусть далее:

$$\alpha = \{n \in \omega \mid n \neq 0 \rightarrow \exists k \in \omega : n = k + 1\}.$$

Ясно, что $\alpha \subseteq \omega$ по построению (одновременно это доказывает существование множества α по теореме 1.7).

С другой стороны, α — прогрессивное множество. Следова-

*Упражнение 1.17.
Докажите $\text{Prog}(\alpha)$*

тельно, по определению ω имеем $\omega \subseteq \alpha$.

Стало быть, $\alpha = \omega$.

Часть (2). Пусть $\alpha = \{n \in \omega \mid \varphi(n)\}$. Ясно, что α — прогрессивное множество, тогда $\omega \subseteq \alpha$. С другой стороны $\alpha \subseteq \omega$, следовательно $\alpha = \omega$, а это и означает, что $\forall n \in \omega \varphi(n)$.

Часть (3), первое утверждение. Для доказательства воспользуемся частью (2), т. е. индукцией. Для $n = 0$ утверждение очевидно, т. к. пустое множество содержится в любом множестве. Пусть для n известно, что $n \in \omega \rightarrow n \subset \omega$. Рассмотрим $n + 1 = n \cup \{n\} \in \omega$. Поскольку $n \in \omega$ (как элемент, предшествующий $n + 1$ — см. часть (1)), имеем $\{n\} \subseteq \omega$. И $n \subset \omega$ по предположению индукции. Отсюда $n \cup \{n\} \subseteq \omega$, при этом $n + 1 \neq \omega$, т. к. $n + 1 \in \omega$ (аксиома AR).

Часть (3), второе утверждение. Докажем индукцией по m следующую формулу: $\varphi(m) \Rightarrow (n \in m) \rightarrow (n \subset m)$. Очевидно, что $\varphi(0)$. Пусть $\varphi(m)$ и $n \in m + 1 = m \cup \{m\}$. Если $n \in m$, то $n \subset m \subset m + 1$, и если $n = m$, то $n \subset m \cup \{m\} = m + 1$. Итак, $\forall n \in m + 1 : n \subset m + 1$.

Часть (4). Необходимость. Пусть $n \in \omega$. Если $n = 0$, то это число по определению. Если n — число, то $n + 1$ — тоже число по определению операции '+1' на числах (раздел 1.1.7). Следовательно, по индукции получаем, что все $n \in \omega$ являются числами грамматики.

Достаточность. Предположим, что a — число, не являющееся обозначением элемента ω . При этом a — набор цифр грамматики. Пусть b означает число такое, что $a = bc$, где c — цифра, т. е. b получается из a удалением последней цифры (используется правило G2 в обратную сторону). Если b обозначает

элемент ω , то необходимо показать, что за конечное число выполнения операции '+1' можно перейти от b к bc . Понятно, что количество таких операций зависит от самого числа b , поскольку $bc - b = 9 \cdot b + c$, но за конечное число шагов действительно можно придти к тому, что вместо b будет $b0$, а дальше за c шагов прийти к bc . Но тогда $a \in \omega$.

Если b также не обозначает элемент ω , повторим для него те же рассуждения, что и для a . Поскольку в числе a конечный набор цифр, рано или поздно мы дойдем до первой цифры в записи a , а проверить, что все цифры являются элементами ω несложно — их можно явно выписать при помощи скобочной записи.

Таким образом, все числа грамматики обозначают элементы ω . \square

Пункт (4) теоремы не совсем укладывается в рамки аксиоматической теории множеств, поскольку основан на правилах грамматики. Однако, после того, как мы определим функции и рекурсивные определения, мы сможем строго формально показать, как можно построить функциональное соответствие между элементами ω и наборами цифр.

Элементы множества ω будем называть *натуральными числами*. По умолчанию переменные n, m, i, j, k, l у нас будут обозначать произвольные натуральные числа.

Упражнение 1.18. В качестве упражнения докажите, что множество ω удовлетворяет аксиомам Пеано PA1, PA2, PA7 из раздела 4.1.5:¹¹

$$P1 : 0 \in \omega;$$

$$P2 : n \in \omega \rightarrow n + 1 \in \omega;$$

$$P3 : n + 1 \neq 0;$$

$$P4 : n, m \in \omega \rightarrow (n + 1 = m + 1) \leftrightarrow (n = m);$$

$$P5 : (X \subseteq \omega) \wedge (0 \in X) \wedge (n \in X \rightarrow n + 1 \in X), \text{ тогда } X = \omega.$$

Постулирование бесконечного множества специального вида напоминает постулирование пустого множества, поскольку сразу же приводит к существованию огромного количества новых множеств, которые можно построить на основе ω .

Так, ничто теперь не мешает нам строить множества вида $\omega_1 = \omega_0 \cup \{\omega_0\}$, $\omega_2 = \omega_0 \cup \{\omega_0, \omega_1\}$ и т. д., где $\omega_0 = \omega$. Затем можно собрать все полученные «омеги» в одно множество (оно существует по аксиоме подстановки и аксиоме бесконечности) ω_ω , которое будет продолжением натурального ряда.

В среде любителей и специалистов по теории множеств известна следующая игра. Собрались два математика и стали придумывать наперегонки

¹¹Аксиомы PA3–PA4 мы пропускаем, поскольку еще не построили операции сложения и умножения.

большие числа, но так, чтобы каждое следующее принципиально отличалось бы от предыдущего. Первый называет 1, второй — 10, первый — число Гугóл, второй — простое число с номером Гугóл, первый — максимальное число последовательности Гудстейна, начавшейся с простого числа с номером Гугóл, второй не выдерживает и называет омегу, первый — континуум, второй — первое слабо недостижимое кардинальное число, первый — строго недостижимое кардинальное число, второй — измеримый по Уламу кардинал. На этом игра закачивается, т. к. громких названий сверхбольших чисел пока никто не определил.

Гугóл = 10^{100}

Другая аналогия, которую нам дает введение множества ω , выводит нас на очередной архетип математики — **архетип порождающего элемента**. Действительно, аксиома бесконечности позволила присоединить к начальным множествам один новый элемент, который тут же породил огромный класс новых множеств. Напоминает известную в алгебре операцию расширения поля с помощью нового элемента. Так мы строим расширения $\mathbb{Q}[\sqrt{2}]$ и тому подобные (см. раздел 3.7.1).

Здесь же можно вспомнить и о семействах множеств, которые порождают топологию или сигма-алгебру измеримых множеств. Обычно в качестве порождающего семейства рассматривают интервалы или круги в зависимости от того, в каком поле (\mathbb{R} или \mathbb{C}) мы находимся (раздел 5.1.1).

В связи последним замечанием уместно рассмотреть очередной архетип математики — **системы множеств**. Общего определения этого понятия не существует, поэтому мы и относим его к архетипам. Но неформально можно определить систему множеств как некоторую определяемую часть булеана $\mathcal{P}(x)$, обладающую рядом определенных свойств. Уже из такого неформального определения видно, что система множеств должна основываться на некотором базисном множестве, которое мы обозначили x . Затем из x выбираются подмножества (элементы $\mathcal{P}(x)$), удовлетворяющие некоторым требованиям. Напомним ряд примеров этого архетипа:

- Топология
- Алгебра множеств и сигма-алгебра
- Фильтры и ультрафильтры

К ним мы еще вернемся позже, а в следующем разделе займемся разработкой инструментария, позволяющего свободно оперировать конструкциями теории множеств для получения всех остальных математических объектов и теорий *внутри теории множеств*.

Приведем теперь полный список аксиом Цермело–Френкеля с использованием упрощающих восприятие обозначений:

$$\mathbf{AV} \quad (a = b) \leftrightarrow \forall x (x \in a \leftrightarrow x \in b);$$

AE $(a = b) \rightarrow \forall c (a \in c \rightarrow b \in c);$

AU $\forall x \neq \emptyset \exists \cup x = \{z \mid \exists y z \in y \in x\};$

AP $\forall x \exists p = \{y \mid y \subseteq x\};$

AF если $\varphi(x, y)$ определяет отображение для всех $x \in a$, то существует $b = \{y \mid \exists x \in a : \varphi(x, y)\};$

AR $\forall x \neq \emptyset \exists y \in x (x \cap y = \emptyset);$

AI $\exists p (\emptyset \in p) \wedge (\forall x \in p : x + 1 \in p).$

Данный список аксиом принято обозначать ZF (по первым буквам отцов-основателей этой теории).

К этим аксиомам можно *опционально* добавлять и некоторые другие аксиомы (например, аксиому выбора (AC) или гипотезу континуума (CH)) или их модификации в различных сочетаниях, но указанные 7 аксиом из теории выбросить просто невозможно. Приведем формулировку аксиомы выбора с использованием сокращений (строгое определение термина «функция» и обозначения $f : a \rightarrow b$ будет дано буквально через 5 страниц):

AC $\forall a \exists f (f : a \rightarrow \cup a) \wedge ((\forall x \neq \emptyset) f(x) \in x).$

Функция f , существование которой гарантирует эта аксиома, называется **функцией выбора**.

Совместны ли аксиомы ZF, а также ZF + AC и другие расширения базового набора? Поскольку, как мы уже видели, в теорию множеств можно без труда поместить арифметику, для которой справедливы теоремы Гёделя, доказать совместность теории множеств невозможно, не выходя за рамки самой теории. Тем не менее, существует ряд расширяющих аксиом (например, аксиома строго недостижимого кардинального числа (SI)), из истинности которых следует совместность теории ZF, что обнадеживает нас, но не доказывает совместность ZF саму по себе. Есть и еще более интересные примеры логических секвенций. Например, справедлива такая

Теорема 1.9. *Из совместности теории ZF невозможно доказать, что аксиому SI невозможно опровергнуть.*

Для более полного знакомства с формальной аксиоматикой теории множеств рекомендуем **видеокурс** И. Пономарева [J_s6DT8ioUc], а также книги [25, 35].

1.3 Основные инструменты

В этом разделе мы вновь включаем в грамматику все правила G1–G20 и прописные латинские буквы различного начертания в качестве объектных переменных. Это даст нам большее разнообразие обозначений, которое в первую очередь послужит более простому запоминанию.

1.3.1 Порядок отношений и отношения порядка

Рассмотрим еще один важный объект теории множеств — упорядоченную пару. Это такой терм $P(a, b)$ с переменными a и b , что $P(a, b) = P(c, d)$ тогда и только тогда, когда $(a = c) \wedge (b = d)$. Конструктивно упорядоченная пара задается (по Куратовскому) следующим образом:

$$(a, b) \equiv \{\{a\}, \{a, b\}\}.$$

При таком определении (a, b) является упорядоченной парой. Понятно, что более простой вариант $\{a, b\}$ не может претендовать на роль упорядоченной пары, поскольку $\{a, b\} = \{b, a\}$. По другой причине на эту роль не годится и $\{a, \{b\}\}$, т. к. в этом случае получим равенство пар $(\{1\}, 0) = (\{0\}, 1)$, хотя при этом $0 \neq 1$ (напомним, что $1 = \{0\}$).

Упражнение 1.19.
Проверьте это, используя аксиому регулярности!

Не исключены и другие способы определения упорядоченной пары множеств, однако данное определение стало общепринятым и самым экономичным в плане использования скобочной символики. Упорядоченная пара — это, пожалуй, первый объект на нашем пути, который является скорее понятием теории множеств, чем какой-то конкретной формулой или объектом, поскольку она может быть реализована различными способами. Но поскольку на упорядоченной паре держится огромное количество конструкций теории множеств, следует все-таки волевым усилием выбрать ее единственную реализацию — по Куратовскому, а все прочие способы игнорировать.

Как недемонстративно!))

Введем собственные обозначения для левой и правой компоненты упорядоченной пары $P = (a, b)$:

$$\text{pr}_1(P) \equiv \cup \cap P, \quad \text{pr}_2(P) \equiv \begin{cases} \cup \cap P, & \text{если } \cup P \setminus \cap P = \emptyset, \\ \cup(\cup P \setminus \cap P), & \text{иначе} \end{cases}$$

Упражнение 1.20.
Проверьте, что $\text{pr}_1((a, b)) = a$ и $\text{pr}_2((a, b)) = b$

Понятно, что, имея упорядоченную пару, можно определять упорядоченные тройки, четверки, энки по правилу $(a, b, c) = (a, (b, c))$ и т.д. Однако при этом сразу же бросается в глаза несимметричность такого определения, а громоздить сюда классы эквивалентных определений совсем уж

избыточно. Поэтому все остальные упорядоченные наборы мы будем определять с помощью функций¹².

Упорядоченная пара открывает нам колоссальные возможности в множестве построении. Введем следующее определение: **прямым** (декартовым) **произведением** множеств A и B называется множество

$$A \times B \equiv \{(x, y) \mid (x \in A) \wedge (y \in B)\}.$$

Упражнение 1.21. Воспользуйтесь тем, что $A \times B \subseteq \mathcal{P}\mathcal{P}(A \cup B)$.

Упражнение: докажите существование и единственность прямого произведения.

Прямое произведение интересно не только само по себе, но и как хороший плацдарм для введения таких понятий как отношение и функция.

В математике отношение между объектами обычно задается формулой, после чего изучаются его свойства. Например, атомарная формула теории множеств $(a = b)$ — это очевидный пример отношения между объектами теории. Формула же позволяет задавать и произвольные n -арные отношения, т.е. между n объектами одновременно. По сути, любая формула $\varphi(x, y)$ с двумя и более свободными переменными x и y определяет отношение.

Но в теории множеств все должно быть множеством! Поэтому для определения понятия отношения вводится его ограниченное определение — отношение на множестве (между множествами).

Подмножество $R \subseteq A \times B$ называется отношением (между) A и B . Если $(x, y) \in R$, то это можно записывать как xRy для краткости. Вместо R часто используются специальные символы, подчеркивающие смысл отношения. Два таких спецсимвола мы используем с момента введения грамматики — это равенство $(=)$ и символ Пеано (\in) , еще один широко известный символ отношения — это знак $<$, и, конечно же, отношение вложения \subseteq . Мы не будем отказывать себе в удовольствии пользоваться общепринятыми значками для обозначения отношений, определяя их при помощи \equiv .

Ясно, что понятие отношения не запрещает нам рассматривать случай $A = B$, когда отношение R действует на каком-то одном множестве, поэтому такой случай мы не выделяем особо.

Отношение R между A и B может иметь следующие свойства, за которыми закреплены специальные термины:

Rel0 Всюду значность: $\forall y \in B \exists x \in A : xRy$;

Rel1 Всюду определенность: $\forall x \in A \exists y \in B : xRy$;

Rel2 Однозначность: $xRy \wedge xRz \rightarrow y = z$;

¹²Заметим — и упорядоченную пару тоже можно задать функцией, но вот функцию без упорядоченной пары уже не построишь.

Rel3 Обратная однозначность: $xRz \wedge yRz \rightarrow x = y$.

Rel4 Антисимметричность: $A = B$ и $(xRy \wedge yRx) \rightarrow (x = y)$;

Rel5 Симметричность: $A = B$ и $xRy \rightarrow yRx$;

Rel6 Рефлексивность: $A = B$ и $\forall x \in a : xRx$;

Rel7 Транзитивность: $A = B$ и $(xRy \wedge yRz) \rightarrow xRz$;

Rel8 Антирефлексивность: $A = B$ и $\forall x \in a : \neg(xRx)$;

Rel9 Связность: $A = B$ и $xRy \vee x = y \vee yRx$;

Перечисленные 10 свойств отношения являются теми кирпичиками, из которых складываются уже привычные нам понятия. На Рис.1.6 показано, как из первых четырех свойств складываются виды функций (стрелки направлены от простого понятия к более сложному). На Рис.1.7 показано, как взаимосвязаны друг с другом различные виды отношений.

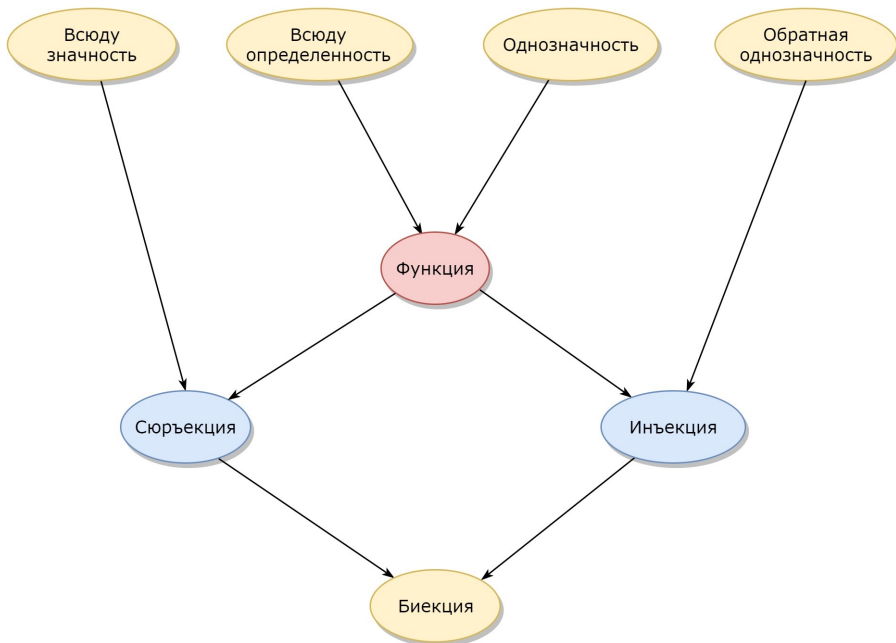


Рис. 1.6: Функциональные свойства отношений

Дадим следующие определения:

F1 Отношение f между A и B называется **функцией** из A в B , если f — всюду определенное и однозначное. Обозначение: $f : A \rightarrow B$;

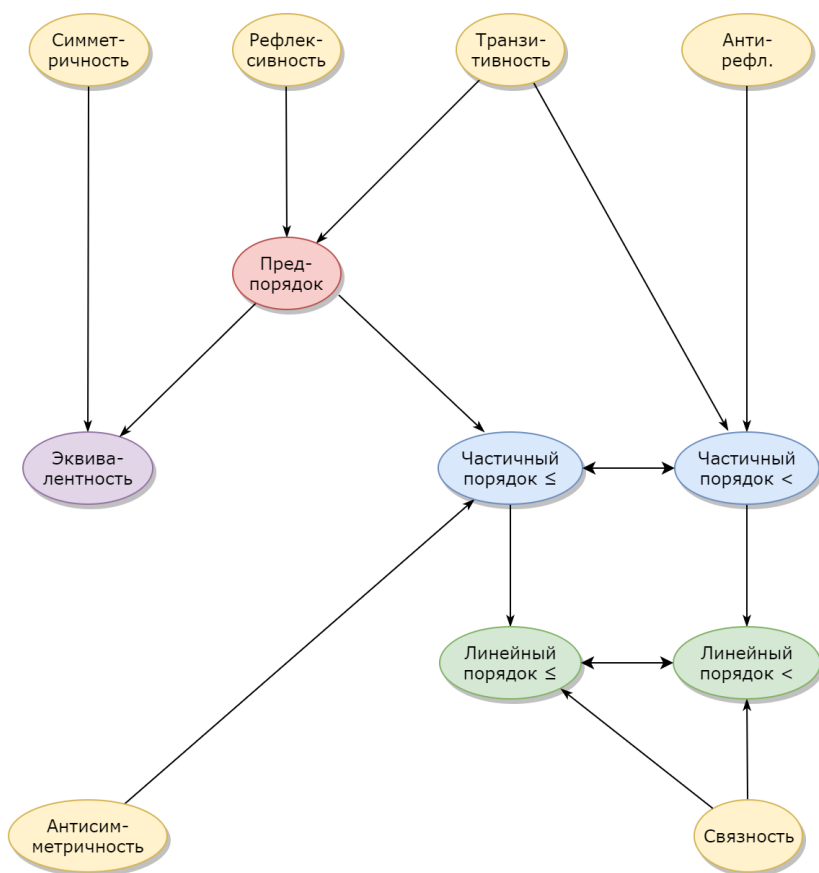


Рис. 1.7: Порядковые свойства отношений

F2 Функция $f : A \rightarrow B$ называется **сюрьекцией** (отображением «на»), если f — всюду значимое отношение. Обозначение: $f : A \xrightarrow{on} B$;

F3 Функция $f : A \rightarrow B$ называется **инъекцией** (отображением «в»), если f — обратно однозначное отношение. Обозначение: $f : A \xrightarrow{in} B$;

F4 Функция $f : A \rightarrow B$ называется **биекцией** (взаимно однозначным соответствием), если она является инъекцией и сюрьекцией. Обозначение: $f : A \leftrightarrow B$.

Утверждение $(x, y) \in f$ для функции f принято записывать через равенство $f(x) = y$.

Более строго функциональный терм $f(x)$ определяется так: | Напомним, что $\bigcup\{a\} = a$.

$$f(x) \Rightarrow \bigcup\{y \mid (x, y) \in f\},$$

множество, стоящее справа, имеет один элемент (в силу однозначности f), так что формула $f(x) = y$ эквивалентна $(x, y) \in f$. Определения, данные выше для функции $f : A \rightarrow B$ предполагают, что A является областью определения f в силу свойства всюду определенности. Но если нам дали функцию, не указав ее область определения и область значений, то как определить их? Для этого вводятся следующие обозначения:

$$\begin{aligned} \text{Pr}_1(R) &\Rightarrow \{x \mid \exists y (x, y) \in R\} = \{\text{pr}_1(p) \mid p \in R\} \text{ (левая проекция отношения)} \\ \text{Pr}_2(R) &\Rightarrow \{y \mid \exists x (x, y) \in R\} = \{\text{pr}_2(p) \mid p \in R\} \text{ (правая проекция отношения)} \\ \text{dom}(f) &\Rightarrow \text{Pr}_1(f) \text{ (область определения)} \\ \text{ran}(f) &\Rightarrow \text{Pr}_2(f) \text{ (область значений)} \end{aligned}$$

Таким образом, зная о множестве только то, что это отношение или функция, можно в рамках аксиоматики определить их проекции и области определения и значений. Это удобно в тех случаях, когда обозначение для функции или отношения возникает без обозначений образующих множеств.

Функциональный терм $f(x)$ в ряде случаев позволяет упростить терм-квантор. Положим:

$$\{f(x) \mid \varphi(x)\} \Rightarrow \{y \mid (y = f(x)) \wedge \varphi(x)\}$$

В частности, $\text{ran}(f) = \{f(x) \mid x \in \text{dom}(f)\}$ и образ множества A относительно функции f и прообраз множества B :

$$fA \Rightarrow \{f(x) \mid x \in A\}, \quad f^{-1}B = \{x \mid f(x) \in B\}.$$

Аналогично, по схеме 1.7 мы предлагаем читателю самостоятельно дать определения различных отношений. Предполагается, что здесь мы рассматриваем отношение на множестве (т. е. случай $A = B$), хотя строго формально ничто не мешает все эти свойства применять и к отношениям между разными множествами. Подчеркнем, что мы одновременно указали два вида порядков — строгий и нестрогий ($<$ и \leq), которые определяются друг через друга согласно следующему свойству.

Теорема 1.10.

(1) Если отношение $<$ является строгим частичным порядком (антирефлексивное + транзитивное), то $\leq (a < b \vee a = b)$ является нестрогим частичным порядком (рефлексивное + транзитивное + антисимметричное);

- (2) Если отношение \leq является нестрогим частичным порядком, то отношение $< (a \leq b \wedge a \neq b)$ является строгим частичным порядком.
- (3) Если отношение $<$ является строгим линейным порядком (строгий частичный порядок + связность), то отношение $\leq (a < b \vee a = b)$ является нестрогим линейным порядком (нестрогий частичный порядок + связность);
- (4) Если отношение \leq является нестрогим линейным порядком, то отношение $< (a \leq b \wedge a \neq b)$ является строгим линейным порядком.

Доказательство. (1) Пусть отношение $a < b$ является строгим частичным порядком (антирефлексивное, транзитивное). Определим отношение $a \leq b$ формулой

$$(a \leq b) \Rightarrow (a < b) \vee (a = b)$$

и проверим свойства такого отношения.

- 1) Оно рефлексивно по определению: $a \leq a$.
- 2) Оно транзитивно, т. к. если $(a \leq b) \wedge (b \leq c)$, то $((a < b) \vee (a = b)) \wedge (b < c) \vee (b = c)$. Во всех 4 случаях выполняется $(a < c) \vee (a = c)$ в силу транзитивности $<$.
- 3) Оно антисимметрично. Действительно, предположим, что $(a \leq b) \wedge (b \leq a)$, но при этом $a \neq b$. Тогда $(a < b) \wedge (b < a)$, и в силу транзитивности $<$ имеем $a < a$, что противоречит антирефлексивности $<$.

(2) Пусть теперь отношение $a \leq b$ является нестрогим частичным порядком (рефлексивное, транзитивное, антисимметричное). Определим отношение $a < b$ формулой

$$(a < b) \Rightarrow (a \leq b) \wedge (a \neq b)$$

и проверим свойства такого отношения:

- 1) Оно антирефлексивно по определению.
- 2) Оно транзитивно. Действительно, пусть $(a < b) \wedge (b < c)$, т. е. $(a \leq b) \wedge (b \leq c) \wedge (a \neq b) \wedge (b \neq c)$. Тогда $a \leq c$ в силу транзитивности \leq . Предположим, что $a = c$, тогда в силу антисимметричности \leq и условия $(a \leq b) \wedge (b \leq c)$ получаем $a = b$, но это противоречит $a < b$. Таким образом, $(a \leq c) \wedge (a \neq c)$, т. е. $a < c$.

(3) и (4) следуют из (1) и (2), а также из того, что связность для строгого и нестрогого порядков определяется одинаково. □

Данная теорема позволяет нам достаточно свободно пользоваться обозначениями строгого или нестрогого порядков, не уточняя каждый раз его свойства.

Если на множестве A задан (частичный или линейный) порядок $<$, то пару $(A, <)$ принято называть (частично или линейно) **упорядоченным множеством**. Часто этот термин применяется и к самому множеству-носителю, т. е. к A , если из контекста понятно, о каком отношении идет речь.

На стыке функциональных и порядковых свойств отношений можно задать еще одно интересное порядково-функциональное отношение — это цикл или, в общем случае, перестановка. Перестановкой можно назвать любую биекцию на множестве A , но с помощью свойства антирефлексивности можно избавиться от вырожденных циклов, т. е. петель. В комбинаторике хорошо известны перестановки конечных множеств и их разложение на циклы.

Например, запись $(125)(3)(4)$ обозначает биекцию, состоящую из одного невырожденного цикла $1 \rightarrow 2 \rightarrow 5 \rightarrow 1$ и двух петель $3 \rightarrow 3$ и $4 \rightarrow 4$.

Нестрогий частичный порядок показывает нам разницу между порядком и предпорядком. Дело в том, что предпорядок допускает симметрию, т. е. aRb и bRa одновременно. В некоторых случаях симметрия даже становится неизбежной. Например, если для треугольника abc задать циклический порядок aRb , bRc , cRa и потребовать, чтобы это был предпорядок, то в силу транзитивности мы сразу получим обратные отношения aRc , bRa , cRb . По сути, оно тут же превратится в отношение эквивалентности.

Отношение эквивалентности приводит нас к понятию разбиения. Пусть на A задано отношение эквивалентности \sim . Для любого элемента $a \in A$ существует единственное множество $[a]_{\sim} = \{x \in A \mid a \sim x\}$, которое называется **классом эквивалентности** элемента a , а множество всех классов эквивалентности элементов A , обозначаемое

$$A/\sim \equiv \{[a]_{\sim} \mid a \in A\},$$

называется **фактор-множеством** множества A по отношению эквивалентности \sim . Ясно, что $A/\sim \subseteq \mathcal{P}(A)$.

Разбиением A называется такое подмножество $D \subseteq \mathcal{P}(A)$, что: $\cup D = A$, $\forall x \in D : x \neq \emptyset$ и $\forall x, y \in D : x \cap y = \emptyset$.

Теорема 1.11. *Фактор-множество является разбиением множества A . Любое разбиение A задает единственное отношение эквивалентности на A .*

Упражнение: докажите теорему.

Упражнение 1.23.

Отношение эквивалентности продолжает поднятую нами тему о равенстве и принадлежности. Оно действует более грубо, чем равенство, поскольку отождествляет не только равные друг другу множества. Таким об-

разом, на данный момент мы располагаем следующими видами «равенства»:

$$\equiv, \quad \equiv\equiv, \quad =, \quad \sim$$

Первое у нас обозначало посимвольное тождество объектных записей, второе — равенство без учета порядка элементов (равенство мультимножеств), третье — равенство без учета порядка и количества копий элементов (обычное равенство множеств), четвертое — отношение эквивалентности произвольной природы.

Напрашивается отношение, которое обозначало бы «равенство», действующее как отношение эквивалентности, но при этом учитывало бы некие дополнительные свойства множества. Такое отношение множеств имеет много определений, но объединяется одним общим термином **изоморфизм**. Изоморфизм (или *структурное подобие*) — очень широкое понятие, отражающее сохранение структуры на множествах при их преобразованиях. Например, изоморфизм групп сохраняет операцию, изоморфизм линейных пространств сохраняет сложение векторов и умножение их на число, изоморфизм графов сохраняет инцидентность вершин и направление дуг. Наконец, изоморфизм множеств с отношением сохраняет отношение.

Распространенность этого понятия (а также смежных с ним — гомоморфизм, гомеоморфизм, автоморфизм), пожалуй, заслуживает того, чтобы внести его в наш перечень архетипов. Итак, **изоморфизм — это архетип, отвечающий за структурное подобие**.

Вообще, факторизация, порожденная каким-либо изоморфизмом или произвольным разбиением, занимает очень важное место в математике, поэтому здесь уместно говорить также об **архетипе факторизации**. Наиболее активно этот архетип проявляется в алгебре, топологии, теории графов.

Пусть на множестве A задано отношение R_A , на множестве B задано отношение R_B . Тогда пары (A, R_A) и (B, R_B) называются **изоморфными**, если существует биекция $f : A \leftrightarrow B$ такая, что $(xR_A y) \leftrightarrow (f(x)R_B f(y))$, где $x, y \in A$. Обозначение: $(A, R_A) \cong (B, R_B)$. Часто пишут просто $A \cong B$, если это не приводит к неоднозначному толкованию записи, хотя правильное было бы писать $R_A \cong R_B$, т. к. изоморфизм отвечает именно за соответствие элементов отношений. Можно даже построить биекцию $F(a, b) = (f(a), f(b))$ между R_A и R_B , которая будет сохранять пары. Поэтому в дальнейшем любой изоморфизм, сохраняющий отношение, мы будем называть **изоморфизмом отношений**.

Упражнение 1.24. Нетрудно видеть, что изоморфизм отношений, как формула на языке ZF, является отношением эквивалентности на отношениях.

В частности, можно поставить вопрос о том, какие линейные порядки изоморфны.

Чтобы сформулировать следующую теорему, введем ряд определений. Для множества A с частичным порядком $<$:

1. Элемент $a \in A$ называется **максимальным**, если $\forall x \in A : \neg(a < x)$.
2. Элемент $b \in A$ называется **минимальным**, если $\forall x \in A : \neg(x < b)$.
3. Элемент $a \in A$ называется **наибольшим**, если $\forall x \in A : (x \leq a)$, обозначение: $\max(A) \Rightarrow \cup\{a \in A \mid \forall x \in A : (x \leq a)\}$.
4. Элемент $b \in A$ называется **наименьшим**, если $\forall x \in A : (b \leq x)$, обозначение: $\min(A) \Rightarrow \cup\{b \in A \mid \forall x \in A : (b \leq x)\}$.
5. A **неограничено сверху**, если в A нет максимального элемента.
6. A **неограничено снизу**, если в A нет минимального элемента.

Наибольший и наименьший элементы единственны, если существуют, поэтому множества $\{a \in A \mid \forall x \in A : (x \leq a)\}$ и $\{b \in A \mid \forall x \in A : (b \leq x)\}$, стоящие в их определении либо, пусты, либо состоят из 1 элемента, а операция $\cup\{a\}$ для синглтона дает сам элемент a . В дальнейшем мы не будем так подробно расписывать обозначения, предлагая читателю самому пройти этот тернистый путь формализации.

В линейном порядке понятия минимального и наименьшего элемента совпадают, аналогично — понятия максимального и наибольшего элемента совпадают. Соответственно, линейно упорядоченное множество неограничено сверху, если в нем нет наибольшего элемента, неограничено снизу — если в нем нет наименьшего элемента.

Пусть $X \subseteq A$, тогда:

7. $a \in A$ является **верхней гранью** X , если $\forall x \in X : x \leq a$. При этом X **ограничено сверху в A** .
8. $b \in A$ является **нижней гранью** X , если $\forall x \in X : b \leq x$. При этом X **ограничено снизу в A** .
9. Точная верхняя грань $\sup(X)$ — это наименьшая верхняя грань, т.е.: $\sup(X) = \min\{a \in A \mid \forall x \in X : x \leq a\}$.
10. Точная нижняя грань $\inf(X)$ — это наибольшая нижняя грань, т.е.: $\inf(X) = \max\{a \in A \mid \forall x \in X : a \leq x\}$.

Если существует $\max(X)$, то он совпадает с $\sup(X)$, и если существует $\min(X)$, то он совпадает с $\inf(X)$.

Приведем простые примеры:

- 1) Множество $\mathcal{P}(A)$ ($A \neq \emptyset$) частично упорядочено отношением \subseteq , имеет наибольший элемент A и наименьший \emptyset , оно ограничено в себе как сверху, так и снизу.

- 2) В множестве натуральных чисел ω с естественным линейным порядком существует единственный минимальный элемент, он же наименьший — 0. Но не существует ни максимального элемента, ни наибольшего. ω неограничено сверху, но если его рассматривать как часть $\omega + 1$, то оно будет ограничено самим собой, т. к. для всех $n \in \omega$ имеем $n < \omega$.
- 3) Пусть $n <_d m$ если n делит m , и рассмотрим $A = \omega \setminus \{0, 1\}$. В этом случае минимальными элементами будут все простые числа, а наименьший элемент не существует. A не ограничено ни сверху, ни снизу. Однако оно будет ограничено снизу в множестве ω , поскольку 1 делит все натуральные числа, включая ноль.
- 4) Разбиения (и отношения эквивалентности) можно сравнивать между собой. Пусть X и Y — два разбиения множества A . Скажем, что $X \leq Y$ (X мельче Y), если для всякого $x \in X$ существует $y \in Y$ такое, что $x \subseteq y$. Ясно, что такое отношение («отношение масштаба») является частичным порядком на множестве всех разбиений A .

Для любых двух разбиений X и Y множества A можно найти их общие верхнюю и нижнюю грани. Так, легко видеть, что

$$\inf\{X, Y\} = \{x \cap y \mid x \in X, y \in Y\}.$$

Упражнение | Упражнение: докажите, что это действительно инфимум (точная нижняя грань).
1.28.

Сложнее дело обстоит с точной верхней гранью. Обозначим $\mathcal{D}(X, Y)$ множество всех верхних граней для $\{X, Y\}$ (оно непусто, т. к. тривиальное разбиение $\{A\}$ является верхней гранью). Пусть также для каждого $a \in A$ и каждого разбиения $D \in \mathcal{D}(X, Y)$ задана функция выбора

$$f_a(D) = \cup\{d \in D \mid a \in d\}.$$

Эта функция выбирает единственный элемент $d \in D$, которому принадлежит a (его единственность гарантируется тем, что D есть разбиение A). Тогда

$$\sup\{X, Y\} = \left\{ \bigcap_{D \in \mathcal{D}(X, Y)} f_a(D) \mid a \in A \right\}. \quad (1.10)$$

Предлагаем читателю самостоятельно доказать, что а) это действительно разбиение и б) оно является минимальным из всех разбиений в $\mathcal{D}(X, Y)$, т. е. $\sup\{X, Y\} = \min \mathcal{D}(X, Y)$. Заметим, что указанная конструкция годится для получения \inf любого множества разбиений, но в данном случае \inf еще и является минимумом $\mathcal{D}(X, Y)$.

Множество A назовем **конечным**, если существует $n \in \omega$ и биекция $f : A \leftrightarrow n$. Множество A назовем **счетным**, если существует биекция $f : A \leftrightarrow \omega$.

Пусть $(A, <)$ — линейно упорядоченное множество и $B \subseteq A$. Множество B **плотно в A** , если любой *непустой* интервал¹³ $(x; y) \equiv \{z \mid x < z < y\}$, где $x, y \in A$, пересекается с B ($B \cap (x; y) \neq \emptyset$). Другими словами, B плотно в A , если между любыми двумя точками A найдется точка из B . Например, \mathbb{Q} плотно в \mathbb{A} и в \mathbb{R} .

A **плотно**, если оно плотно в себе, т.е. между любыми двумя точками A найдется третья точка из A . $\mathbb{Q}, \mathbb{A}, \mathbb{R}$ — плотные (при обычном линейном порядке).

Теорема 1.12. *Все счетные неограниченные сверху и снизу плотные линейно упорядоченные множества порядково изоморфны.*

Поясним. Пусть у нас имеется два множества A и B , которые счетны (т.е. все их элементы можно перенумеровать натуральными числами), на них заданы линейные порядки $<_A$ и $<_B$ такие, что в обоих множествах нет ни наибольшего, ни наименьшего элемента, и эти множества плотны, тогда существует изоморфизм $f : A \leftrightarrow B$, сохраняющий порядок.

Доказательство этой теоремы можно посмотреть в [11]. Из этой теоремы следует, например, что множество \mathbb{Q} с обычным линейным порядком и множество \mathbb{A} всех алгебраических чисел с обычным линейным порядком порядково изоморфны.

Пусть $(L, <)$ — линейно упорядоченное множество и $A, B \subset L$, причем $A \neq \emptyset$, $B \neq \emptyset$, $A \cap B = \emptyset$, $A \cup B = L$, $A \leqslant B$ ¹⁴. Тогда пара (A, B) называется **сечением** л.у.м. L , множество A — нижним классом сечения, B — верхним классом сечения.

Линейный порядок $<$ на множестве L называется **непрерывным** (множество L с таким порядком непрерывно), если каково бы ни было его сечение, либо в нижнем классе сечения существует наибольший элемент, а в верхнем нет наименьшего, либо в верхнем классе существует наименьший элемент, а в нижнем нет наибольшего (такие сечения называются **дедекиндовыми**).

Теорема 1.13. *Следующие утверждения эквивалентны:*

- (1) Л.у.м. $(L, <)$ непрерывно;
- (2) $(L, <)$ плотно и всякое непустое ограниченное сверху множество в L

¹³Мы обозначаем интервал метасимволом, очень похожим на обозначение упорядоченной пары: отличие состоит лишь в разделителе. Нужно очень аккуратно пользоваться такой символикой, если вы имеете дело с машиной. Французы предлагают экзотическое для русской школы обозначение $]x, y[$, но вполне логичное, если вспомнить наше обозначение замкнутых отрезков.

¹⁴Здесь и далее сравнение множеств означает сравнение их элементов с квантором всеобщности: $X \leqslant Y$ ($X < Y$) означает, что $\forall x \in X \forall y \in Y : x \leqslant y$ ($x < y$). То же относится к сравнению множества и элемента: $c \leqslant Y$.

имеет точную верхнюю грань;

(3) $(L, <)$ плотно и всякое непустое ограниченное снизу множество в L имеет точную нижнюю грань.

Доказательство. Покажем, что непрерывное множество плотно в себе. Пусть $x < y$. Предположим, что $(x; y) \cap L = \emptyset$, рассмотрим сечение (A, B) такое, что $A = \{z \mid z \leq x\}$, $B = \{z \mid z \geq y\}$. Такое сечение не является дедекиндовым, поскольку в нем верхний и нижний классы имеют свой наименьший и наибольший элемент, соответственно. Это противоречит определению непрерывности. Следовательно, любой интервал $(x; y)$ содержит точки L , т. е. L плотно в себе.

(1) \Rightarrow (2). Пусть непустое X ограничено сверху, т. е. $\exists y \in L : X \leq y$. Пусть Y — множество всех верхних граней X . Положим $A = L \setminus Y$, $B = Y$. Тогда (A, B) — сечение L и $X \leq B$, причем в A нет наибольшего элемента (если бы таковой был, он был бы также верхней гранью X , а значит, лежал бы в $Y = B$). Из непрерывности L получаем, что тогда $\exists \min B$, т. е. $\exists \sup X$.

(2) \Rightarrow (3). Пусть непустое Y ограничено снизу, т. е. $\exists x \in L : x \leq Y$. Пусть X — множество всех нижних граней Y . Очевидно, что X ограничено сверху, и пусть $c = \sup(X)$ (он существует в силу (2)). Нетрудно видеть, что тогда $c = \inf(Y)$.

(3) \Rightarrow (1). Пусть (A, B) — сечение L . В силу (3) $\exists c = \inf(B)$. Ясно, что $A \leq c$. Если $c \in A$, то в A есть наибольший элемент, а в B нет наименьшего. Если $c \in B$, то в B есть наименьший элемент. Допустим, что в A есть наибольший элемент $a = \max(A)$. Тогда $a < c$, а интервал $(a; c)$ не пересекается с L (все элементы $l \in L$ либо $l \leq a$, либо $l \geq c$), а это противоречит требованию плотности L . Значит, в A нет наибольшего элемента. Таким образом, (A, B) является дедекиндовым сечением, т. е. L непрерывно. \square

Отметим, что плотное л.у.м. не всегда непрерывно. Например, \mathbb{Q} плотно, но не непрерывно (сечение для $\sqrt{2}$ не дедекиндово).

Линейно упорядоченное множество $(L, <)$ называется **вполне** (или полностью) **упорядоченным**, если любое его непустое подмножество имеет наименьший элемент.

Упражнение 1.29. Нетрудно видеть, что любое вполне упорядоченное множество не непрерывно.

Теорема 1.14. Множество ω вполне упорядочено отношением \in .

Данную теорему можно доказать непосредственно по определению ω с помощью теоремы 1.8 и аксиомы регулярности AR, а также как очевидное следствие более общей теоремы об ординалах (см. ниже теорему 1.16).

Добавим еще несколько слов об отношениях на множествах. Пусть R — отношение между A и B .

11. **Обратное к R отношение** определяется так: $R^{-1} = \{(y, x) \mid (x, y) \in R\}$, оно является подмножеством $B \times A$.

Если $<$ — частичный порядок, то обратное к нему отношение обозначается $>$, аналогично: \leq и \geq обратны друг другу. Отношение, обратное частичному порядку, является частичным порядком, к линейному — линейным. Обратное отношение к полному порядку, вообще говоря, не является полным порядком, поскольку все минимальные элементы при этом становятся максимальными. Обратное само себе отношение ($R = R^{-1}$) — это симметричное отношение. В частности, отношение эквивалентности само себе обратно. Если отношение f^{-1} для функции $f : A \rightarrow B$ является функцией, то говорят, что функция f обратима, а f^{-1} называют обратной к f функцией. При этом запись $f^{-1}Y$ означает прообраз множества Y :

$$f^{-1}Y = \{x \mid f(x) \in Y\}.$$

Все отношения, которые мы рассмотрели, можно разделить на 3 вида: отношения порядка, отношение эквивалентности и функции. Первые два вида действуют на одном множестве, третье — на двух (или между двумя). От двух к трем и более множествам можно перейти, например, с помощью прямого произведения, но по сути это все равно будет отношение на новом одном множестве. Тем не менее, существует еще один весьма условный тип отношений — инцидентность. Он основан как раз на комбинировании разнородных объектов и представляет собой подмножество в прямом произведении (двух и более множеств). Опираясь на данные выше определения, скажем, что инцидентность — это отношение между множествами, не обязательно являющееся функцией.

Суть инцидентности состоит в том, что мы рассматриваем связи разнородных объектов друг с другом, чтобы в дальнейшем изучать сохранение инцидентности при преобразованиях исходных множеств.

Например, для многогранника вводится понятие **флаг**. Это конструкция, которая состоит из одной грани H , одного ребра E и одной вершины V , причем вершина V является вершиной ребра E , а ребро E является ребром грани H . После чего мы начинаем отображать многогранник сам на себя различными способами и смотреть, какие его преобразования сохраняют инцидентность компонентов флага — вершины, ребра и грани.

Второй пример: графы. В графе каждому ребру инцидентны две вершины. Если у нас имеется изоморфизм графов (биекции между ребрами и между вершинами), то он предполагает сохранение инцидентности.

Аналогично можно рассматривать в геометрии инцидентные прямую и точку, плоскость и прямую, кривую и касательные и т.д., и смотреть, при каких преобразованиях эта инцидентность сохраняется.

1.3.2 Трансфинитная рекурсия

Мы уже выше сталкивались со свойством транзитивности элементов множества (теорема 1.8). Положим теперь:

$$\text{Trans}(x) \equiv \forall y \in x : y \subset x.$$

Из теоремы 1.8, например, следует, что множество ω и все натуральные числа транзитивны.

Далее, множество α называется **ординалом**, если оно транзитивно и всего его элементы транзитивны:

$$\text{Ord}(\alpha) \equiv \text{Trans}(\alpha) \wedge (\forall \beta \in \alpha : \text{Trans}(\beta)).$$

Ординалы принято обозначать буквами греческого алфавита (обычно начальными: $\alpha, \beta, \gamma, \delta$), которые были введены в нашу грамматику правилом ГЗ.

Рассмотрим ряд свойств ординалов.

Ord1: ω и все натуральные числа — ординалы;

Ord2: $\text{Ord}(\alpha) \rightarrow \text{Ord}(\alpha + 1)$;

Ord3: $\text{Ord}(\alpha) \rightarrow \forall \beta \in \alpha : \text{Ord}(\beta)$;

Ord4: Если X — множество ординалов, то $\text{Ord}(\cup X) \wedge \text{Ord}(\cap X)$

Упражнение 1.30. | Предлагаем читателю самостоятельно разобраться в доказательстве этих свойств.

Следующая теорема является обобщением арифметической индукции, доказанной нами в теореме 1.8.

Теорема 1.15 (Трансфинитная индукция). Пусть $\varphi(x)$ — формула языка ZF, x — свободная переменная. Если для любого ординала β

$$(\forall \alpha \in \beta : \varphi(\alpha)) \rightarrow \varphi(\beta), \quad (1.11)$$

то для любого ординала β истинно $\varphi(\beta)$.

Доказательство. Предположим, что для некоторого ординала $\neg \varphi(\beta)$, и положим

$$X \equiv \{\alpha \in \beta \mid \neg \varphi(\alpha)\}.$$

Если $X = 0$, то $\forall \alpha \in \beta : \varphi(\alpha)$, и посылка (1.11) индукции дает нам $\varphi(\beta)$.

Если $X \neq 0$, то по аксиоме AR $\exists \alpha \in X : \alpha \cap X = 0$. Пусть $\gamma \in \alpha$. Поскольку β транзитивно, $\alpha \subset \beta$ и, следовательно, $\gamma \in \beta$. Но при этом $\gamma \notin X$ (т. к. $\alpha \cap X = 0$), откуда $\varphi(\gamma)$. Таким образом, $\forall \gamma \in \alpha : \varphi(\gamma)$, откуда в силу посылки (1.11) получаем, что $\varphi(\alpha)$, а это противоречит тому, что $\alpha \in X$. \square

Теорема 1.16. *Ординалы вполне упорядочены отношением \in .*

Доказательство. Нужно показать, что (1) \in антирефлексивно, (2) \in транзитивно, (3) \in связно и (4) любое непустое множество ординалов содержит наименьший элемент по отношению \in .

(1) выполняется в силу аксиомы регулярности AR.

(2) выполняется в силу транзитивности ординалов (по определению).

Покажем (3), а именно, что для любых ординалов α, β имеет место связность отношения \in :

$$\varphi(\alpha, \beta) \Rightarrow \text{Ord}(\alpha) \wedge \text{Ord}(\beta) \rightarrow (\alpha \in \beta) \vee (\alpha = \beta) \vee (\beta \in \alpha). \quad (1.12)$$

Для доказательства этого факта нам потребуется провести двойную индукцию — по α и β одновременно.

Индукция по α : требуется показать, что если $\forall \gamma \in \alpha : \varphi(\gamma, \beta)$, то $\varphi(\alpha, \beta)$.

Пусть выполнено предположение индукции:

$$\forall \gamma \in \alpha : \varphi(\gamma, \beta) \quad (1.13)$$

Доказательство $\varphi(\alpha, \beta)$ также поведем индукцией, но уже по параметру β . Для этого требуется показать, что если $\forall \delta \in \beta : \varphi(\alpha, \delta)$, то $\varphi(\alpha, \beta)$. Пусть выполнено предположение вложенной индукции:

$$\forall \delta \in \beta : \varphi(\alpha, \delta) \quad (1.14)$$

Для любых множеств, в том числе ординалов, выполняется свойство трихотомии: $(\alpha = \beta) \vee (\alpha \setminus \beta \neq 0) \vee (\beta \setminus \alpha \neq 0)$.

1) если $\alpha = \beta$, то очевидно $\varphi(\alpha, \beta)$.

2) если $\alpha \setminus \beta \neq 0$, то существует $\gamma \in \alpha \setminus \beta$, в частности $\gamma \in \alpha$, поэтому в силу предположения (1.13) возможны три случая:

2.1) $\gamma \in \beta$ — невозможно в силу $\gamma \in \alpha \setminus \beta$;

2.2) $\gamma = \beta$, тогда $\beta \in \alpha$ и, следовательно, $\varphi(\alpha, \beta)$;

2.3) $\beta \in \gamma$, тогда $\beta \in \alpha$, т. к. $\gamma \in \alpha \wedge \text{Trans}(\alpha)$, откуда получаем $\varphi(\alpha, \beta)$.

3) если $\beta \setminus \alpha \neq 0$, то существует $\delta \in \beta \setminus \alpha$, в частности $\delta \in \beta$, поэтому в силу предположения (1.14) возможны три случая:

3.1) $\delta \in \alpha$ — невозможно в силу $\delta \in \beta \setminus \alpha$;

3.2) $\delta = \alpha$, тогда $\alpha \in \beta$ и, следовательно, $\varphi(\alpha, \beta)$;

3.3) $\alpha \in \delta$, тогда $\alpha \in \beta$, т. к. $\delta \in \beta \wedge \text{Trans}(\beta)$, откуда получаем $\varphi(\alpha, \beta)$.

Итак, из предположения (1.14) следует $\varphi(\alpha, \beta)$, и вложенная индукция завершается. Тогда из предположения (1.13) следует $\varphi(\alpha, \beta)$, и основная индукция также завершена.

Таким образом, отношение \in связно на ординалах и, следовательно, линейно упорядочивает их.

Покажем (4). Пусть X — непустое множество ординалов. По аксиоме регулярности AR существует ординал $\alpha \in X$ такой, что $\alpha \cap X = 0$. Пусть $\beta \in X$, тогда в силу связности \in на ординалах $(\alpha \in \beta) \vee (\alpha = \beta) \vee (\beta \in \alpha)$. Если $\beta \in \alpha$, то $\alpha \cap X \neq 0$ — противоречит выбору α . Следовательно, либо $\alpha = \beta$, либо $\alpha \in \beta$, откуда следует, что α — наименьший в X .

Таким образом, отношение принадлежности вполне упорядочивает все ординалы. \square

Следствие 1.3.

ω и натуральные числа вполне упорядочены отношением \in .

В силу доказанной теоремы отношение \in на ординалах принято обозначать символом ' $<$ '. Кроме того, еще одно отношение на ординалах отождествляется с принадлежностью. Справедлива

Теорема 1.17. Если α и β — ординалы, то $\alpha \in \beta \leftrightarrow \alpha \subset \beta$.

Доказательство. Импликация $\alpha \in \beta \rightarrow \alpha \subset \beta$ очевидна в силу транзитивности ординалов. Пусть $\alpha \subset \beta$, и предположим, что $\neg(\alpha \in \beta)$, тогда либо $\beta \in \alpha$, либо $\beta = \alpha$ — в обоих случаях имеем $\beta \subseteq \alpha$. Противоречие. \square

Дадим следующее определение: ординал α называется **последующим**, если $\alpha = \beta + 1$ при некотором ординале β ; ординал α называется **предельным**, если $\alpha > 0$ и α — не последующий.

Упражнение 1.31. Приведем еще ряд свойств ординалов, которые мы оставим в качестве упражнения для самостоятельного развития навыков обращения с этими сущностями:

Ord5: Если X — множество ординалов, то $\sup(X) = \cup X$, $\min(X) = \cap X$;

Ord6: Если $n > 0$, то n — последующий ординал; ω — предельный;

Ord7: α — последующий тогда и только тогда, когда $\alpha = \sup(\alpha) + 1$;

Ord8: α — предельный тогда и только тогда, когда $\alpha = \sup(\alpha)$;

Ord9: Не существует наибольшего ординала;

Ord10: Не существует множества всех ординалов.

Следующая теорема дает нам очень мощный инструмент построения объектов в теории множеств.

Архетип
мощности?

Теорема 1.18 (Трансфинитная рекурсия). Пусть формула $\varphi(x, y)$ определяет отображение, т. е. $\varphi(x, y) \wedge \varphi(x, z) \rightarrow y = z$ для любых x, y, z . Тогда для любого ординала α существует единственная функция h_α , определенная на α и такая, что для всех $\beta < \alpha$:

$$h_\alpha(\beta) = \cup\{y \mid x = \{h(\gamma) \mid \gamma < \beta\} \wedge \varphi(x, y)\} \quad (1.15)$$

Доказательство. Сначала введем более удобные обозначения. Определим функциональный терм

$$f(x) \doteq \cup\{y \mid \varphi(x, y)\}.$$

Очевидно, что $y = f(x)$ тогда и только тогда, когда $\varphi(x, y)$. Далее, если задана функция $g : A \rightarrow B$ и $a \subseteq A$, то

$$ga \doteq \{y \in B \mid \exists x \in a : y = g(x)\}$$

называется образом множества a относительно функции g .

В этих обозначениях равенство (1.15) записывается так:

$$h_\alpha(\beta) = f(h_\alpha\beta), \quad \beta < \alpha. \quad (1.16)$$

Требуется доказать существование и единственность такой рекурсивной функции.

(1) Докажем единственность. Пусть при некотором α существует h_α и h'_α , удовлетворяющие (1.16). Предположим, что они различны в точке $\beta < \alpha$. Пусть далее

$$X \doteq \{\gamma \leq \beta \mid h_\alpha(\gamma) \neq h'_\alpha(\gamma)\},$$

и пусть $\gamma_0 = \min(X)$. Тогда

$$h_\alpha(\gamma_0) = f(h_\alpha\gamma_0) = f\{h_\alpha(\delta) \mid \delta < \gamma_0\} = f\{h'_\alpha(\delta) \mid \delta < \gamma_0\} = h'_\alpha(\gamma_0),$$

противоречие. Единственность доказана.

(2) Докажем существование. Предположим, что для некоторого ординала не существует функция, удовлетворяющая (1.16), и обозначим через α наименьший из таких ординалов.

(2.1) Пусть α — предельный ординал. И пусть $H_\alpha = \{h_\beta \mid \beta < \alpha\}$. Такое множество существует по аксиоме переноса AF, т. к. в силу единственности h_β для каждого β сопоставление $\beta \mapsto h_\beta$ однозначно.

Положим $h_\alpha = \cup H_\alpha$. Нетрудно видеть, что для $\gamma < \alpha$ и любых двух ординалов β, β' таких, что $\gamma < \beta < \beta' < \alpha$ (они существуют в силу предельности α) имеем равенство $h_\beta(\gamma) = h'_{\beta'}(\gamma)$ (доказательство аналогично части (1)). Так что h_α — функция.

Кроме того, $h_\alpha(\gamma) = h_\beta(\gamma)$ при любом $\beta \in (\gamma; \alpha)^{15}$, откуда следует, что h_α удовлетворяет равенству (1.16):

$$h_\alpha(\gamma) = h_\beta(\gamma) = f(h_\beta\gamma) = f(h_\alpha\gamma).$$

(2.2) Пусть $\alpha = \beta + 1$. Положим $h_\alpha(\gamma) = h_\beta(\gamma)$, если $\gamma < \beta$ и $h_\alpha(\beta) = f(h_\beta\beta)$. Нетрудно видеть, что и в этом случае h_α удовлетворяет (1.16). \square

Заметим, что если у нас определены рекурсивные функции h_α для любого ординала α , то мы имеем и определяемое отображение h , заданное на всех ординалах и удовлетворяющее той же самой рекурсивной формуле $h(\alpha) = f(h\alpha)$. Такое отображение можно задать функциональным термом $h(\alpha) = f(h_\alpha\alpha)$.

Частным случаем трансфинитной рекурсии является счетная рекурсия, определенная для натуральных чисел. Например, рекурсию $x_0 = a$ и $x_{n+1} = r(x_n)$ можно определить следующим образом:

$$\begin{aligned} f(\{\}) &= (0, a), \\ f(X) &= \cup\{(\alpha + 1, r(x)) \mid ((\alpha, x) \in X) \wedge (\alpha = \sup\{\beta \mid (\beta, x) \in X\})\}, \end{aligned}$$

Рассмотрим функцию h_ω . Ясно, что $h_\omega(0) = f(\{\}) = (0, x_0)$, $h_\omega(1) = f(\{(0, x_0)\}) = \cup\{(1, r(x_0))\} = (1, x_1)$, и т.д. В общем случае имеем

$$h_\omega(n + 1) = f(\{(0, x_0), (1, x_1), \dots, (n, x_n)\}) = \cup\{(n + 1, r(x_n))\} = (n + 1, x_{n+1}).$$

Таким образом, h_ω задает последовательность пар вида (n, x_n) , откуда нетрудно получить саму последовательность x_n . Аналогично можно задать счетную рекурсию, где каждое следующее значение зависит от нескольких или всех предыдущих.

Наконец, рассмотрим связь ординалов и порядков.

Теорема 1.19. *Если $(X, <)$ — вполне упорядоченное множество, то существует единственный порядково изоморфный ему ординал.*

Доказательство. Рассмотрим функцию $f : \mathcal{P}(X) \rightarrow X$ такую, что

$$f(A) = \min(X \setminus A), \quad f(X) = \min(X).$$

На основе данной функции для каждого ординала α построим рекурсивную функцию $h_\alpha(\beta) = f(h_\alpha\beta)$.

1) Если h_α — сюръекция и h_β — не сюръекция при всех $\beta < \alpha$, то h_α — монотонная биекция. Действительно, т. к. h_β — не сюръекция, то $X \setminus h_\alpha\beta \neq \emptyset$ при любом $\beta < \alpha$, откуда следует, что если $\gamma < \beta < \alpha$, то $h_\alpha(\gamma) < h_\alpha(\beta)$.

¹⁵Не путаем пару (a, b) и интервал $(a; b)$!

2) Существует ординал α , при котором h_α — сюръекция. Действительно, если бы это было не так, мы бы получили инъективное определяемое отображение со всех ординалов на подмножество множества X , но тогда обратное ему отображение было бы сюръекцией с множества на все ординалы, и по аксиоме АФ существовало бы множество ординалов, что неверно, как было показано выше.

3) из 1) и 2) следует, что существует монотонная биекция между X и некоторым ординалом α , т. е. $(X, <) \cong \alpha$. \square

Следствие 1.4. *Если два ординала порядково изоморфны, то они равны.*

Ординал, единственный изоморфный вполне упорядоченному множеству $(X, <)$, называется **порядковым типом** $(X, <)$. Порядковый тип обозначается функциональным термом $|(X, <)|$ или $\text{Ord}(X, <)$, иногда пользуются упрощенным вариантом $|X|$ или $\text{Ord}(X)$, если понятно из контекста, о каком именно отношении порядка на X идет речь. Для ординалов по умолчанию всегда предполагается порядок, заданный отношением \in , который всегда обозначается $<$ и опускается в обозначении ординалов как упорядоченных множеств. Так что $|\alpha| = \alpha$.

Не путать с модулем и мощностью!!!

Изоморфизм между вполне упорядоченными множествами и ординалами позволяет применять трансфинитную индукцию и рекурсию на любых вполне упорядоченных множествах.

Постепенно мы начинаем наблюдать тот факт, что чем дальше мы уходим от формализма аксиоматики, тем чаще мы начинаем пользоваться различными значками, значение которых зависит от контекста. Это можно сравнить с локальными обозначениями для переменных и функций, которые вводит программист при написании фрагмента кода, и которые работают только в рамках данного фрагмента, а за его пределами могут либо иметь иной смысл (как, например, знак модуля $|\cdot|$ в анализе), либо вообще не иметь смысла (как, например, скобочные записи множеств). Тем не менее, мы вынуждены увеличивать количество обозначений, иначе наш математический текст станет похож на машинный код, который невозможно будет воспринимать осмысленно, а места будет занимать в десятки и сотни раз больше. Удивительное свойство человека компактифицировать (упаковывать) информацию как нельзя лучше проявляется в математике и программировании.

Но вернемся к нашим инструментам. Последний инструмент, который мы здесь рассмотрим, — это прямое произведение произвольного набора множеств.

Рассмотрим совокупность $X = \{X_\lambda \mid \lambda \in \Lambda\}$ множеств, проиндексирован-

ных ординалами $\lambda \in \Lambda$, где Λ — множество ординалов.¹⁶ Формально это означает, что имеется функция $F : \Lambda \rightarrow X$ и $F(\lambda) = X_\lambda$. По аксиоме AF X — множество. Тогда терм

$$\prod_{\lambda \in \Lambda} X_\lambda \equiv \{f \mid (f : \Lambda \rightarrow \cup X) \wedge (f(\lambda) \in X_\lambda)\}$$

обозначает **прямое произведение** всех X_λ .

То есть, $\prod_{\lambda \in \Lambda} X_\lambda$ состоит из всех функций, выбирающих для каждого ординала λ какое-то значение из X_λ .

Отметим, что существование таких функций в общем случае гарантирует аксиома выбора, которую мы еще не рассматривали. Это одна из таких

Странно, что нет упражнения для самостоятельного изучения аксиомы выбора :)

аксиом теории множеств, которая дает возможность доказать ряд интереснейших и полезных теорем, но истинность которой невозможно установить в ZF, поэтому можно как принимать ее, так и отвергать. Что касается теории начальных множеств, то там мы имеем дело только с конечными объектами, и для них алгоритм построения функций выбора можно считать вполне определенным — можно указать, каким способом за конечное число шагов выбрать элемент из множества, если оперировать их скобочными записями или гамма-деревьями. Поэтому отношение математиков к данной аксиоме ухудшается по мере роста величины множеств, к которым она применяется: для конечных объектов — ОК, для счетных — вроде бы, можно применять, для несчетных — остерегайтесь.

Ниже мы еще вернемся к аксиоме выбора и ее следствиям (см. раздел 4.2).

Предполагая, что $\prod_{\lambda \in \Lambda} X_\lambda$ не пусто, можно заметить, что его элементы — это кортежи (или векторы) длины Λ , компонентами которых являются элементы множеств X_λ . Если предположить, что все X_λ попарно различны между собой, то $\prod_{\lambda \in \Lambda} X_\lambda$ означает прямое произведение элементов множества X , которое можно обозначить короче: $\prod X$. Правда, при этом предполагается, что элементы X вполне упорядочены некоторым отношением.

Если бы мы имели дело с мультимножествами, где допускается повторимость элементов, то $\prod_{\lambda \in \Lambda} X_\lambda$ было бы произведением элементов мультимножества (с учетом кратностей), пронумерованных некоторым ординалом Λ «вперемежку» (т.е. не обязательно кратные элементы стояли бы в этом произведении подряд).

Прямое произведение — одна из таких конструкций теории множеств, где

¹⁶Произвольное множество ординалов вполне упорядочено отношением \in , а значит, имеет единственный порядковый тип, являющийся ординалом. Поэтому, не ограничивая общности, можно считать, что Λ есть ординал.

очевидные, казалось бы, вещи перестают работать. Например, ассоциативность произведения не работает в буквальном смысле:

$$\Pi\{A, B, C\} \neq \Pi(\{\Pi\{A, B\}, C\}) \neq \Pi(\{A, \Pi\{B, C\}\}),$$

поскольку в первом случае элемент произведения будет иметь вид (a, b, c) , во втором — $((a, b), c)$, в третьем — $(a, (b, c))$.

Конечно, мы можем без особого труда построить биекцию между всеми этими произведениями, причем если на одной из них задана какая-то конструкция вроде порядка или топологии, то мы ее также без труда изоморфно перенесем на две другие, но «осадок останется» — вроде бы равные по сути объекты на деле получаются всего лишь изоморфными. Это своего рода плата за логическую строгость и за то, чтобы иметь возможность подвести под математику общий логический базис.

*Упражнение 1.32.
Постройте эту биекцию.*

Рассмотрим еще один частный случай прямого произведения — степень. Пусть все $X_\lambda = A$. Тогда обозначим

$$A^\Lambda \equiv \prod_{\lambda \in \Lambda} X_\lambda$$

Нетрудно видеть, что A^Λ попросту обозначает множество всех функций $f : \Lambda \rightarrow A$.

И совсем частный случай — $\Lambda = n$, тогда все функции $f : n \rightarrow A$ имеют вид $f = \{(0, a_0), (1, a_1), \dots, (n-1, a_{n-1})\}$, такие функции принято называть кортежами или векторами и обозначать проще:

$$(a_0, \dots, a_{n-1}) \equiv \{(0, a_0), (1, a_1), \dots, (n-1, a_{n-1})\}.$$

Заметим, что здесь мы используем внеграмматический знак многоточие, и по правилам нам следовало бы писать $\{(k, a_k) \mid 0 \leq k < n\}$. Условимся использовать многоточие тогда, когда терм с многоточием можно заменить грамматически корректным кванторным термом, в котором индексирующая переменная пробегает некоторое конечное линейно упорядоченное множество (например, натуральное число) или из построения формулы, определяющей терм-квантор, очевиден алгоритм перебора всех элементов данного множества.

Здесь нам приходится констатировать очередной досадный факт: дело в том, что прямое произведение $A \times A$ и A^2 (где $2 = \{0, 1\}$) — это по сути одно и то же множество, однако же $A \times A$ определяется как набор упорядоченных пар вида (a_0, a_1) , где $a_0, a_1 \in A$, а A^2 — это набор функций, каждая из которых сама по себе является набором упорядоченных пар $f = \{(0, a_0), (1, a_1)\}$, что, вообще говоря, совсем не одно и то же. Но поскольку между такими

конструкциями можно построить отношение эквивалентности, очень близкое к равенству (т. е. такое, которое будет сохранять все свойства, кроме тех, которые основаны на способе построения элементов), мы не будем каждый раз уточнять природу множества A^2 , считая ее ясной из контекста.

То же самое можно сказать о множестве A^1 , вместо которого ничто не мешает нам рассматривать исходное множество A .

Упражнение 1.33. В качестве упражнения проверьте следующие свойства: $A^0 = 1$, $0^0 = 1$, $0^A = 0$, где A — любое непустое множество.

Для элементов A^ω также существуют привычные обозначения, поскольку это не что иное как последовательности в множестве A . $f : \omega \rightarrow A$ можно записать как $\{a_n\}_{n=0}^\infty$ или еще проще $\{a_n\}$, предполагая, что это недописанный до конца терм-квантор.

Так как терм A^Λ обозначает не только прямое произведение, но и просто множество всех функций из Λ в A , то принято в общем случае обозначать через A^B множество всех функций вида $f : B \rightarrow A$ (здесь B не обязано быть ординалом).

Отсюда легко перейти к множеству 2^A , которое представляет собой множество всех функций, определенных на A и принимающих значения 0 или 1. Эти функции принято называть индикаторными и обозначать 1_B , где B — прообраз 1 относительно данной функции, т. е. $1_B(x) = 1$ тогда и только тогда, когда $x \in B$, а для $x \in A \setminus B$ $1_B(x) = 0$. Каждой функции 1_B взаимно однозначно соответствует множество B , поэтому между 2^A и $\mathcal{P}(A)$ существует естественная биекция. Возможно, поэтому булеан еще называют степенным множеством.

Ну, наконец-то! Рассмотрим теперь один конкретный пример. Пусть A — некоторое конечное множество, т. е. существует натуральное n такое, что $A = \{a_0, \dots, a_{n-1}\}$.

Упражнение 1.34. Рассмотрим множество всех конечных произведений $\{A^k \mid k \in \omega\}$. И далее пусть A^* обозначает $\cup \{A^k \mid k \in \omega\}$.

Докажите, что это множество. Ясно, что элементы A^* — это все возможные кортежи (b_1, \dots, b_k) , где $b_i \in A$, при $k = 0$ элемент равен \emptyset .

Множество A^* называется **звездой Клин** (или замыканием Клини). Если считать, что $A = A^1$, то $A \subseteq A^*$.

Таким образом, если A — алфавит в некоторой грамматике (точнее, его теоретико-множественный образ), то A^* — это множество всех слов над этим алфавитом.

В частности, если $C = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, то C^* обозначает все записи нашей грамматики, которые построены по правилам G1–G2, т. е. Числа.

Теперь, чтобы окончательно обосновать пункт (4) теоремы 1.8, нам осталось построить рекурсивное определение для обозначений натуральных чисел.

Итак, нам нужно построить функцию $f : C^* \rightarrow \omega$, которая позволяла бы по Числу-записи однозначно устанавливать соответствующее ему натуральное

число-множество, и наоборот, по числу-множеству определять все возможные его записи.

Для начала определим операцию склейки кортежей:

$$(a_1, \dots, a_n) \bowtie (b_1, \dots, b_m) = (a_1, \dots, a_n, b_1, \dots, b_m)$$

Формально, пусть за первый кортеж отвечает функция $a = \{(0, a_1), (1, a_1), \dots, (n-1, a_n)\}$, за второй — $b = \{(0, b_1), (1, b_2), \dots, (m-1, b_m)\}$.

Построим функцию c по правилу: $c(n+k) = b(k)$, т.е. $c = \{(n, b_1), (n+1, b_2), \dots, (n+m-1, b_m)\}$. Тогда нужный нам кортеж-склейка — это сумма $a \cup c$.

Далее, построим рекурсивную функцию $h : \omega \rightarrow C^*$ по следующим правилам. Пусть предшествующее m число $m-1$ соответствует кортежу $c = (c_1, \dots, c_n)$, т.е. $h(m-1) = c$, если $m > 0$, тогда положим:

$$h(m) = \begin{cases} (0), & \text{если } m = 0 & (h0) \\ (c_1 + 1), & \text{если } n = 1, c_1 \in \{0, \dots, 8\} & (h1) \\ (0, 1), & \text{если } n = 1, c_1 = 9 & (h2) \\ (c_1 + 1, c_2, \dots, c_n), & \text{если } n > 1, c_1 \in \{0, \dots, 8\} & (h3) \\ (0) \bowtie h(h^{-1}(c_2, \dots, c_n) + 1), & \text{если } n > 1, c_1 = 9 & (h4) \end{cases}$$

Можно доказать по индукции, что (c_0, c_1, \dots, c_n) и (c_1, \dots, c_n) принадлежат образу $hm = \{h(k) \mid k < m\}$, и тогда h задана корректно как рекурсия. Приведем несколько примеров значений h :

$$\begin{aligned} h(0) &= (0), & \text{правило (h0)} \\ h(1) &= (1), & c = (0), \text{ правило (h1)} \\ h(2) &= (2), & c = (1), \text{ правило (h1)} \\ h(9) &= (9), & c = (8), \text{ правило (h1)} \\ h(10) &= (0, 1), & c = (9), \text{ правило (h2)} \\ h(11) &= (1, 1), & c = (0, 1), \text{ правило (h3)} \\ h(20) &= (0) \bowtie h(h^{-1}((1)) + 1) = (0, 2), & c = (9, 1), \text{ правило (h4)} \\ h(100) &= (0) \bowtie h(h^{-1}((9)) + 1) = (0, 0, 1), & c = (9, 9), \text{ правило (h4)} \end{aligned}$$

Рекурсивная функция h каждому числу $n \in \omega$ ставит в соответствие кортеж из цифр, записанных в обратном порядке (меньшие разряды стоят левее). При этом она не покрывает все возможные кортежи из C^* , т.к. среди них есть кортежи вида (c_1, \dots, c_n) , где $c_n = 0$. Чтобы построить обещанную выше функцию $f : C^* \rightarrow \omega$, определим сначала операцию инверсии:

$$\overleftarrow{(c_1, \dots, c_{n-1}, c_n)} = (c_n, c_{n-1}, \dots, c_1),$$

формально, $\overleftarrow{c}(k) = c_{n-k+1}$.

Далее,

$$f(c) = \begin{cases} 0, & \text{если } c = 0 \\ f(c'), & \text{если } c = (0) \bowtie c', c' \neq 0 \\ h^{-1}(\overleftarrow{c}), & \text{иначе} \end{cases}$$

Например, $f((0, 0, 1, 0, 0)) = f((0, 1, 0, 0)) = f((1, 0, 0)) = h^{-1}((0, 0, 1)) = 100$. Итак, функция f каждому кортежу из C^* (включая единственный элемент множества $C^0 = \{0\}$) ставит в соответствие число из ω , причем это соответствие взаимно однозначное с точностью до начальных нулей в записи, которые отбрасываются.

Таким образом, мы полностью алгоритмизировали сопоставление наборов цифр (грамматических Чисел) и натуральных чисел аксиоматической теории множеств.

И главную роль в этом сыграла рекурсия, а также ординал ω — первый бесконечный ординал.

Кроме вполне упорядочения на ординалах, как и на любых множествах, можно рассматривать еще одно отношение. Это **отношение равномощности**. Назовем два множества равномощными ($A \sim B$), если между ними существует биекция. Легко проверить, что отношение равномощности является отношением эквивалентности на множествах, но в общем случае мы не сможем выделить классы эквивалентности по этому отношению, т. к. они не будут множествами. Например, все множества вида $\{\alpha\}$, где α — ординал, равномощны, но их класс эквивалентности будет заведомо больше, чем совокупность всех ординалов, которая множеством не является.

Теорема 1.20 (Кантора). *Любое множество X неравномощно $\mathcal{P}(X)$.*

Доказательство. Предположим, что это не так и что существует биекция $f : \mathcal{P}(X) \leftrightarrow X$. Пусть

$$Y = \{x \in X \mid x \notin f^{-1}(x)\}.$$

Вопрос: $f(Y) \in Y$ или нет?

Пусть $f(Y) = y \in Y$, тогда по определению Y имеем: $y \notin f^{-1}(y)$, откуда $y \notin Y$.

Пусть $y \notin Y = f^{-1}(y)$, тогда по определению Y имеем $y \in Y$.

В обоих случаях мы получили противоречие. □

Теорема Кантора говорит нам о том, что каким бы большим ни было множество X , всегда можно найти еще большее — его булеан. В частности, ω неравномощно 2^ω . Последнее множество интересно тем, что оно равномощно множеству всех действительных чисел \mathbb{R} , то есть *континуум несчетен*.

Эта теорема демонстрирует нам часто употребляемый в математике **архетип трансцендентного объективизма**, с помощью которого строятся контрпримеры к каким-либо неконструктивным или слишком общим определениям. Это, пожалуй, первый методологический архетип, который мы встречаем. Его суть состоит в том, что мы задаем некоторое множество, предполагая, что оно включает все описываемые объекты, а затем приходим к противоречию, тем самым показывая, что построить такое множество нельзя. Иначе говоря, определение объектов получается трансцендентным в том смысле, что их генеральная совокупность не определяется конструктивными методами.



*Георг
Фердинанд
Кантор*

В теореме Кантора мы видим, что множество X «не в силах» перечислить все свои подмножества, и в этом смысле $\mathcal{P}(X)$ является трансцендентной надстройкой над X . Другой пример — парадокс лжеца. Утверждение «все лгут» трансцендентно в том смысле, что, кто бы из людей его ни произнес, он не может быть ни прав, ни неправ. Еще пример — парадокс Расселла в наивной теории множеств: множество $\{x \mid x \notin x\}$ существует и не существует одновременно, т. к. оно не может ни принадлежать себе, ни не принадлежать.

*Доктор
Хаус? :)*

Частный случай работы архетипа трансцендентного объективизма доставляет диагональное доказательство в любой его форме. Например, доказательство Кантора о несчетности интервала $(0; 1)$.

Тот же архетип работает и при доказательстве бесконечности простых чисел, когда в предположении об их конечности строится новое число, которое вновь оказывается простым, хотя оно больше любого простого из данного конечного набора. При данном предположении это число оказывается и трансцендентным и объективно существующим одновременно.

Архетип показывает глубокую пропасть между бытовым и строго математическим подходом к предметной области. Действительно, в бытовой логике слово «все» как правило означает все мыслимые в данный момент объекты, причем мыслимые субъектом, который это высказывание формирует. Так, в парадоксе лжеца тот, кто говорит «все лгут» (субъект), не имеет ввиду себя, а только тех, кого он воспринимает как объекты своего утверждения. Поэтому парадокс нивелируется.

В отношении же математических объектов бытовая логика предполагает, что само пространство этих объектов переменчиво, оно может модифицироваться субъектом исследования тогда, когда он «строит» новые множества. Поэтому построение множества Расселла является всего лишь динамическим расширением понятия «множество», и потому оно никак не может принадлежать самому себе, поскольку его «еще не было» до момента его «сотворения».

В то же время, математический подход к предметной области предполагает, что все объекты уже заданы аксиоматикой (т. е. они объективны), и любое

построение есть не что иное как нахождение нужного объекта среди имеющихся, но никак не создание того, чего раньше не было. Это отличие мешает многим воспринять и доказательство теоремы Кантора, и доказательство несчетности континуума, и доказательство бесконечности простых чисел.

Разница в «размерах» множеств (в смысле биективного соответствия) наводит на мысль о том, что можно построить шкалу мощностей так же, как мы построили шкалу полных порядков — ординалы. И часть этой шкалы нам известна — все натуральные числа попарно неравномощны и образуют шкалу конечных мощностей.

Определение: ординал τ называется **кардиналом**, если $\forall \alpha < \tau : \alpha \not\sim \tau$.

Теорема 1.21. *Для любого ординала α существует равномощный ему кардинал.*

Доказательство. Пусть $X = \{\beta \leq \alpha \mid \alpha \sim \beta\}$ и $\tau = \min(X)$. Очевидно, что τ — кардинал, равномощный ординалу α . \square

Если между множеством X и кардиналом τ существует биекция, то мы полагаем по определению $\|X\| = \tau$, где функциональный терм $\|X\|$ называется **мощностью множества** X . Мощность множества, если существует, то определяется однозначно. Заметим также, что если существует $\|X\|$, то множество X можно вполне упорядочить — достаточно перенести порядок с кардинала (являющегося ординалом) при помощи той же биекции.

Итак, кардиналы — это ординалы специального вида, упорядочение которых является самым компактным среди всех равномощных ординалов. Приведем несколько свойств кардиналов, которые мы по традиции оставляем читателю на самостоятельное изучение:

Card1 Ординал ω и все натуральные числа являются кардиналами;

Card2 Бесконечный кардинал является предельным ординалом;

Card3 Ординалы $\omega + 1$, $\omega + 1 + 1$, $\omega + 1 + 1 + 1$ и т.д. не являются кардиналами;

Card4 Мощность множества определяется однозначно и $\|\tau\| = \tau$, где τ — кардинал;

Card5 Если $\text{Ord}(\alpha)$ и $\tau = \|\alpha\|$, то $\tau \leq \alpha$;

Card6 Множество X можно вполне упорядочить тогда и только тогда, когда у него существует мощность;

Card7 Если существует $\|\mathcal{P}(X)\|$, то существует $\|X\|$ и $\|X\| < \|\mathcal{P}(X)\|$;

Card8 Если существует инъекция $f : A \rightarrow B$ и существуют мощности A и B , то $\|A\| \leq \|B\|$;

Card9 Если X — множество кардиналов, то $\cup X$ — кардинал и $\sup X = \cup X$.

Пусть τ — кардинал и пусть существует наименьший кардинал среди кардиналов $\chi > \tau$. Такой кардинал обозначим τ^+ . Отметим, что существование τ^+ для любого τ не выводится в теории ZF. Если $\tau < \omega$, то, очевидно, $\tau^+ = \tau + 1$.

Запахло аксиомой выбора!

Бесконечные кардиналы принято нумеровать с помощью ординалов:

$$\aleph_0 = \omega, \aleph_1 = \omega^+, \aleph_2 = \aleph_1^+, \dots, \aleph_\omega = \sup_{k < \omega} \aleph_k$$

Больше того, можно построить рекурсивное определение для «алефов», удовлетворяющее условиям:

$$\begin{aligned} \aleph_0 &= \omega \\ \aleph_{\alpha+1} &= \aleph_\alpha^+ \\ \aleph_\alpha &= \sup_{\beta < \alpha} \aleph_\beta, \text{ если } \alpha \text{ — предельный ординал,} \end{aligned}$$

правда, оно будет обрываться на таком ординале α , для которого кардинал \aleph_α^+ не существует. Вопрос о существовании последующих кардиналов мы отложим до разбирательств с аксиомой выбора.

Последнее, что мы рассмотрим в этой части книги, среди основных инструментов, — это теорема о равномощности.

Теорема 1.22 (Кантора–Бернштейна–Шредера). *Если $f : A \rightarrow B$ и $g : B \rightarrow A$ — инъекции, то множества A и B равномощны.*

Доказательство этой теоремы больше техническое, чем концептуальное, поэтому мы приведем здесь лишь его идею¹⁷. Возьмем функцию $H : \mathcal{P}(A) \rightarrow \mathcal{P}(A)$ такую, что:

$$H(X) = A \setminus g[B \setminus fX], \quad X \subseteq A,$$

где $fX = \{y \mid \exists x \in X : y = f(x)\}$ — образ X , $g[Y] = \{x \mid \exists y \in Y : x = g(y)\}$ — образ Y .

Требуется показать, что существует корень уравнения $H(X) = X$. В этом случае две инъекции f и g действуют одна на X , а вторая на $Y = B \setminus fX$ так, что их образы и прообразы не пересекаются, но покрывают целиком множества A и B . Точнее, $X \cap gY = \emptyset$, $X \cup gY = A$ и $fX \cap Y = \emptyset$, $fX \cup Y = B$. Соответственно, функция $f|_X \cup g^{-1}|_{A \setminus X}$ — искомая биекция между A и B . \square

Отметим, что ординалы в теории ZF играют примерно такую же роль, как натуральные числа в теории «начальных» множеств. Во-первых, это вполне упорядоченная шкала множеств, во-вторых, она ничем не ограничена сверху

¹⁷Подробнее см., например, в [11, 25]

(натуральные числа без аксиомы бесконечности обладают таким же свойством), в-третьих, актуализация собрания всех ординалов как единого объекта теории потребует выхода за пределы аксиоматики ZF (актуализация ω потребовала ввести аксиому бесконечности). Отличие состоит в том, что новой аксиомой здесь уже не поможешь — чтобы ввести в теорию новый объект «класс всех ординалов», придется ввести новое понятие: *класс множеств*.

Классы множеств вводит аксиоматика Гёделя–Бернайса [24]. В такой теории грамматика оперирует двумя видами объектов — множества и классы, причем всякое множество есть класс, но не всякий класс есть множество. Собственно классы — это сверхбольшие совокупности множеств, например, класс всех множеств, класс ординалов и т.д. При этом в теории с классами по-прежнему мы не встречаем парадокс Расселла, просто потому, что класса всех классов не существует. Чтобы ввести совокупность

*Еще лучше
— башню
видов
надклассов,
перенумерован-
ных видами
элементами
 ω ...*

всех классов, потребуется сделать еще один шаг — ввести суперклассы как очередной вид объектов.

В этой книге мы не будем углубляться в такого рода построения, поскольку обычной теории множеств нам вполне достаточно для описания большинства математических конструкций, но иногда мы будем использовать понятие класса тех или иных множеств для упрощения нашего повествования.

1.3.3 Перечень инструментов

Приведем список рассмотренных в данном разделе книги основных инструментов:

1. Упорядоченная пара
2. Прямое произведение двух множеств
3. Отношения:
 - (a) частичный порядок
 - (b) линейный порядок
 - (c) полный порядок
 - (d) эквивалентность, класс эквивалентности
4. Функции:
 - (a) инъекция
 - (b) сюръекция
 - (c) биекция
 - (d) обратная функция

5. Изоморфизм (порядковый)
6. Ординал
 - (a) последующий ординал
 - (b) предельный ординал
7. Трансфинитная индукция
8. Трансфинитная рекурсия
9. Порядковый тип
10. Прямое произведение (обобщенное)
 - (a) степень множества
 - (b) звезда Клини
11. Кардинал
12. Мощность множества

1.4 Универсумы и мультимножества

Выше мы определили ординалы, которые образуют стержень в классе всех множеств, позволяющий их измерять: сравнить их размер (мощность) и порядки на них.

*Скромно
молчим про
аксиому
выбора*

На этот стержень мы хотим теперь нанизать множества специального вида, которые были бы максимально широкими, т.е. охватывали бы все возможные множества определенного размера. Такие множества называются **универсальными** или, проще, **универсумами**. Мы с ними уже имели дело в разделе 1.1.9 в отношении «начальных» множеств. Настало время ввести аналогичное понятие в аксиоматической теории множеств. Дадим следующее рекурсивное определение:

$$V1 : V_0 = \emptyset;$$

$$V2 : V_{\alpha+1} = \mathcal{P}(V_\alpha);$$

$$V3 : V_\alpha = \cup \{V_\beta \mid \beta < \alpha\}, \text{ если } \alpha \text{ — предельный ординал.}$$

Нетрудно видеть, что при $\alpha = n \in \omega$ множество V_α совпадает с n -ым универсумом до-2 кратности, который был определен в разделе 1.1.9. При этом сохраняется формула $V_n = \mathcal{P}^n(\emptyset)$.

Соответственно, $V_\omega = \cup \{\mathcal{P}^n(\emptyset) \mid n \in \omega\}$ — универсум всех «начальных» множеств.

Несколько простых свойств универсумов:

V4 : Если α — предельный ординал, то $\text{Prog}(V_\alpha)$;

V5 : Всякий универсум транзитивен: $\text{Trans}(V_\alpha)$;

V6 : Если $\alpha < \beta$, то $V_\alpha \subset V_\beta$ (V_α строго монотонно по α);

V7 : Всякое множество является элементом какого-то универсума.

Упражнение 1.35. | Свойства V4–V6 мы оставляем в качестве упражнения читателю, а свойство V7 оформим в виде теоремы.

Теорема 1.23. *Для любого множества x существует ординал α такой, что $x \in V_\alpha$.*

Доказательство. Предположим, что это не так и пусть x такое, что для любого α : $x \notin V_\alpha$. Построим рекурсивно функцию f :

$$f(0) = x, \quad f(n+1) = \cup f(n), \quad n < \omega.$$

Метод бесконечного спуска — см. стр. 244.

Положим далее $T(x) = \cup \{f(n) \mid n < \omega\}$ ¹⁸.

Множество $T(x)$ транзитивно. Пусть далее

$$M = \{z \in T(x) \cup \{x\} \mid \forall \alpha \text{ Ord}(\alpha) \rightarrow z \notin V_\alpha\}.$$

Очевидно, что $M \neq \emptyset$. Пусть $z \in M$ такое, что $z \cap M = \emptyset$ (аксиома AR). Тогда для любого $y \in z$ существует α такой, что $y \in V_\alpha$ (в противном случае $y \in z \in T(x) \rightarrow y \in T(x) \rightarrow y \in M \rightarrow z \cap M \neq \emptyset$).

Построим отображение:

$$r(y) = \min\{\alpha \mid y \in V_\alpha\}.$$

Ясно, что $y \in V_{r(y)}$. Далее положим $\gamma = \sup\{r(y) \mid y \in z\}$. Очевидно, что для всех $y \in z$ имеем $y \in V_\gamma$, т. е. $z \subseteq V_\gamma$. Откуда следует, что $z \in V_{\gamma+1}$, а значит, $z \notin M$. Противоречие. \square

Как следствие из этой теоремы, мы можем сразу же определить ранг множества:

$$\text{rank}(x) = \min\{\alpha \mid x \subseteq V_\alpha\}$$

и добавим еще несколько простых свойств универсумов:

V8 : $\text{rank}(\alpha) = \alpha$ для любого ординала α ;

V9 : $x \in V_\alpha$ тогда и только тогда, когда $\text{rank}(x) < \alpha$;

V10 : $\|V_n\| = 2 \uparrow \uparrow (n-1)$, $n < \omega$.

Упражнение 1.36.
Докажите V8–V10.

¹⁸ $T(x)$ называют **транзитивным замыканием** x .

Последние два свойства повторяют утверждения (1) и (2) теоремы 1.2.

Напомним, что в случае «начальных» множеств ранг множества a отвечал за высоту гамма-дерева этого множества, которая определялась максимальной длиной цепочки принадлежности $a \ni b \ni c \ni \dots \ni z$. Для ω гамма-дерево — это корень с дочерними вершинами — натуральными числами, у каждого из которых есть свое конечное гамма-дерево. Понятно, что гамма-дерево ω получается бесконечным, однако, как того и требует аксиома регулярности, все цепочки вида $\omega \ni a \ni b \ni c \ni \dots \ni z$ в нем конечные. При этом не существует максимальной по длине цепочки. Какое бы n мы ни взяли, найдется цепочка длины больше n . Поэтому можно считать, что высота гамма-дерева множества ω равна ω .

Итак, α -ый универсум — это множество, содержащее в качестве элементов все множества «высоты» меньше α . Здесь мы уже видим, как начинает проявляться важнейший из архетипов — **архетип числа**. Пока ординалы еще не обрели арифметической структуры (этим мы займемся в следующей главе), но они уже могут эффективно использоваться для *перечисления*. Причем, помогает им в этом трансфинитная рекурсия — еще один архетипический представитель. Таким образом, мы можем здесь зафиксировать архетип числа в самом общем виде.

Важнейшим из архетипов для нас является число!

Универсумы (некоторые) удобны тем, что их можно использовать для моделирования различных теоретико-множественных аксиоматик для доказательства их совместности (непротиворечивости). Так, универсум V_ω служит моделью для «начальных» множеств, которые задаются аксиомами ZF без аксиомы бесконечности. Конечно, суждение такого вида тривиально, но зато наглядно демонстрирует традиционный способ доказательства совместности некоторой теории: если есть теория T , непротиворечивость которой доказана (или принимается как постулат), и в теории T построена модель теории T' (т. е. внутри теории T задано некоторое множество M , на котором определены отношения и функции, соответствующие атомарным формулам теории T' , так, что на M реализуются аксиомы теории T'), то теория T' также совместна.

Наиболее интересны в плане моделирования универсумы с индексом, равным какому-нибудь сверхбольшому кардинальному числу. Это позволяет погрузить в модель не только саму теорию ZF, но и дополнительные аксиомы.

Собственно, следующим нашим шагом будет построение модели мультимножеств в теории ZF. При этом нам придется использовать не только конечные, но и бесконечные кратности, и вот почему. Пусть, например, у нас есть мультимножества вида $m_k = \{k \bullet a\}$, где k — кратность элемента a , $k < \omega$. Требуется построить $\cup\{m_k \mid k < \omega\}$. Очевидно, что в полученном множестве должен быть элемент a с кратностью ω . Таким образом, либо мы должны запрещать произвольные операции объединения мультимножеств, либо до-

пускать бесконечные кратности. Второй вариант является, очевидно, более органичным в ZF.

Как и в разделе 1.1.9, построим универсумы мультимножеств с кратностями $< \lambda$, где λ на этот раз будет произвольным ординалом. Дадим рекурсивное определение (не путать степень и индекс кратности!):

$$\begin{aligned} mV_0^\lambda &= \emptyset \\ mV_{\alpha+1}^\lambda &= \{f \mid (f : D \rightarrow \lambda \setminus \{0\}) \wedge (D \subseteq V_\alpha^\lambda)\} \\ mV_\alpha^\lambda &= \cup \{V_\beta^\lambda \mid \beta < \alpha\}, \text{ если } \alpha \text{ — предельный ординал} \end{aligned}$$

Функции–элементы универсумов mV_α^λ будем называть **мультимножествами до- λ кратности**. Все рассмотренные в разделе 1.1.9 мультимножества, основанные на скобочной записи, являются мультимножествами до- ω кратности и полностью моделируются универсумом mV_ω^ω .

Обозначение mV введено для того, чтобы отличать данные универсумы от построенных ранее для «начальных» мультимножеств. Дело в том, что здесь сами универсумы mV являются обычными множествами, а их элементы — это функции. То есть мультимножества — это функции специального вида. Таким образом, элементом мультимножества, представленного функцией f является элемент области определения функции f , а кратностью этого элемента в данном мультимножестве является значение функции f на данном элементе. Поэтому в определении мультимножеств на значения функции накладывается ограничение: она не может принимать нулевое значение, т. к. нулевая кратность должна означать отсутствие элемента в мультимножестве.

Несколько примеров:

$$\begin{aligned} mV_0^2 &= \emptyset \\ mV_1^2 &= \{\emptyset\} \\ mV_2^2 &= \{\emptyset, \{(0, 1)\}\} \\ mV_3^2 &= \{\emptyset, \{(0, 1)\}, \{(\{(0, 1)\}, 1)\}, \{(0, 1), (\{(0, 1)\}, 1)\}\} \end{aligned}$$

В универсуме mV_3^2 элементами являются 4 функции: пустая, $f = \{(0, 1)\}$ — переводящая 0 в 1, $\{(f, 1)\}$ — переводящая f в 1, $\{(0, 1), (f, 1)\}$ — переводящая 0 и f в 1.

Заметим, что $mV_0^2 \subseteq mV_1^2 \subseteq mV_2^2 \subseteq mV_3^2$. На самом деле, всегда верно, что $mV_\alpha^\lambda \subseteq mV_\beta^\lambda$, если $\alpha < \beta$.

Действительно, предположим, что это не так, и пусть β — наименьший такой ординал, для которого существует $\alpha < \beta$ такой, что $mV_\alpha^\lambda \not\subseteq mV_\beta^\lambda$. Для данного β выберем наименьший $\alpha < \beta$, удовлетворяющий этому условию. Ясно, что β не может быть предельным ординалом (поскольку в этом случае $mV_\alpha^\lambda \subseteq mV_\beta^\lambda$ по определению). Следовательно, $\beta = \gamma + 1$.

Пусть $f \in mV_\alpha^\lambda$. Пусть далее α' — минимальный ординал такой, что $f \in mV_{\alpha'}^\lambda$ (если α — предельный, то $\alpha' < \alpha$). Очевидно, что α' не может быть предельным, следовательно, $\alpha' = \alpha'' + 1$. Тогда по определению имеем: $\text{dom}(f) \subseteq mV_{\alpha''}^\lambda$. Но в силу выбора минимальности α и β получаем, что $mV_{\alpha''}^\lambda \subseteq mV_\alpha^\lambda \subseteq mV_\gamma^\lambda$. Но тогда по определению $f \in mV_{\gamma+1}^\lambda = mV_\beta^\lambda$. Таким образом, $mV_\alpha^\lambda \subseteq mV_\beta^\lambda$. Противоречие.

Следовательно, отображение mV_α^λ монотонно по α . Более того, это отображение строго монотонно.

Атомарные формулы на мультимножествах принимают вид:

*Упражнение 1.37.
Проверьте строгую монотонность.*

$$\begin{aligned} f \dot{=} g &\Leftrightarrow f = g \\ f \in^\alpha g &\Leftrightarrow (f \in \text{dom}(g)) \wedge (g(f) = \alpha) \end{aligned}$$

Заметим, что никакое натуральное число (и тем более никакой ординал) не является мультимножеством. Поэтому мультимножество не может быть использовано как обозначение кратности элемента мультимножества.

Почему нельзя было определить мультимножества проще? Например, в качестве мультимножества взять пару (a, f) , где a — произвольное множество, а f — функция на нем со значениями-ординалами.

*Упражнение 1.38.
Проверьте это!*

Дело в том, что если в этом случае рассмотреть формулу $(a, f) \in^k (b, g)$, то либо нужно требовать, чтобы $(a, f) \in b$, и тогда b — уже не произвольное множество, либо $a \in b$, но тогда нет никакой разницы между $(a, f) \in^k (b, g)$ и $(a, f') \in^k (b, g)$, т. к. f и f' (кратности элементов a) никак не участвуют в определении принадлежности.

По сути мы реализуем первый случай: $(a, f) \in b$, только все несущие множества строим рекурсивно, отсекая в них все лишнее.

Определим следующие отношения и операции над мультимножествами:

$$\begin{aligned} f \in g &\Leftrightarrow \exists \alpha > 0 : f \in^\alpha g \\ f \subseteq g &\Leftrightarrow (\text{dom}(f) \subseteq \text{dom}(g)) \wedge (\forall h \in \text{dom}(f) : f(h) \leq g(h)) \\ f \sqcup g &\Leftrightarrow \{(h, \alpha) \mid (h \in \text{dom}(f) \cap \text{dom}(g) \rightarrow \alpha = \max(f(h), g(h))) \wedge \\ &\quad \wedge (h \in \text{dom}(f) \setminus \text{dom}(g) \rightarrow \alpha = f(h)) \wedge \\ &\quad \wedge (h \in \text{dom}(g) \setminus \text{dom}(f) \rightarrow \alpha = g(h))\} \\ f \sqcap g &\Leftrightarrow \{(h, \alpha) \mid (h \in \text{dom}(f) \cap \text{dom}(g)) \wedge (\alpha = \min(f(h), g(h)))\} \\ f \otimes g &\Leftrightarrow \{((h_1, h_2), \alpha) \mid (h_1 \in \text{dom}(f)) \wedge (h_2 \in \text{dom}(g)) \wedge (\alpha = f(h_1) \cdot g(h_2))\}, \end{aligned}$$

где $f \otimes g$ обозначает прямое произведение мультимножеств.¹⁹

¹⁹Произведение ординалов $f(h_1) \cdot g(h_2)$ будет определено в следующей главе.

Поскольку мультимножества — это функции, мы можем позволить себе пользоваться более привычной нотацией для их определения:

$$(f \uplus g)(h) = \begin{cases} \max(f(h), g(h)), & \text{если } h \in \text{dom}(f) \cap \text{dom}(g) \\ f(h), & \text{если } h \in \text{dom}(f) \setminus \text{dom}(g) \\ g(h), & \text{если } h \in \text{dom}(g) \setminus \text{dom}(f) \end{cases}$$

$$(f \sqcap g)(h) = \min(f(h), g(h)), \text{ если } h \in \text{dom}(f) \cap \text{dom}(g)$$

$$(f \otimes g)(h_1, h_2) = f(h_1) \cdot g(h_2), \text{ если } h_1 \in \text{dom}(f) \wedge h_2 \in \text{dom}(g)$$

Пусть F — некоторое непустое мультимножество. Тогда положим:

$$(\uplus F)(h) \Rightarrow \sup\{f(h) \mid (f \in F) \wedge (h \in \text{dom}(f))\}$$

$$(\sqcap F)(h) \Rightarrow \min\{f(h) \mid (f \in F) \wedge (h \in \bigcap_{g \in F} \text{dom}(g))\}$$

В последней формуле предполагается, что $\bigcap_{g \in F} \text{dom}(g)$ не пусто, в противном случае следует положить по определению, что $\sqcap F = 0$.

Мультимножества до-2 кратности естественным образом изоморфны обычным множествам в том смысле, что между ними можно установить взаимно однозначное соответствие, переводящее отношение \in на множествах в \in^1 на мультимножествах, и отношение $=$ на множествах в отношение \doteq на мультимножествах. Для этого достаточно построить биекции $h_\alpha : V_\alpha \leftrightarrow mV_\alpha^2$ по следующим правилам:

$$h_\alpha(\emptyset) = \emptyset,$$

$$h_\alpha(x) = \{(f, 1) \mid y \in x \wedge h_\alpha(y) = f\}, \text{ где } x \in V_\alpha.$$

Например, функция h_3 биективно переводит универсум $V_3 = \{0, 1, \{1\}, \{0, 1\}\}$ (где, как мы помним, $1 = \{0\}$) в универсум

$$mV_3^2 = \{0, \{(0, 1)\}, \{(f, 1)\}, \{(0, 1), (f, 1)\}\}, \text{ где } f = \{(0, 1)\}.$$

Натуральные числа при этом соответствуют следующим мультимножествам:

$$h_\omega(n) = \{(h_\omega(0), 1), (h_\omega(1), 1), \dots, (h_\omega(n-1), 1)\},$$

т.е. каждая следующее значение $h_\omega(n)$ — это функция, переводящая все предыдущие значения h_ω в единицу.

Анализ свойств мультимножеств мы оставляем за рамками данной книги, отсылая читателя, например, к книге [26]. Здесь нам было важно показать, что в обычной теории множеств существует модель мультимножеств.

Интересно, что если теперь рассматривать мультимножества, у которых все элементы имеют кратность 1, то мы получим модель **ZF** внутри самой **ZF**.

Более того, в теории **ZF** мы можем построить модель грамматики начальных множеств и ввести равенство объектов грамматики так, как мы это делали в разделе 1.1.1, то есть $a \equiv b$, если записи a и b посимвольно совпадают.

Выше мы уже имели пример реализации правил грамматики на множествах теории **ZF**, когда строили соответствие грамматических чисел и натуральных чисел. Ничто не мешает пополнить множество A всеми терминальными символами и рассматривать A^* как совокупность слов в грамматике. В частности, мы можем отдельно реализовать исчисление высказываний и изучить его свойства (на этом основано доказательство теоремы о полноте исчисления высказываний).

То есть, если сначала мы изучали множества при помощи грамматики и матлогики, то теперь можем изучать грамматику и матлогику при помощи теории **ZF** (и матлогики). Таким образом, фундамент математики замыкается сам на себя и как бы повисает в воздухе, поскольку ничего более фундаментального человечеству пока придумать не удалось.

Вложение какой-либо аксиоматической теории в язык **ZF** называется моделированием этой теории в **ZF**, а множество, которое реализует объекты этой теории, называется ее моделью. Больше того, сама математика (как семейство наук) является не чем иным как очень гибким языком моделирования реальности, а погружение всех моделей в **ZF** только придает им строгости и, тем самым, усиливает нашу веру в адекватность представлений о реальности. В этом смысле уже никак нельзя считать основание математики «подвешенным в воздухе», поскольку оно возникло и продолжает развиваться на стыке логики и моделирования естественно-научных и гуманитарных знаний.

*Любители
теории
категорий,
наверное, не
согласятся*

*А здесь
будут
против
сторонники
Бурбаки :)*

Финал главы 1

Итак, мы изучили основу современной математики — теорию множеств, и даже коснулись несколько необычной ее версии, включающей мультимножества.

Мы также увидели глубокую связь между теорией множеств и арифметикой, которая означает, что, в общем-то, не так важно, что ставить в основание математики, вопрос лишь в удобстве интерпретации.

В дальнейшем мы будем по-прежнему рассматривать теорию множеств как корень, а арифметику — как одну из ветвей математического дерева. Поэтому все построения будут основаны на теоретико-множественной терминологии.

В связи с этим уместно от собственно множеств перейти к различным надстройкам над ними, позволяющим образовывать системы множеств.

Одна такая надстройка нам уже известна — это отношение. Она используется в качестве основы прямое произведение множеств, в котором выбирается некоторое подмножество.

Существует и другой подход к надстройкам — сначала строится булеан (множество всех подмножеств), а затем из него выбирается некоторое подмножество, в итоге получается некоторое выделенное семейство подмножеств на исходном множестве.

Хм... а в
машине
стоит
4-тактный
ДВС...

Иначе говоря, мы *всегда наблюдаем двухтактную модель построения математических систем*: **1)** на основном множестве строится некоторая всеобъемлющая (генеральная) совокупность (прямое произведение, либо множество всех подмножеств), **2)** из этой совокупности выбирается некоторое подмножество элементов, связанных определенным набором правил (создается структура определенного рода, говоря словами Н. Бурбаки).

Через прямое произведение мы задаем отношения на множествах, графы, функции, операции и операторы и т.д., т.е. такие конструкции, которые работают по правилу *точка-точка* (в общем случае).

Через множество подмножеств мы задаем некоторые специальные системы подмножеств — алгебру множеств, фактор-множество, топологию, фильтры и т.п.

В линейном
простран-
стве, правда,
это не сразу
проявляет-
ся...

Обычно множество с системой подмножеств называется **пространством**, поскольку ее конструкция использует «пространные» объекты — подмножества.

Существуют и разного рода переплетения двух тактов процедуры надстроек. Например, в линейной алгебре мы сначала (такт 1) строим прямое произведение базового множества (\mathbb{R}^n), затем (такт 2) на его элементах задаем операции и норму, а далее строим булеан (снова такт 1), и в нем (снова такт 2) с помощью нормы определяем топологию на этом пространстве.

Вероятностное пространство (и любое пространство с мерой) строится симметричным способом: сначала (такт 1) строится булеан, на нем (такт 2) определяется система подмножеств (алгебра или сигма-алгебра), затем (снова такт 1) рассматривается прямое произведение системы подмножеств с

И все-таки
двигатель
4-тактный
:)

какой-нибудь числовой структурой, и уже в нем (снова такт 2) выбирается некоторая функция (например, мера — специальное подмножество прямого произведения алгебры множеств и действительной прямой).

На протяжении всех следующих глав мы будем строить математические объекты, в основном придерживаясь изложенной схемы — явно или неявно.

Числа

В первой главе мы уже имели честь познакомиться с натуральными числами и даже свели к ним (посредством универсального кода) «начальные» множества. И раз уж множества у нас являются одним из основных архетипов математики, то, по-видимому, и архетип числа стоит где-то рядом.

На самом деле, число — это намного более древнее и фундаментальное понятие, которое пронизывает не только всю математику, но и множество наук. Для большинства людей вообще сама математика ассоциируется прежде всего с числами.

Число как абстрактное понятие играет, по-видимому, несколько ролей в нашем сознании:

1. дискретизация информации;
2. шкалирование;
3. алгоритмизация.

Под дискретизацией информации мы понимаем возможность разложить все многообразие известных фактов на кластеры со схожими числовыми параметрами. Близость чисел отвечает здесь за сходство качественных характеристик, которым приписаны те или иные числа.

Шкалирование — это возможность сравнить факты, действия, обстоятельства, предметы, которым по тем или иным причинам приписаны числа, т. е. произведена их «оцифровка».

Наконец, алгоритмизация обычно означает работу с оцифрованными объектами в соответствии со структурой самих чисел, т. е. перенос с чисел на объекты порядка, арифметических операций, разложений чисел на компоненты и т. п. Это позволяет не только сравнивать уже оцифрованные объекты,

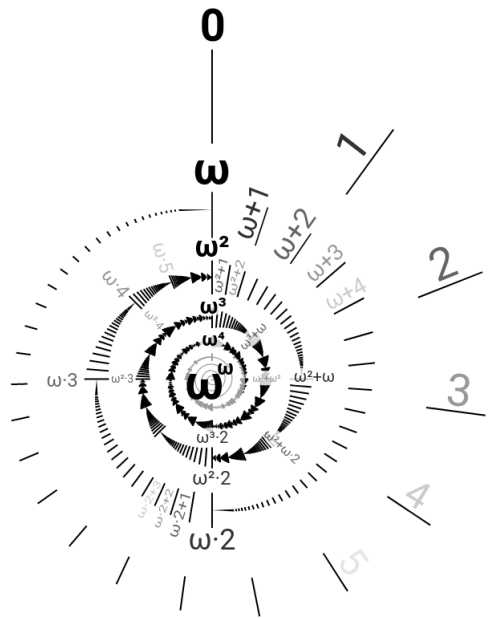


Рис. 2.1: Ординал ω^ω (изображение свободно от авторских прав согласно CC0).

но и предсказывать оцифровку новых объектов информации, получаемых из ранее известных путем синтеза, а значит, предсказывать и их физические свойства и поведение.

Скорее всего, опытный философ приведет еще десяток-другой примеров, в том числе исторических, того, насколько значимы для цивилизации числа. Но нас здесь будут занимать прежде всего вопросы шкалирования (упорядочения) и алгоритмизации (вычислений), которые имманентны всем математическим числам и сопутствующим исчислениям.

Для начала мы вернемся к основе — натуральным числам и их естественному продолжению — ординалам. Затем коснемся кардинальной арифметики, и только потом построим остальные известные со школы числовые системы.

В этой главе мы все еще стараемся придерживаться грамматических правил, введенных в первой главе, но позволим себе больше вольностей в использовании символики. Тем не менее, следует помнить о том, что все это лишь сокращающие обозначения для термов и формул, заданных в языке и аксиоматике ZF.

2.1 Арифметика порядковых чисел

Арифметические операции определим сразу для ординалов, а для натуральных чисел они будут заданы автоматически (поскольку натуральные числа — это конечные ординалы). Более подробно с определением операций и доказательством их свойств можно познакомиться, например, в [25].

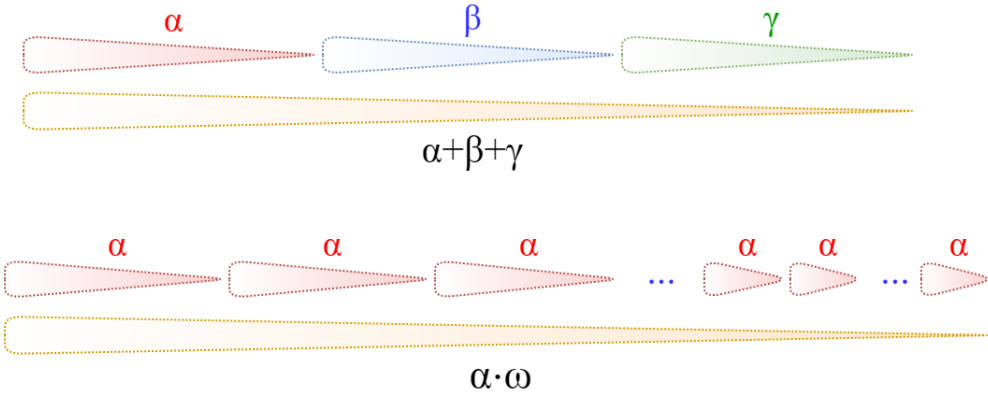


Рис. 2.2: Визуальное представление суммы и произведения.

2.1.1 Сложение

Ординалы складываются последовательной стыковкой друг к другу в порядке их написания или индексации ординалом. Например, сумма $\alpha + \beta$ озна-

чает, что мы поставили следом за α порядковый тип β и взяли результирующий порядковый тип. Сумма $\alpha + \beta + \gamma$ означает, что мы следом за α поставили порядковый тип β , затем к нему пристыковали порядковый тип γ , и в качестве суммы взяли результирующий порядковый тип.

Формально, пусть есть последовательность ординалов $\{\alpha_\kappa \mid \kappa < \Lambda\}$, проиндексированная ординалом Λ (напоминаем, что это не что иное как функция, определенная на ординале Λ и принимающая значения в классе ординалов). Построим новое множество¹

$$\bigsqcup_{\kappa < \Lambda} \alpha_\kappa \rightleftharpoons \bigcup_{\kappa < \Lambda} \alpha_\kappa \times \{\kappa\} = \{(\beta, \kappa) \mid (\kappa < \Lambda) \wedge (\beta < \alpha_\kappa)\},$$

и введем на нем порядок $<$ следующим способом:

$$(\beta, \kappa) < (\gamma, \iota), \text{ если } (\kappa < \iota) \vee ((\kappa = \iota) \wedge (\beta < \gamma))$$

Иначе говоря, все элементы слагаемого с меньшим номером строго меньше всех элементов слагаемого с большим номером, а внутри одного слагаемого используется их собственный порядок.

Нетрудно доказать, что множество $\bigsqcup_{\kappa < \Lambda} \alpha_\kappa$ вполне упорядочено этим отношением. Его порядковый тип называется **суммой ординалов α_κ в порядке Λ** и обозначается

Упражнение 2.1.

$$\sum_{\kappa < \Lambda} \alpha_\kappa \rightleftharpoons \left| \left(\bigsqcup_{\kappa < \Lambda} \alpha_\kappa, < \right) \right| \quad (2.1)$$

В случае, когда Λ — натуральное число, допустимо использовать запись $\alpha_0 + \alpha_1 + \dots + \alpha_n$, а также варианты $\alpha + \beta$, $\alpha + \beta + \gamma$ и т.п.

Помимо прямого определения суммы ординалов существует рекурсивный способ определения сложения двух ординалов:

$$\begin{aligned} \alpha + 0 &= \alpha \\ \alpha + (\beta + 1) &= (\alpha + \beta) \cup \{\alpha + \beta\} \\ \alpha + \beta &= \sup\{\alpha + \gamma \mid \gamma < \beta\}, \text{ если } \beta \text{ — предельный ординал} \end{aligned}$$

¹Такое множество еще называют **дизъюнктивной суммой**, поскольку она делает все слагаемые α_κ попарно непересекающимися за счет приписывания их элементам нумерации.

и, на его основе, — сложения многих ординалов в порядке Λ :

$$\begin{aligned} \sum_{\varkappa < \Lambda} \alpha_{\varkappa} &= 0, \text{ если } \Lambda = 0 \\ \sum_{\varkappa < \Lambda} \alpha_{\varkappa} &= \left(\sum_{\varkappa < \lambda} \alpha_{\varkappa} \right) + \alpha_{\lambda}, \text{ если } \Lambda = \lambda + 1 \\ \sum_{\varkappa < \Lambda} \alpha_{\varkappa} &= \sup \left\{ \sum_{\varkappa < \lambda} \alpha_{\varkappa} \mid \lambda < \Lambda \right\}, \text{ если } \Lambda \text{ — предельный ординал} \end{aligned} \quad (2.2)$$

Теорема 2.1. *Варианты определений сложения ординалов, заданные при помощи формул (2.1) и (2.2), эквивалентны.*

Для доказательства этой теоремы достаточно воспользоваться трансфинитной индукцией и аккуратно построить соответствие между дизъюнктивной суммой из формулы (2.1) и суммой ординалов, возникающей на шаге индукции.

Сложение ординалов обладает следующими свойствами:

| Упражнение
2.2.

OS1 : $\alpha + 0 = 0 = 0 + \alpha$ (левый и правый нейтральные элементы)

OS2 : $(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma)$ (ассоциативность)

OS3 : если $\alpha < \beta$, то $\gamma + \alpha < \gamma + \beta$ (строгая монотонность по второму слагаемому)

OS4 : строгой монотонности по первому слагаемому нет: $1 < 2$, но $1 + \omega = 2 + \omega$

OS5 : если $\alpha \leq \beta$, то $\alpha + \gamma \leq \beta + \gamma$ (нестрогая монотонность по первому слагаемому)

OS6 : если $\alpha + \beta = \alpha + \gamma$, то $\beta = \gamma$ (сокращение слева)

OS7 : сокращение справа не всегда верно: $1 + \omega = 2 + \omega = \omega$, но $1 \neq 2$

OS8 : для любого α , если $\beta < \alpha$, то $\alpha = \beta + \gamma$ (вычитание слева: $-\beta + \alpha = \gamma$), причем γ определяется однозначно

OS9 : вычитание справа в общем случае не работает: не существует γ такого, что $\gamma + 2019 = \omega$

Чтобы продемонстрировать стандартную методику доказательств на ординалах, покажем справедливость свойства OS8.

Теорема 2.2. Для любого ординала α , если $\beta < \alpha$, то существует ординал γ такой, что $\alpha = \beta + \gamma$.

Доказательство. Имеем: $0 \leq \beta$, откуда по свойству OS5 $\alpha \leq \beta + \alpha$. Пусть далее $\gamma = \min\{\gamma' \leq \alpha \mid \alpha \leq \beta + \gamma'\}$.

Понятно, что $\alpha \leq \beta + \gamma$. С другой стороны, если $\delta < \gamma$, то $\alpha > \beta + \delta$.

Пусть $\gamma = \delta + 1$, тогда $\beta + \delta < \alpha \leq (\beta + \delta) + 1 = \beta + \gamma$, где последнее равенство следует из определения суммы. Но для любого $\delta' \in \beta + \gamma$ имеем $\delta' \leq \beta + \delta$, откуда $\beta + \gamma \subseteq \alpha$, т. е. $\beta + \gamma \leq \alpha$. Следовательно, $\alpha = \beta + \gamma$.

Пусть γ — предельный ординал. Тогда $\beta + \gamma = \sup\{\beta + \delta \mid \delta < \gamma\} \leq \alpha$, что вместе с неравенством $\alpha \leq \beta + \gamma$ дает $\alpha = \beta + \gamma$. \square

Эту теорему можно доказать проще: достаточно взять множество $\alpha \setminus \beta$, которое вполне упорядочено, а значит, имеет порядковый тип γ . Нетрудно показать, что в таком случае $\alpha = \beta + \gamma$ по определению (2.1). Но приведенное выше доказательство более характерно для ординальной арифметики и всех фактов, которые основаны на ней.

Введенная операция сложения на ординалах естественным образом порождает сложение натуральных чисел.

Упражнение 2.3.
Докажите, что
 $2 + 2 = 4$.

2.1.2 Умножение

Чуть сложнее, но концептуально так же, определяется умножение ординалов. $\alpha\beta$ означает, что ординал α сам с собой суммируется β раз, т. е. $\alpha\beta = \sum_{\delta < \beta} \alpha$.

Другой способ задать произведение двух ординалов — это взять прямое произведение $\alpha \times \beta$ и ввести на нем инверсный лексикографический порядок²:

$$(\gamma, \delta) < (\gamma', \delta'), \text{ если } (\delta < \delta') \vee ((\delta = \delta') \wedge (\gamma < \gamma')). \quad (2.3)$$

То, что эти два определения эквивалентны, видно из:

$$\sum_{\delta < \beta} \alpha = \left| \bigsqcup_{\delta < \beta} \alpha \right| = |\{(\gamma, \delta) \mid (\gamma < \alpha) \wedge (\delta < \beta)\}| = \left| \bigcup_{\delta < \beta} \alpha \times \{\delta\} \right|,$$

поскольку в определении дизъюнктивной суммы все пары совпадают с парами в определении прямого произведения.

²Прямой лексиграфический порядок подразумевает, что ведущую роль в сравнении кортежей играют начальные компоненты: $1122 < 1211$, тогда как при инверсном лексикографическом порядке ведущую роль играют последние компоненты: $aabb < aaab$.

Ничем не
лучше был
бы обратный
вариант.

Трудно сказать, почему было выбрано такое определение произведения, где элементы второго множителя играют роль индекса для копий первого множителя, но для его запоминания достаточно знать формулу $\omega \cdot 2 = \omega + \omega$ («омега дважды» — это «омега плюс омега»).

Рекурсивно произведение двух ординалов задается следующим способом:

$$\begin{aligned}\alpha \cdot 0 &= 0 \\ \alpha(\beta + 1) &= \alpha\beta + \alpha \\ \alpha\beta &= \sup\{\alpha\gamma \mid \gamma < \beta\}, \text{ если } \beta \text{ — предельный ординал}\end{aligned}$$

и, на его основе, — умножение последовательности ординалов $\{\alpha_\kappa \mid \kappa < \Lambda\}$ в порядке Λ :

$$\begin{aligned}\prod_{\kappa < \Lambda} \alpha_\kappa &= 1, \text{ если } \Lambda = 0 \\ \prod_{\kappa < \Lambda} \alpha_\kappa &= \left(\prod_{\kappa < \lambda} \alpha_\kappa \right) \alpha_\lambda, \text{ если } \Lambda = \lambda + 1 \\ \prod_{\kappa < \Lambda} \alpha_\kappa &= \sup \left\{ \prod_{\kappa < \lambda} \alpha_\kappa \mid \lambda < \Lambda \right\}, \text{ если } \Lambda \text{ — предельный ординал}\end{aligned}$$

В случае, когда Λ — натуральное число, допустимо использовать запись $\alpha_0 \alpha_1 \dots \alpha_n$, а также варианты $\alpha\beta$, $\alpha\beta\gamma$, $\alpha \cdot \beta$, $\alpha \cdot \beta \cdot \gamma$ и т.п.

Умножение ординалов обладает следующими свойствами:

Упражнение
2.4.

ОР1 : $\alpha \cdot 0 = 0 = 0 \cdot \alpha$

ОР2 : $\alpha \cdot 1 = \alpha = 1 \cdot \alpha$

ОР3 : $(\alpha\beta)\gamma = \alpha(\beta\gamma)$

ОР4 : если $\alpha < \beta$ и $\gamma > 0$, то $\gamma\alpha < \gamma\beta$ (строгая монотонность по второму множителю)

ОР5 : строгой монотонности по первому множителю нет: $1 < 2$, но $1\omega = 2\omega$

ОР6 : если $\alpha \leq \beta$, то $\alpha\gamma \leq \beta\gamma$ (нестрогая монотонность по первому множителю)

ОР7 : если $\alpha\beta = \alpha\gamma$ и $\alpha > 0$, то $\beta = \gamma$ (сокращение слева)

ОР8 : сокращение справа не всегда верно: $1\omega = 2\omega$, но $1 \neq 2$

ОР9 : если $\alpha\beta = 0$, тогда $\alpha = 0$ или $\beta = 0$

ОР10 : $\alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma$ (дистрибутивный закон слева)

ОР11 : дистрибутивный закон справа не всегда выполняется:
 $(\omega + 1) \cdot 2 \neq \omega \cdot 2 + 2$

ОР12 : для любого α , если $\beta > 0$, то $\alpha = \beta\gamma + \delta$, где $\delta < \beta$ (деление слева с остатком: $\beta^{-1}\alpha$), причем γ и δ определяются однозначно

ОР13 : деление справа с остатком не всегда работает: не существует α такого, что $\alpha\omega \leq \omega^\omega \leq (\alpha + 1)\omega$

Введенная операция умножения на ординалах естественным образом порождает умножение натуральных чисел.

Докажем для примера свойство ОР12.

Упражнение
2.5.
Докажите,
что $2 \cdot 2 = 4$.

Теорема 2.3 (Деление с остатком). *Для любого ординала α , если $\beta > 0$, то $\alpha = \beta\gamma + \delta$, где $\delta < \beta$ причем γ и δ определяются однозначно.*

Доказательство. Пусть $\alpha_0 = \min\{\alpha' \leq \alpha \mid \beta\alpha' > \alpha\}$.

Если α_0 — предельный ординал, то $\beta\alpha_0 = \sup\{\beta\alpha' \mid \alpha' < \alpha_0\} \leq \alpha$, что противоречит определению α_0 : $\beta\alpha_0 > \alpha$. Следовательно, $\alpha_0 = \gamma + 1$ при некотором γ , тогда

$$\beta\gamma \leq \alpha < \beta\alpha_0 = \beta(\gamma + 1) = \beta\gamma + \beta.$$

В этом случае положим $\delta = \min\{\delta' \leq \beta \mid \beta\gamma + \delta' \geq \alpha\}$. Очевидно, что при этом $\alpha = \beta\gamma + \delta$ и $\delta < \beta$.

Единственность разложения. Предположим, что

$$\alpha = \beta\gamma + \delta = \beta\gamma' + \delta', \quad \delta < \beta, \quad \delta' < \beta.$$

Пусть $\gamma > \gamma'$, тогда о свойстве OS8 имеем: $\gamma = \gamma' + \sigma$, $\sigma \geq 1$. Откуда:

$$\beta\gamma + \delta = \beta(\gamma' + \sigma) + \delta = \beta\gamma' + \beta\sigma + \delta \geq \beta\gamma' + \beta > \beta\gamma' + \delta'$$

в силу свойства сложения OS3.

То есть $\alpha = \beta\gamma + \delta > \beta\gamma' + \delta' = \alpha$. Противоречие. Аналогично исключается случай $\gamma' > \gamma$.

Итак, $\gamma = \gamma'$, и, следовательно, $\beta\gamma + \delta = \beta\gamma + \delta'$. Отсюда по свойству OS6 производим сокращение слева и получаем $\delta = \delta'$. \square

Ординал γ называется **частным**, а ординал δ — **остатком**. Если остаток равен 0, то говорят, что α делится на β **нацело** (слева) и пишут: $\beta \mid \alpha$.³

³Важно понимать, что в арифметике ординалов деление слева и справа отличаются в силу некоммутативности умножения, поэтому и деление с остатком, и деление нацело в ординалах всегда подразумевает, что делитель стоит слева, а частное — справа. Однако, при сужении арифметики ординалов до области натуральных чисел коммутативность умножения появляется, и данное уточнение уже не требуется.

Например, разделим $\omega^2 \cdot 3 + \omega \cdot 5 + 8$ на $\omega + 2$ с остатком:

$$\omega^2 \cdot 3 + \omega \cdot 5 + 8 = (\omega + 2)(\omega \cdot 3 + 5) + 6,$$

т. е. неполным частным тут будет $\omega \cdot 3 + 5$, а остатком 6, поскольку $(\omega + 2) \cdot 5 = \omega \cdot 5 + 2$ (напомним, что дистрибутивный закон справа не работает, поэтому умножение на 5 нужно представить как сложение 5 раз, после чего всюду заменить $2 + \omega$ на ω).

Возможность деления с остатком позволяет рассмотреть на ординалах **алгоритм Евклида**. Пусть $\alpha_0 > \alpha_1 > 0$. Тогда по свойству ОР12 получаем следующие выкладки:

$$\begin{aligned} \alpha_0 &= \alpha_1 \gamma_1 + \alpha_2, & \alpha_2 &< \alpha_1, \\ \alpha_1 &= \alpha_2 \gamma_2 + \alpha_3, & \alpha_3 &< \alpha_2, \\ &\dots & \dots & \\ \alpha_{n-2} &= \alpha_{n-1} \gamma_{n-1} + \alpha_n, & \alpha_n &< \alpha_{n-1}. \\ \alpha_{n-1} &= \alpha_n \gamma_n + \alpha_{n+1}, & \alpha_{n+1} &< \alpha_n. \end{aligned} \tag{2.4}$$

В итоге у нас получается строго убывающая последовательность ординалов

$$\alpha_0 > \alpha_1 > \dots > \alpha_n > \dots,$$

которая не может быть бесконечной в силу вполне упорядоченности ординалов. Это значит, что на каком-то n последовательность положительных остатков прервется: $\alpha_n > 0$, $\alpha_{n+1} = 0$. И мы получим $\alpha_{n-1} = \alpha_n \gamma_n$, т. е. $\alpha_n | \alpha_{n-1}$.

Но тогда легко видеть, что $\alpha_{n-2} = \alpha_{n-1} \gamma_{n-1} + \alpha_n = \alpha_n (\gamma_n \gamma_{n-1} + 1)$, т. е. $\alpha_n | \alpha_{n-2}$. Раскручивая эти равенства вверх по последовательности (2.4), мы приходим к тому, что α_n делит нацело как α_1 , так и α_0 , т. е. является их общим делителем.

С другой стороны, если какое-то β делит α_0 и α_1 , т. е. $\alpha_0 = \beta \gamma$, $\alpha_1 = \beta \delta$, то $\beta \gamma = \beta \delta \gamma_1 + \alpha_2$. В этом случае $\alpha_2 = \beta (\delta \gamma_1 + \varepsilon)$, где ε единственным образом (в силу свойства OS8) определяется из равенства $\gamma = \delta \gamma_1 + \varepsilon$ (то, что $\delta \gamma_1 < \gamma$, следует из свойств умножения). Но тогда $\beta | \alpha_2$. Спускаясь по равенствам (2.4) вниз и используя аналогичные рассуждения, получаем, что β делит все ординалы $\alpha_0, \alpha_1, \dots, \alpha_n$. А это означает, что $\beta \leq \alpha_n$. Отсюда следует, что α_n является **наибольшим общим делителем** (слева) для ординалов α_0 и α_1 .

Таким образом, для любых двух ненулевых ординалов существует НОД (левый), который можно определить с помощью алгоритма Евклида (2.4). В частности, это работает в обычных натуральных числах.

Для примера выше имеем следующий алгоритм Евклида:

$$\begin{aligned}\omega^2 \cdot 3 + \omega \cdot 5 + 8 &= (\omega + 2)(\omega \cdot 3 + 5) + 6 \\ \omega + 2 &= 6 \cdot \omega + 2 \\ 6 &= 2 \cdot 3 + 0\end{aligned}$$

Таким образом, число 2 является НОД ординалов $\omega^2 \cdot 3 + \omega \cdot 5 + 8$ и $\omega + 2$.

Другой пример: $\text{НОД}(\omega^2 \cdot 3 + \omega \cdot 5, \omega \cdot 4) = \omega$.

Заметим, что в целых числах мы получаем еще один бонус, а именно — представление НОД в виде линейной комбинации с целыми коэффициентами от исходных двух чисел. И это легко увидеть из равенств (2.4), последовательно сворачивая их снизу вверх. Так, сначала мы имеем $\alpha_n = \alpha_{n-2} - \alpha_{n-1}\gamma_{n-1}$, затем подставляем выражение для α_{n-1} через вышестоящие «альфы», и т.д. В итоге мы получим линейную комбинацию

$$\alpha_n = \alpha_0\beta_0 + \alpha_1\beta_1,$$

где коэффициенты β_0 и β_1 могут быть отрицательными (в \mathbb{Z} это не является проблемой). Возможность такого представления НОД мы еще увидим в других числовых кольцах (см. раздел 3.3).

2.1.3 Степень

Частный случай произведения — степень. Если все $\alpha_k = \alpha$, то по определению положим

$$\alpha^\beta \doteq \prod_{\varkappa < \beta} \alpha_\varkappa.$$

Как видим, в случае ординалов степень обозначает вовсе не множество функций, а вполне конкретный ординал. Поэтому иногда множество функций вида $f : \beta \rightarrow \alpha$ обозначают $\alpha^{(\beta)}$, чтобы отличить от арифметической операции.

Степень можно задать и с помощью функций. Для этого рассматривается множество функций $f : \beta \rightarrow \alpha$ с конечным **носителем** (т.е. $f(\gamma) \neq 0$ на конечном подмножестве β), которые сравниваются в инверсном лексикографическом порядке, такое упорядочение будет полным (именно из-за конечности носителя функций), а его порядковый тип совпадет с ординалом α^β .

Рекурсивно степень ординала задается следующим способом:

$$\begin{aligned}\alpha^0 &= 1 \\ \alpha^{\beta+1} &= \alpha^\beta \alpha \\ \alpha^\beta &= \sup\{\alpha^\gamma \mid \gamma < \beta\}, \text{ если } \beta \text{ — предельный ординал}\end{aligned}$$

Стрелочная нотация Кнута в случае ординалов:

$$\begin{aligned}\alpha \uparrow\uparrow 0 &= 1 \\ \alpha \uparrow\uparrow (\beta + 1) &= \alpha^{\alpha \uparrow\uparrow \beta} \\ \alpha \uparrow\uparrow \beta &= \sup\{\alpha \uparrow\uparrow \gamma \mid \gamma < \beta\}, \text{ если } \beta - \text{предельный ординал}\end{aligned}$$

Свойства степеней ординалов:

Упражнение
2.6.

OW1 : $\alpha^0 = 1$

OW2 : $0^\alpha = 0$, если $\alpha > 0$

OW3 : $1^\alpha = 1$

OW4 : $\alpha^1 = \alpha$

OW5 : $\alpha^\beta \cdot \alpha^\gamma = \alpha^{\beta+\gamma}$

OW6 : $(\alpha^\beta)^\gamma = \alpha^{\beta \cdot \gamma}$

OW7 : существуют α , β и γ такие, что $(\alpha\beta)^\gamma \neq \alpha^\gamma \cdot \beta^\gamma$, например, $(\omega \cdot 2)^2 = \omega \cdot 2 \cdot \omega \cdot 2 = \omega^2 \cdot 2 \neq \omega^2 \cdot 4$

OW8 : если $\alpha < \beta$ и $\gamma > 1$, то $\gamma^\alpha < \gamma^\beta$ (строгая монотонность по показателю)

OW9 : строгой монотонности по основанию нет: $2 < 3$, но $2^\omega = 3^\omega = \omega$

OW10 : нестрогая монотонность: $\alpha \leq \beta \rightarrow \alpha^\gamma \leq \beta^\gamma$

OW11 : если $\alpha^\beta = \alpha^\gamma$ и $\alpha > 1$, то $\beta = \gamma$ (логарифмирование)

OW12 : извлечение корня не всегда верно: $1^\omega = 2^\omega$, но $1 \neq 2$

OW13 : для любых $\alpha > 0$ и $\beta > 1$: если $\alpha < \beta^\gamma$ при некотором γ , то $\alpha = \beta^\eta \cdot \varkappa + \delta$, где $0 < \varkappa < \beta$, $\eta < \gamma$ и $\delta < \beta^\eta$ (разложение по основанию), причем η , \varkappa и δ определяются однозначно

Упражнение 2.7. Докажите, что $2^2 = 4$. Введенная операция возведения в степень на ординалах естественным образом порождает возведение в степень натуральных чисел.

Докажем последнее свойство.

Теорема 2.4. Для любых ординалов $\alpha > 0$ и $\beta > 1$: если $\alpha < \beta^\gamma$ при некотором ординале γ , то

$$\alpha = \beta^\eta \cdot \varkappa + \delta, \quad (2.5)$$

где $0 < \varkappa < \beta$, $\eta < \gamma$ и $\delta < \beta^\eta$.

Доказательство. Пусть $\eta_1 = \min\{\gamma' \mid \beta^{\gamma'} > \alpha\}$ (это множество не пусто в силу условия $\alpha < \beta^\gamma$).

Предположим, что η_1 — предельный ординал, тогда $\beta_1^\eta = \sup\{\beta^{\gamma'} \mid \gamma' < \eta_1\} \leq \alpha$. Противоречие.

Следовательно, $\eta_1 = \eta + 1$, при этом $\beta^\eta \leq \alpha < \beta^{\eta+1}$.

По теореме (2.3) разделим α на β^η с остатком:

$$\alpha = \beta^\eta \cdot \varkappa + \delta, \quad \text{где } \delta < \beta^\eta.$$

Ясно, что $\varkappa < \beta$, т.к. иначе получим $\beta^\gamma \cdot \varkappa \geq \beta^{\eta+1} > \alpha$, что невозможно. Таким образом, существование разложения (2.5) доказано.

Единственность ординалов η, δ, \varkappa при заданных α и β следует из теоремы 2.3 и того, что η нельзя уменьшить при условии $\varkappa < \beta$. \square

2.1.4 Разложения ординалов

Пользуясь теоремой 2.4 мы можем построить рекурсивную последовательность троек ординалов $(\eta_i, \varkappa_i, \delta_i)$ такую, что

$$\alpha = \beta_0^\eta \cdot \varkappa_0 + \dots + \beta_i^\eta \cdot \varkappa_i + \delta_i.$$

Для этого нужно получаемый каждый раз остаток δ_i вновь представлять в виде $\delta_i = \beta_{i+1}^\eta \cdot \varkappa_{i+1} + \delta_{i+1}$.

Пользуясь свойствами арифметики ординалов, нетрудно показать, что мы получим строго убывающую последовательность степеней $\eta_0 > \eta_1 > \dots > \eta_j$, которая не может быть бесконечной (в силу вполне упорядоченности ординалов). Это означает, что в какой-то момент δ_i станет меньше β^0 , т.е. нулем, и дальнейшее разложение будет невозможно. При этом все коэффициенты в разложении будут определены единственным образом.

Итак, любой ординал α , для которого выполняется неравенство $\alpha < \beta^\gamma$ при некотором γ (в частности, натуральное число), может быть единственным образом разложен в сумму степеней по основанию β , где β — произвольный ординал (понятно, что имеет смысл рассматривать только $\alpha > \beta > 1$, иначе разложение будет тривиальным), в виде:

$$\alpha = \beta^{\eta_0} \cdot \varkappa_0 + \dots + \beta^{\eta_j} \cdot \varkappa_j,$$

где $\eta_0 > \eta_1 > \dots > \eta_j \geq 0$ и $\beta > \varkappa_0, \dots, \varkappa_j > 0$.

Подчеркнем, что порядок слагаемых в этой сумме существенен, т. к. $2 + \omega = \omega$, но $\omega + 2 > \omega$.

Положим теперь $\beta = \omega$, тогда будет справедливой следующая

Теорема 2.5 (Канторовская нормальная форма). *Если при некотором γ $\alpha < \omega^\gamma$, то*

$$\alpha = \omega^{\eta_0} \cdot k_0 + \dots + \omega^{\eta_i} \cdot k_j,$$

где ординалы $\eta_0 > \eta_1 > \dots > \eta_j \geq 0$, натуральные числа $k_0, \dots, k_j > 0$.

Ординал η_0 называется степенью α и удовлетворяет неравенству $\eta_0 \leq \alpha$, причем равенство $\eta_0 = \alpha$ достигается тогда и только тогда, когда $\alpha = \omega^\alpha$. Для таких ординалов α канторовская нормальная форма не сводит исходный ординал к конечной арифметической комбинации меньших ординалов, а значит, разложение α по супероснованию ω (в виде башен степеней) в этом случае невозможно.

Остается выяснить, как много существует «хороших» бесконечных ординалов, для которых возможно разложение по супероснованию ω в полном соответствии с таковым разложением натуральных чисел.

Дадим следующее определение. **ε -числами** называются ординалы, удовлетворяющие равенству $\varepsilon = \omega^\varepsilon$. Наименьшее из таких чисел обозначается ε_0 , и оно совпадает с ординалом $\omega \uparrow \uparrow \omega$.

Если $\alpha < \varepsilon_0$, то он однозначно раскладывается по супероснованию ω при помощи канторовской нормальной формы. Действительно, пусть $\eta_{0,0}$ — степень ординала α (по основанию ω), тогда $\eta_{0,0} < \alpha$ и, следовательно, к $\eta_{0,0}$ снова применимо разложение по основанию ω . Пусть $\eta_{0,1}$ — степень $\eta_{0,0}$, тогда $\alpha > \eta_{0,0} > \eta_{0,1}$, и т.д. Мы снова видим строго убывающую последовательность ординалов, которая не может быть бесконечной, т. е. за конечное число шагов мы дойдем до степени — натурального числа, и на этом разложение закончится.

Итак, все бесконечные ординалы $< \varepsilon_0$ однозначно раскладываются по супероснованию ω в виде конечного терма, содержащего только арифметические операции, натуральные числа и символ ω , что позволяет производить

*Компьютер
все еще с
нами!*

компьютерное моделирование арифметики таких ординалов. Например,

$$\omega \left(\omega^{\left(\omega^{12 \cdot 5 + \omega^2 \cdot 1024 + 1} \right) + \omega^{2018} + 32768} \right) \cdot 123321 + \omega^{\omega^\omega} \cdot 6 + 2019$$

представляет собой такой ординал.

Имея некоторый опыт работы с мультимножествами в ходе рассмотрения теоремы Гудстейна в разделе 1.1.10, мы теперь понимаем, что ординалы из интервала $[0; \varepsilon_0)$ взаимно однозначно нумеруют конечные мультимножества с

произвольными конечными кратностями элементов, а значит, позволяют нумеровать и произвольные формулы в языках первого порядка, на чем основывается ординальная теория доказательств (ординальный анализ). Например, Г. Генцен⁴ в 1936 году установил [33], что непротиворечивость арифметики Пеано PA может быть доказана трансфинитной индукцией до ординала ε_0 . Однако до сих пор остается открытым вопрос о вычислении теоретико-доказательственного ординала самой теории множеств с присоединенной аксиомой выбора (ZFC).

Мы вновь с удовлетворением наблюдаем, как тесно связаны между собой логика, арифметика и теория множеств, которые вместе являют собой мощный фундамент всей современной математики, а в широком смысле слова — и всей науки.

2.1.5 Доказательство теоремы Гудстейна

Итак, мы имеем теперь все необходимое, чтобы доказать теорему 1.4. Для этого достаточно показать, что приведенные в разделе 1.1.10 леммы 1.1 и 1.2 остаются справедливыми, если заменить число N на ω и воспользоваться свойствами арифметики ординалов.

Центральное место в доказательстве леммы 1.1 занимает неравенство (1.7), которое необходимо получить для случая $m = \omega$, а именно:

$$\left(\omega_n^{\check{\mathbf{S}}_{(s_j-1)}} + \dots + \omega_n^{\check{\mathbf{S}}_{(0)}} \right) (n-1) < \omega_n^{\check{\mathbf{S}}_{(s_j)}}, \quad (2.6)$$

где $\omega > s_j > \dots > s_0 > 0$.

Подчеркнем, что и здесь, и в определении функции f_n равенствами (1.5) важен порядок слагаемых — от большей степени ω к меньшей. Если порядок слагаемых нарушить, то все степени, предшествующие максимальной, исчезнут по свойствам сумм ординалов. Итак,

$$\begin{aligned} \left(\omega_n^{\check{\mathbf{S}}_{(s_j-1)}} + \dots + \omega_n^{\check{\mathbf{S}}_{(0)}} \right) (n-1) &\leq \\ &\leq \underbrace{\left(\omega_n^{\check{\mathbf{S}}_{(s_j-1)}} + \dots + \omega_n^{\check{\mathbf{S}}_{(s_j-1)}} \right)}_{s_j \text{ раз}} (n-1) = \\ &= \omega_n^{\check{\mathbf{S}}_{(s_j-1)}} \cdot s_j \cdot (n-1) < \omega_n^{\check{\mathbf{S}}_{(s_j-1)}} \omega = \omega_n^{\check{\mathbf{S}}_{(s_j-1)+1}} \leq \omega_n^{\check{\mathbf{S}}_{(s_j)}} \end{aligned}$$

⁴Кстати, это он ввел обозначение \forall для квантора всеобщности

Здесь мы воспользовались тем, что $\omega^\alpha \cdot k < \omega^\alpha \cdot \omega$ (свойство ОР4 — строгая монотонность умножения ординалов по второму множителю).

Отметим еще одно важное логическое место в доказательстве леммы 1.1. В нем шаг индукции основан на том, что если $T \geq n^{s_i}$, то $s_j < T$, и значит, к s_j можно применять предположение индукции. Но и для ординалов, получаемых с помощью функции $\overset{\omega}{S}_n$ это также справедливо, поскольку все они меньше ε_0 , т. е. первого ординала, для которого $\varepsilon_0 = \omega^{\varepsilon_0}$.

Итак, лемма 1.1 верна при $m = \omega$, т. е. $\overset{\omega}{S}_n(t)$ строго возрастает по $t < \omega$. Отсюда сразу же слeует лемма 1.2, которая утверждает, что величина $f_n = \overset{\omega}{S}_n(g_n)$, где g_n — числа Гудстейна, строго убывает с ростом n .

Но теперь ничто не ограничивает рост n , следовательно, f_n определена для всех $n < \omega$, а это означает, что при некотором n получится $f_n = 0$, что и доказывает окончательно теорему Гудстейна.

2.1.6 «Начальные» мультимножества

Здесь мы завершим разговор о «начальных» мультимножествах, т. е. таких мультимножествах, которые записываются конечной скобочной записью. Они были определены нами еще до аксиом ZF, в рамках исходной грамматики с добавлением правил G18 и G19 в подразделе 1.1.8.

Напомним, что скобочная запись мультимножества является одновременно и записью «начального» множества, поскольку при записи множества мы условились не принимать во внимание кратность равных элементов. Иначе говоря, $\{\{\}, \{\}\}$ — это и запись мультимножества, в котором единственный элемент $\{\}$ входит с кратностью 2, и запись множества, эквивалентная записи $\{\{\}\}$.

Для моделирования «начальных» мультимножеств в ZF нам достаточно слегка подправить определение универсумов мультимножеств, которые мы определили в разделе 1.4.

Положим по определению:

$$\begin{aligned} fmV_0 &= \{\} \\ fmV_{n+1} &= \{f : D \rightarrow \omega \setminus \{0\} \mid (D \subseteq fmV_n) \wedge (D - \text{конечное})\} \\ fmV_\omega &= \bigcup \{fmV_n \mid n < \omega\} \end{aligned}$$

Универсум fmV_ω — это подмножество универсума mV_ω^ω , и его элементы отличаются тем, что это функции, определенные на конечных множествах, т. е. все входящие в него мультимножества имеют не только конечную крат-

ность элементов, но и конечный носитель⁵.

Нетрудно проверить, что мультимножествам из универсума fmV_ω можно взаимно однозначно с помощью индукции по гамма-дереву скобочной записи сопоставить «начальные» мультимножества. При этом естественным образом определятся и введенные ранее операции и отношения на мультимножествах: \in^n , \subseteq , \cap , \cup , \otimes .

Кроме того, мультимножествам из универсума fmV_ω также взаимно однозначно ставится в соответствие Code^ω , который является не чем иным как канторовской нормальной формой для ординалов $< \varepsilon_0$.

Приведем несколько примеров:

$$\begin{aligned}\text{Code}^\omega(\emptyset) &= 0 \\ \text{Code}^\omega(\{0\}) &= 1 \\ \text{Code}^\omega(\{k \bullet 0\}) &= k \\ \text{натуральные числа на этом исчерпываются} \\ \text{Code}^\omega(\{1\}) &= \omega \\ \text{Code}^\omega(\{\{1\}\}) &= \omega^\omega \\ \text{Code}^\omega(\{n \bullet 0, m \bullet \{0, 0\}, \{1, 1\}\}) &= \omega^{\omega^2} + \omega^2 \cdot m + n\end{aligned}$$

Формальное определение ω -кода для начальных мультимножеств:

$$\begin{aligned}\text{Code}^\omega(\{\}) &= 0 \\ \text{Code}^\omega(f) &= \sum_{g \in \text{dom}(f)} \omega^{\text{Code}^\omega(g)} \cdot f(g),\end{aligned}$$

где суммирование происходит в порядке убывания кодов $\text{Code}^\omega(g)$.

В качестве развлечения мы предлагаем читателю построить мультимножество, соответствующее оглавлению данной книги, считая все подразделы третьего уровня (у которых нет своих подразделов) пустыми множествами, а затем написать его ω -код.

2.2 Кардинальная арифметика

Если до этого мы рассматривали арифметические действия над порядковыми типами, то теперь посвятим некоторое время арифметике кардинальных чисел, отвечающих за мощность множества. Мы уже видели расхождения арифметики бесконечных чисел с привычной со школы арифметикой, например, некоммутативность сложения и умножения, невозможность сокращения

⁵Носителем мультимножества принято называть множество, получаемое из мультимножества редуцированием всех кратностей до 1. В нашем случае носитель мультимножества $f : D \rightarrow \omega \setminus \{0\}$ — это $\text{dom}(f)$

справа. С другой стороны, практически без ограничений мы увидели теорему о разложении ординалов в сумму степеней по произвольному основанию, что бесспорно является сильным сходством двух арифметик. С кардиналами также будет ряд особенностей, но совсем другого характера. В некотором смысле их арифметика проще даже чем арифметика натуральных чисел постольку, поскольку и сам объект — кардинальное число — проще, чем ординал. В приложении приведена таблица В.2, где вкратце перечислены для сравнения основные свойства рассматриваемых арифметик.

Напомним, что символом \sim мы обозначаем отношение равномощности множеств: $a \sim b$, если существует биекция $f : a \leftrightarrow b$. Кроме того, была доказана основная теорема 1.22 о мощностях: если существуют инъекции $f : a \rightarrow b$ и $g : b \rightarrow a$, то $a \sim b$. Наконец, кардиналом называется всякий ординал τ , для которого $(\alpha < \tau) \rightarrow \neg(\alpha \sim \tau)$.

Если для обозначения ординалов используются, как правило, начальные буквы греческого алфавита, то для обозначения кардиналов — $\mu, \nu, \xi, \rho, \sigma, \tau, \chi$. Кроме того, часто бывает полезна нумерация с помощью алефов: \aleph_0, \aleph_1 и т.д., которая устанавливает порядковый изоморфизм между ординалами и кардиналами.

Сложение кардиналов определяется следующим образом. Пусть имеется последовательность кардиналов $\{\tau_x \mid x \in X\}$, проиндексированная произвольным множеством X . Тогда, если существует мощность прямой суммы $\sqcup_{x \in X} \tau_x$, то она называется суммой кардиналов τ_x и обозначается:

$$\sum_{x \in X} \tau_x \equiv \left\| \bigsqcup_{x \in X} \tau_x \right\|.$$

Отметим важное отличие суммы кардиналов от суммы ординалов: порядок слагаемых здесь не играет роли, и поэтому в качестве индексирующего множества может выступать не только ординал или кардинал, но вообще любое множество. С другой стороны, определение не гарантирует существования суммы для произвольной последовательности кардиналов.

Упражнение 2.9. Для того, чтобы проверить, что сумма кардиналов не зависит от порядка элементов индексирующего множества, достаточно показать, что между множествами $\sqcup_{x \in X} \tau_x$ и $\sqcup_{x \in X} \tau_{f(x)}$, где $f : X \leftrightarrow X$, существует биекция.

Свойства суммы кардиналов:

Упражнение 2.10.

CS1 если $X = \emptyset$, то $\sum_{x \in X} \tau_x = 0$;

CS2 $\kappa + 0 = \kappa = 0 + \kappa$ (нейтральный элемент по сложению);

CS3 $\kappa + \mu = \max\{\kappa, \mu\}$, если $(\kappa \geq \omega) \vee (\mu \geq \omega)$, кроме того, $\aleph_\alpha + \aleph_\beta = \aleph_{\max\{\alpha, \beta\}}$;

CS4 $\kappa + \mu = \mu + \kappa$ и вообще $\sum_{x \in X} \tau_x = \sum_{x \in X} \tau_{f(x)}$, если $f : X \leftrightarrow X$ (коммутативность);

CS5 $(\kappa + \mu) + \nu = \kappa + (\mu + \nu)$ и вообще $\sum_{i \in I} \sum_{x \in X_i} \tau_x = \sum_{x \in X} \tau_x$, где $X = \bigcup_{i \in I} X_i$ — разбиение⁶ X (ассоциативность);

CS6 если $\kappa \leq \mu$, то $\kappa + \nu \leq \mu + \nu$ и $\nu + \kappa \leq \nu + \mu$ (нестрогая монотонность);

CS7 строгая монотонность не всегда верна: $1 + \omega \not\leq 2 + \omega$ и $\omega + 1 \not\leq \omega + 2$;

CS8 сокращение на одинаковое слагаемое не всегда верно: $1 + \omega = 2 + \omega$, но $1 \neq 2$;

CS9 если $\kappa < \mu$ и $\mu \geq \omega$, то $\mu - \kappa = \mu$ (вычитание);

CS10 если $\tau \geq \omega$, $\tau_x \leq \tau$ и $\|X\| \leq \tau$, то $\sum_{x \in X} \tau_x \leq \tau$ (следует из CW9 ниже).

Вычитание кардиналов можно определить явно через теоретико-множественную операцию. Пусть $x = \mu \setminus \kappa$, где $\kappa < \tau$, а разность — это обычная разность множеств. Тогда множество x вполне упорядочено (как часть кардинала μ), а значит, имеет мощность, равную некоторому кардиналу δ , который по определению называется разностью $\mu - \kappa$. При этом ясно, что $\mu = \kappa + \delta$, и если $\mu \geq \omega$, то $\mu = \max\{\kappa, \delta\}$, откуда следует $\delta = \mu$, т. е. $\mu - \kappa = \mu$.

Умножение кардиналов определяется следующим образом. Пусть имеется последовательность кардиналов $\{\tau_x \mid x \in X\}$, проиндексированная произвольным множеством X . Тогда, если существует мощность прямого произведения $\prod_{x \in X} \tau_x$, то она называется произведением кардиналов τ_x и обозначается:

$$\prod_{x \in X} \tau_x \equiv \left\| \prod_{x \in X} \tau_x \right\|,$$

проще говоря, она обозначается точно так же, как прямое произведение множеств, так что по умолчанию произведение кардиналов мы будем считать именно кардиналом (результатом операции умножения на кардиналах), и лишь в исключительных случаях — прямым произведением кардиналов как множеств.

⁶Напомним, что $\{X_i \mid i \in I\}$ называется разбиением X , если $X = \bigcup_{i \in I} X_i$ и $(i \neq j) \rightarrow (X_i \cap X_j = \emptyset)$.

Свойства произведения кардиналов:

Упражнение
2.11.

CP1 если $X = \emptyset$, то $\prod_{x \in X} \tau_x = 1$;

CP2 $\kappa \cdot 0 = 0 = 0 \cdot \kappa$ (умножение на ноль);

CP3 $\kappa \cdot 1 = \kappa = 1 \cdot \kappa$ (нейтральный элемент по умножению);

CP4 $\kappa \cdot \mu = \max\{\kappa, \mu\}$, если $(\kappa \geq \omega) \vee (\mu \geq \omega)$, кроме того, $\aleph_\alpha \cdot \aleph_\beta = \aleph_{\max\{\alpha, \beta\}}$;

CP5 $\kappa \cdot \mu = \mu \cdot \kappa$ и вообще $\prod_{x \in X} \tau_x = \prod_{x \in X} \tau_{f(x)}$, если $f : X \leftrightarrow X$ (коммутативность);

CP6 $(\kappa \cdot \mu) \cdot \nu = \kappa \cdot (\mu \cdot \nu)$ и вообще $\prod_{i \in I} \prod_{x \in X_i} \tau_x = \prod_{x \in X} \tau_x$, где $X = \bigcup_{i \in I} X_i$ — разбиение X (ассоциативность);

CP7 если $\kappa \leq \mu$, то $\kappa \cdot \nu \leq \mu \cdot \nu$ и $\nu \cdot \kappa \leq \nu \cdot \mu$ (нестрогая монотонность);

CP8 строгая монотонность не всегда верна: $1 \cdot \omega \not\leq 2 \cdot \omega$ и $\omega \cdot 1 \not\leq \omega \cdot 2$;

CP9 сокращение на одинаковый множитель не всегда верно: $1 \cdot \omega = 2 \cdot \omega$, но $1 \neq 2$;

CP10 если $0 < \kappa < \mu$ и $\mu \geq \omega$, то $\mu / \kappa = \mu$ (деление);

CP11 если $\tau_x < \xi_x$ при всех $x \in X$, то $\sum_x \tau_x < \prod_x \xi_x$ (теорема Ю. Кёнига);

CP12 $(\kappa + \mu) \cdot \tau = \kappa \cdot \tau + \mu \cdot \tau$ (дистрибутивность).

Упражнение
2.12.
Докажите,
что $\|X\|$
существует.

Если кардинал μ можно представить как сумму $\sum_{x \in X} \kappa_x$, где все $\kappa_x = \kappa < \mu$, то частное μ / κ можно определить как $\|X\|$.

Возведение кардиналов в степень можно определить как через произведение одинаковых кардиналов, так и через мощность множества μ^κ , т. е. множества всех функций вида $f : \kappa \rightarrow \mu$. При этом мы вновь вынуждены принять по умолчанию соглашение, что под термом μ^κ мы прежде всего понимаем кардинальную степень κ кардинала μ , и лишь в исключительных случаях — множество функций.

Свойства степеней кардиналов:

Упражнение
2.13.

CW1 $\mu^0 = 1$, в частности, $0^0 = 1$;

CW2 $0^\kappa = 0$, если $\kappa > 1$;

CW3 $1^\kappa = 1$;

CW4 $\mu^{\kappa+\nu} = \mu^\kappa \cdot \mu^\nu$;

$$\text{CW5 } \mu^{\varkappa \cdot \nu} = (\mu^{\varkappa})^{\nu};$$

$$\text{CW6 } (\mu\nu)^{\varkappa} = \mu^{\varkappa} \cdot \nu^{\varkappa};$$

$$\text{CW7 } \text{если } 0 < \nu \text{ и } \varkappa \leq \mu, \text{ то } \nu^{\varkappa} \leq \mu^{\varkappa} \text{ (монотонность по показателю);}$$

$$\text{CW8 } \text{если } \varkappa \leq \mu, \text{ то } \varkappa^{\nu} \leq \mu^{\nu} \text{ (монотонность по основанию);}$$

$$\text{CW9 } \varkappa^2 = \varkappa \text{ и } \varkappa^n = \varkappa, \text{ если } n < \omega \leq \varkappa;$$

$$\text{CW10 } \text{если существует } \|\mathcal{P}(X)\|, \text{ то } \|\mathcal{P}(X)\| = 2^{\|\mathcal{X}\|};$$

$$\text{CW11 } \varkappa < 2^{\varkappa};$$

$$\text{CW12 } 2^{\varkappa} = \mu^{\varkappa}, \text{ если } \varkappa \geq \omega \text{ и } 2 \leq \mu \leq \varkappa.$$

Мы не будем подробно останавливаться на доказательстве всех перечисленных свойств арифметики кардиналов, отметим лишь, что центральным свойством здесь можно назвать CW9, которое впервые было доказано Гессенбергом, в книге [25] оно доказывается с помощью разложения бесконечного кардинала по степеням ω . Существует и еще один способ доказательства — трансфинитной индукцией по кардиналам. В основе индукции лежит утверждение $\omega^2 = \omega$, которое доказывается явным построением биекции между парами натуральных чисел (n, m) и натуральными числами k по следующей формуле: $k = n + (n + m + 1)(n + m)/2$ (канторовская диагональная нумерация).

Несомненно, более уместный в этой книге

Как видим, арифметика кардиналов очень похожа на арифметику обычных (натуральных) чисел, но все портят два свойства: $\varkappa + \mu = \max\{k, t\}$ и $\varkappa \cdot \mu = \max\{k, t\}$ в случае бесконечного \varkappa или μ . Эти свойства делают операции сложения и умножения неоднозначно обратимыми и лишают нас удовольствия строить арифметические разложения кардиналов, подобные канторовскому разложению ординалов и аналогичному разложению натуральных чисел.

С этой точки зрения арифметику ординалов следует считать более близкой к «истинной» арифметике, даже несмотря на ее «однобокость» (выполнение некоторых свойств только слева или только справа).

2.3 Немного теории чисел

Уделим некоторое внимание внутренней структуре ординала ω , т. е. натуральным числам. А именно, приведем несколько основных свойств и опишем ряд проблем, связанных с арифметикой.

О да! Нерешенные проблемы есть даже внутри ω !

Прежде всего, дадим определение: натуральное число называется **простым**, если оно имеет ровно 2 делителя — единицу и само себя. Нужно подчеркнуть, что единица не относится к простым числам, поскольку имеет ровно один делитель. Единицу неудобно относить к простым числам, т. к. она «испортила» бы единственность разложения любого числа по степеням простых, которое называется основной теоремой арифметики.

Теорема 2.6 (Основная теорема арифметики). *Любое натуральное число $n \geq 2$ единственным образом раскладывается в конечное произведение степеней простых:*

$$n = p_1^{\alpha_1} \dots p_k^{\alpha_k},$$

где $p_1 < \dots < p_k$ — простые числа, степени $\alpha_1, \dots, \alpha_k \in \omega \setminus \{0\}$

Доказательство данной теоремы не такое уж тривиальное, как может показаться, хотя в школьной математике эта теорема преподносится как очевидный факт. Мы его пропустим в этом разделе, поскольку оно требует доказательства еще некоторого ряда утверждений о делимости. Но в разделе 3.3 о гауссовых числах мы полностью докажем аналогичную теорему 3.7.

Основная теорема арифметики (в ее общем понимании, т. е. в различных числовых структурах) имеет колоссальное значение как внутри самой арифметики, так и для математики в целом.

Комментарий 2.

Хочу привести один неожиданный пример использования основной теоремы арифметики в своей практике программирования. Отладка программ, как известно, занимает чуть ли не вдвое больше времени, чем первичное написание кода. Так вот, предположим, что у нас в программе имеется точка ветвления вычислений — нужно вычислить величину A циклом в зависимости от выполнения одного из условий C_1, \dots, C_k (внутри цикла). Мы вычисляем значение A , но оно получилось явно не такое, как мы ожидали. Отладку нужно начать с того, чтобы определить, какой из кейсов C_1, \dots, C_k сработал и сколько раз (в скольких итерациях цикла). Первое, что приходит в голову — это задать целочисленный массив размерности k и наращивать в нем счетчики столько раз, сколько срабатывает тот или иной кейс. Решение простое, но требует дополнительной работы с массивом и его отображением в результатах выполнения кода, а это не всегда удобно, если работаешь с шаблонизатором.

Эlegantное решение заключается в том, чтобы установить переменную $P = 1$ и затем каждый раз, когда срабатывает кейс C_i , умножать P на i -ое простое число (2 — первое, 3 — второе, 5 — третье, 7 — четвертое, 11 — пятое и т. д.). В итоге получится некое довольно большое число P , но оно будет в себе содержать ровно столько же информации, сколько тот самый массив из простого решения. Все, что нам нужно после этого — это аккуратно вывести данное число на превью

шаблонизатора, а затем разложить его на простые множители любым онлайн-сервисом. Мы получим полный перечень сработавших кейсов (номера простых чисел) с указанием количества их срабатывания (степени простых чисел). При этом мы будем использовать всего лишь одну целочисленную переменную вместо работы с массивом.

Несомненно, яркой стороной арифметики натуральных чисел является арифметика вычетов (или остатков). Пусть у нас имеется число $m \geq 2$, тогда, в силу все той же теоремы 2.4 любое число n можно представить в виде $n = m \cdot k + d$, где $0 \leq d < m$. При этом d называется **остатком** от деления n/m (а k — [неполным] частным). Говорят, что два числа n и n' **сравнимы по модулю m** , если их остатки от деления на m совпадают. Этот факт записывается в виде:

$$n \equiv n' \pmod{m}.$$

Например, числа 52 и 122 сравнимы по модулю 10, поскольку оба имеют остаток 2 при делении на 10.

При нахождении остатка, какое бы большое ни было исходное число, нас не интересует его часть, делящаяся на модуль. Например, $365 \bmod 7$ сравнимо с 1, поскольку $365 = 350 + 14 + 1$. Именно поэтому если сегодня у нас на календаре понедельник, то ровно через год (если мы не переходим через високосный февраль) на календаре будет вторник (понедельник+1), а ровно через 4 года будет суббота, т. е. три года нам дадут +3, а високосный год даст +2, итого +5 (суббота = понедельник+5). Наконец, каждые 28 лет календарь полностью повторяется.

Отношение сравнимости по определенному модулю является отношением эквивалентности на множестве натуральных (и, вообще говоря, целых) чисел.

Кроме того, операции сложения и умножения не нарушают это отношение, т. е.

$$(a \equiv b) \wedge (c \equiv d) \rightarrow (a + c \equiv b + d) \wedge (ac \equiv bd) \wedge (a^n \equiv b^n).$$

Таким образом, можно корректно ввести операции сложения и умножения над классами эквивалентности, порожденными фиксированным модулем m . Для целых чисел такая структура имеет специальное обозначение $\mathbb{Z}/m\mathbb{Z}$.

Два числа n и m называются **взаимно простыми** ($n \perp m$ согласно Д. Кнуту [2]), если они не имеют общих делителей, кроме 1.⁷ Например, $15 \perp 28$. В то же время всегда $n \perp p$, если p — простое и $n \neq 0$. В случае

⁷Это равносильно тому, что разложения данных чисел по степеням простых не имеют общих простых чисел.

Не путать с
налоговыми
вычетами!

Упражнение
2.14.

$a \perp m$ можно выполнять сокращение на a по модулю m :

$$(a \perp m) \wedge (ab \equiv ac \pmod{m}) \rightarrow b \equiv c \pmod{m}$$

Из свойств сравнимости по модулю 9 (модулю 3) вытекает, например, что число n делится на 9 (на 3) тогда и только тогда, когда его сумма цифр делится на 9 (на 3):

$$c_0 10^n + c_1 10^{n-1} + \dots + c_n \equiv c_0 + c_1 + \dots + c_n \pmod{9(3)}.$$

Теорема 2.7 (Китайская теорема об остатках). Пусть m_1, \dots, m_k попарно взаимно простые и $m = m_1 \dots m_k$. Тогда

$$(n \equiv n' \pmod{m}) \leftrightarrow \forall j (n \equiv n' \pmod{m_j}).$$

Иначе говоря, вместо оперирования остатками по одному большому модулю можно оперировать наборами остатков по взаимно простым модулям, произведение которых равно исходному модулю. Приведем пример:

mod 6	mod 2	mod 3
0	0	0
1	1	1
2	0	2
3	1	0
4	0	1
5	1	2

Каждому остатку от деления на 6 взаимно однозначно сопоставляется пара остатков от деления на 2 и на 3. Более того, сложение и умножение остатков можно производить независимо по компонентам: например, $4 + 5 = 9 \equiv 3 \pmod{6}$, в то же время сложение пар $(0, 1) + (1, 2)$ по своим модулям дает пару остатков $(0+1, 1+2) \equiv (1, 0) \pmod{(2, 3)}$. Как видим, остатку 3 действительно соответствует пара $(1, 0)$.

Еще один замечательный факт.

Теорема 2.8. Пусть $n \perp k$ ($n, k > 0$), тогда остатки чисел $k, 2k, \dots, (n-1)k$ по модулю n попарно различны и отличны от 0, т. е. образуют множество $\{1, \dots, n-1\}$.

Доказательство. Предположим, что $ks \equiv kl \pmod{n}$, когда $0 < s < l < n$. Тогда $k(l-s) = nt$ при некотором t . Так как $n \perp k$, по основной теореме арифметики получаем, что k делит t , т. е. $k \leq t$. Очевидно, что $l-s < n$. Но тогда $k(l-s) < nt$. Противоречие.

Аналогично проверяется, что ни одно из чисел $k, 2k, \dots, (n-1)k$ не кратно n . □

И еще один.

Теорема 2.9 (Малая теорема Ферма). Пусть p — простое число и $n \perp p$, тогда

$$n^{p-1} \equiv 1 \pmod{p}.$$

Например, $10^{2016} \equiv 1 \pmod{2017}$, поскольку 2017 — простое число.

Малая теорема Ферма дает возможность получить обратный по умножению элемент в арифметике с остатками по простому модулю. Действительно, если у нас есть остаток d , то очевидно, что $d \in \{0, 1, \dots, p-1\}$ и $d \perp p$. Тогда через d^{-1} обозначим остаток $d^{p-2} \pmod{p}$. В силу МТФ имеем: $dd^{-1} = d^{p-1} \equiv 1 \pmod{p}$. На этом факте строится теория полей вычетов по простому модулю.

Приведем очень простое доказательство МТФ, основанное на биноме Ньютона.

- (1) Биномиальный коэффициент $\binom{p}{k}$ делится на p при $k = 1, \dots, p-1$.
- (2) Далее действуем по индукции:

$$1^p \equiv 1 \pmod{p}$$

$$2^p = 2 + \sum_{k=1}^{p-1} \binom{p}{k} \equiv 2 \pmod{p}$$

Предполагая, что $n^p \equiv n \pmod{p}$, получаем:

$$(n+1)^p = 1 + n^p + \sum_{k=1}^{p-1} \binom{p}{k} n^k \equiv n+1 \pmod{p}$$

- (3) Наконец, предполагая, что $n \perp p$, из полученного равенства $n^p \equiv n \pmod{p}$ сокращением на n получаем $n^{p-1} \equiv 1 \pmod{p}$.

Малая теорема Ферма используется в вероятностном **тесте Ферма** на простоту числа. Пусть у нас задано некоторое число N , о котором мы еще не знаем, простое оно или нет. Возьмем произвольное n , взаимно простое с N (можно брать n , не кратное N , ведь в случае простого N они будут взаимно просты). Вычислим выражение $n^{N-1} \pmod{N}$. Если N простое, то это выражение равно 1.

Можно провести серию таких экспериментов, случайно выбирая число n . Если в какой-то момент мы получим число, отличное от 1, то N точно не является простым. Если же достаточно длительная серия экспериментов (скажем, 100) дает 1, то число, скорее всего, является простым.

Числа N , не являющиеся простыми, но для которых выполняется тождество $n^{N-1} \equiv 1 \pmod{N}$ для всех $n \perp N$, называются числами Кармайкла.

Именно для таких чисел тест Ферма даст сбой. Но поскольку числа Кармайкла встречаются редко,⁸ тест с большой долей вероятности дает правильный ответ.

Оценка времени работы теста Ферма составляет $O(\ln N)$, в то время как алгоритм полного перебора занимает время порядка $O(\sqrt{N})$.

Более общим утверждением является теорема Эйлера. Пусть $\varphi(n)$ обозначает количество чисел меньших n и взаимно простых с ним:

$$\varphi(n) = |\{k | (0 \leq k < n) \wedge (k \perp n)\}|$$

Например, $\varphi(1) = 1$ (единица взаимно проста с любым натуральным числом, в том числе с нулем), $\varphi(p) = p - 1$, если p — простое, $\varphi(24) = 8$ и т.д.

Теорема 2.10 (Эйлера). Если $n \perp m$, то $n^{\varphi(m)} \equiv 1 \pmod{m}$.

Функция Эйлера из-за сложности ее вычисления играет ключевую роль в работе алгоритма RSA шифрования с открытым ключом, который используется для шифрования доступа в различных цифровых сервисах.

Банки,
E-mail,
Госуслуги и
т.д.

Схема работы RSA состоит в следующем. Пусть имеется два очень больших простых числа p и q (скажем, 1024-битовые, т. е. порядка 10^{300}). Положим $n = pq$, тогда $\varphi(n) = (p-1)(q-1)$. Выберем небольшое число $a < \varphi(n)$, взаимно простое с $\varphi(n)$ (обычно выбираются первые простые числа Ферма 17, 257, 65537, т. к. в этом случае очень эффективен алгоритм быстрого возведения в степень). Вычислим обратное к нему число $b = a^{-1}$ по модулю $\varphi(n)$. Пары чисел (a, n) и (b, n) называются, соответственно, открытым и закрытым ключами алгоритма RSA. Длина ключа при этом составляет порядка 2048 бит (более 600 десятичных цифр). Закрытый ключ держится в секрете.

Пусть теперь нам требуется зашифровать сообщение m (это некое число, т. к. в компьютере любой текст является числом), причем $m < n$ и $m \perp n$ (последнее означает, что $p \neq m \neq q$). Мы его преобразуем к числу $c = m^a \pmod{n}$. Абонент получает письмо в виде числа c и применяет к нему закрытый ключ:

$$c^b = m^{ab} = m^{1+k\varphi(n)} = m \pmod{n},$$

где мы воспользовались тем, что $m^{\varphi(n)} \equiv 1 \pmod{n}$.

Это — классическое описание алгоритма. В настоящее время он усилен добавлением шифрования сеансового ключа, который имеет ограниченное время жизни.

Если число b неизвестно абоненту, то подобрать его — сложная вычислительная задача. Поскольку $b = a^{-1} \pmod{\varphi(n)}$, основная вычислительная

⁸первые числа Кармайкла с 3 простыми сомножителями: 561, 1105, 1729, 2465, 2821, 6601, 8911, с 4 простыми множителями первое число Кармайкла равно 41041, с 5-ю — 825265.

проблема дешифровки сводится к поиску $\varphi(n)$ по известному числу n , что, в свою очередь, сводится к поиску простых делителей p и q (задача факторизации), а эта задача при известных на сегодняшний день алгоритмах факторизации имеет сложность порядка

$$\exp((c + o(1))k^{\frac{1}{3}} \log^{\frac{2}{3}} k), \quad c < 2,$$

где k — длина ключа, в нашем примере это 2048. В настоящее время ключи меньшей длины уже считаются ненадежными.

Коль скоро мы упомянули Пьера Ферма, грех обойти и его великую теорему, хотя она уже не относится к теории сравнимости.



Пьер Ферма

Теорема 2.11 (Великая теорема Ферма). *Уравнение $x^n + y^n = z^n$ не имеет решений в целых ненулевых числах при степени $n > 2$.*

Теорема была полностью доказана в 1994 году Эндрю Уайлсом, хотя результаты для $n = 3, 4, 5$ были известны еще в XIX веке (1659 — $n = 4$ (Ферма), 1770 — $n = 3$ (Эйлер), 1825 — $n = 5$ (Лежандр)).

Примечательно, что незадолго до доказательства Великой теоремы Ферма (в 1980-х) на свет родилась **АВС-гипотеза**, которая гласит: для любого $\varepsilon > 0$ существует $K(\varepsilon)$ такое, что если $a + b = c$ и натуральные a, b, c попарно взаимно просты, то $c \leq K(\varepsilon)(\text{rad}(abc))^{1+\varepsilon}$, где под радикалом понимается число abc , в разложении по степеням простых которого все степени простых заменены на единицу (т. е. $\text{rad}(p_1^{\alpha_1} \cdots p_k^{\alpha_k}) = p_1 \cdots p_k$).

Например, $4 + 9 = 13$, тогда

$$13 \leq \text{rad}^2(4 \cdot 9 \cdot 13) = \text{rad}^2(2^2 \cdot 3^2 \cdot 13) = (2 \cdot 3 \cdot 13)^2 \quad (\varepsilon = 1).$$

Из АВС-гипотезы следует теорема Ферма при достаточно больших n . Действительно, допустим, что $x^n + y^n = z^n$, причем x, y, z попарно взаимно просты. И пусть $\varepsilon = 1$. Тогда числа $a = x^n$, $b = y^n$, $c = z^n$ удовлетворяют условиям АВС-гипотезы, из которой следует, что $z^n \leq K(1)\text{rad}^2(x^n y^n z^n) = K(1)\text{rad}^2(xyz)$. Но $\text{rad}(xyz) \leq xyz \leq z^3$. Отсюда получаем, что $z^n \leq K(1)z^6$, т. е. $n \leq N$, где N определяется из условия $N > 6 + \ln K(1)$.

В 2017 году появилось 300-страничное доказательство АВС-гипотезы [67], принадлежащее японскому математику Синъити Мотидзуки. В 2018-м году Петер Шольце и Якоб Стикс заявили, что нашли неустранимую ошибку в этом доказательстве. Мотидзуки с ним не согласен, так что указанное доказательство пребывает в неопределенном статусе. Тем не менее, начало покорению этой математической вершины положено.

Сиранузы
здесь — от
Сиранузского
университе-
та, а не от
древних
греков.

Еще одна знаменитая проблема теории чисел, чем-то схожая с теоремой Гудстейна, носит название сиранузской проблемы или гипотезы Коллатца. Берем любое натуральное число n . Если оно четное, то делим на 2, а если нечетное, то умножаем на 3 и прибавляем 1 (получаем $3n + 1$), затем применяем данный алгоритм снова и снова. Например, для числа 3 имеем такую последовательность:

$$3 \rightarrow 10 \rightarrow 5 \rightarrow 16 \rightarrow 8 \rightarrow 4 \rightarrow 2 \rightarrow 1.$$

Для стартового числа $n = 27$ последовательность составляет 111 шагов и в максимуме достигает числа 9232.

В приложении (см. листинг C.1) приведен листинг программы на языке Python, которая вычисляет эту последовательность для любого стартового числа.⁹

Гипотеза заключается в том, что независимо от стартового числа n за конечное число шагов в итоге получится 1 (и цикл $4 \rightarrow 2 \rightarrow 1$). По сути это означает, что операция деления на 2 встречается в среднем в ≥ 1.6 раз чаще, чем операция умножения на 3, т. е. среди возникающих чисел $3n + 1$ довольно часто попадают кратные 4 и более высоким степеням двойки.

Многочисленные компьютерные вычисления до сих пор не смогли найти контрпример к этой гипотезе. Доказательство также не найдено. Существует оценочное доказательство, основанное на (нестрогом) предположении о вероятности возникновения четного числа в последовательности Коллатца. Это доказательство приведено в комментарии 23 на стр. 503.

2.4 Числовые структуры

В финале главы 1 мы обсуждали основные «архитектурные» приемы при строительстве математических объектов: 1) из основного множества создается либо прямое произведение (в том числе многократное), либо булеан, 2) затем в полученном множестве выбирается некоторое специальное подмножество (или несколько подмножеств). Операции 1) и 2) могут повторяться некоторое количество раз в произвольном порядке.

Такое строительство объектов можно определить скорее как искусственное, чем естественное, поскольку мы сознательно подгоняем архитектуру будущего множества под нужные нам свойства. Иначе говоря, зная заранее некоторые свойства объектов, мы строим для них *модель*.

В этом смысле натуральные числа и их продолжение — ординалы и кардиналы — скорее можно считать естественными (имманентными) объектами

⁹Обратите внимание, что сам итеративный шаг цикла `while` занимает 3 строчки, а все остальное — проверочный обвес.

в теории множеств, поскольку они получаются путем выделения из всей совокупности множеств вполне упорядоченного ряда объектов, конструкция которых гарантируется непосредственно аксиомой бесконечности.

Теперь, для того, чтобы построить какие-то дополнительные числовые структуры, нам потребуется их смоделировать, т. е. создать искусственно.

Для того, чтобы подчеркнуть разницу между естественными и искусственными объектами теории множеств, остановимся на простом примере. Следуя аксиоме бесконечности, мы определили натуральные числа по фон Нейману и ординал ω , который из них состоит, а также задали операции сложения и умножения, отталкиваясь всего лишь от основного кирпичика — пустого множества, и применяя теоретико-множественную операцию построения следующего множества $x \cup \{x\}$, которое назвали $x + 1$. Вся остальная арифметика — следствие аксиом и неограниченного прибавления единицы к пустому множеству!

Таким образом, мы определили натуральные числа как естественные объекты теории. Однако, как уже отмечалось, для арифметики существует самостоятельный формализм — аксиомы Пеано, моделью которого является ординал ω с операциями сложения и умножения ординалов. Если мы теперь скажем, что **натуральные числа** — это элементы множества \mathbb{N} с операциями $+$ и \cdot , удовлетворяющими аксиомам Пеано, то мы определим эти натуральные числа не как конкретные (естественные) объекты теории множеств, а как формальное понятие, имеющее конкретные реализации в теории множеств.

\mathbb{N} — понятие, ω — его реализация (модель).

Более точно разницу между естественными объектами ZF и прочими формальными конструкциями можно определить следующим образом: естественные объекты описываются (определяются) формулами на языке ZF без каких-либо ограничений, т. е. все они могут быть элиминированы до атомарных отношений \in и $=$, в то время как искусственные образования и определяются, и допускают описание своей внутренней структуры только на языке тех отношений и функций, которые заданы их базовыми формулами, т. е. *аксиоматикой*.

Иначе говоря, при описании ординала ω мы пользуемся всей мощью логического аппарата теории ZF, а при описании \mathbb{N} мы пользуемся строго аксиомами Пеано, которые устанавливают логические зависимости между операциями $+$, \cdot и переменными, обозначающими натуральные числа. Соответственно, при изучении ω мы вправе пользоваться фон Неймановской структурой натурального числа $n = \{i \mid i < n\}$, а в аксиоматике Пеано такой фокус не пройдет, поскольку она не использует символ (и отношение) принадлежности и, соответственно, терм-квантор тоже.

Тем не менее, метод определения новых объектов через список аксиом очень удобен и продуктивен. Таким способом мы вводим новые

Ординал —
тоже
понятие, а ω
— его
реализация.

Забавно, что
символ \in
ввел именно
Пеано ;)

понятия в математике, а в теории множеств находим их модель, чтобы удостовериться в непротиворечивости этого нового понятия.

Дальнейшее наше повествование будет отличаться от всего предыдущего как раз тем, что мы будем вводить новые понятия с помощью их свойств (аксиом) и строить их модели (реализации) в ZF. В результате мы усложним себе жизнь нетривиальной архитектурой моделирования в языке ZF, но при этом все вопросы о непротиворечивости вводимых понятий (и их аксиоматик) сведем к одному единственному вопросу о непротиворечивости самой теории множеств.

Заметим, что схожий подход к архитектуре математики предложен в свое время группой французских математиков под псевдонимом Н. Бурбаки¹⁰, которые ввели понятие математической структуры как некоторого абстрактного формализма с минимально-достаточной аксиоматикой, имеющего реализации в конкретных математических теориях. Ими выделены *порождающие структуры* (les structures-mères) для алгебры, топологии и упорядоченных множеств, а также дочерние *сложные* (multiples) *структуры*, которые могут включать аксиомы сразу двух порождающих структур и какие-то дополнительные, создавая тем самым все большее разнообразие математических теорий.

Мы не будем здесь столь категоричны и зафиксируем лишь отличие абстрактного понятия какой-либо структуры от ее реализации на некотором объекте теории ZF, и безотлагательно перейдем к их описанию и моделированию.

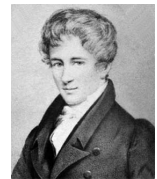
2.4.1 Группы, кольца, поля

Пусть имеется непустое множество G с заданной на его квадрате функцией $f : G \times G \rightarrow G$. Такую функцию, переводящую элементы G в элементы G , принято называть **операцией** и вместо термина $f(a, b)$ записывать afb (по аналогии с записью отношений). Пара (G, f) называется:

GR1 **группоидом**;

GR2 (левой/правой) **квазигруппой**, если возможно (левое/-правое) деление: уравнение $afx = b$ и/или $xfa = b$ имеет единственное решение;

GR3 **полугруппой**, если операция f ассоциативна: $(afb)fc = af(bfc)$;



Нильс
Хенрик
Абель

¹⁰В разное время в группу входили Клод Шевалле, Жан Дьедонне, Шолем Мандельброт, Андре Вейль, Лоран Шварц, Александр Гротендик и другие. Группе Бурбаки принадлежит введение в обиход символов \mathbb{N} , \mathbb{Z} , \mathbb{Q} , \mathbb{R} , \mathbb{C} . В настоящее время группа возобновила свою работу.

GR4 **моноидом**, если это полугруппа с *нейтральным элементом*, т.е. существует элемент $e \in G$ такой, что $efg = g = gfe$;

GR5 **группой**, если это моноид, в котором каждый элемент обратим: $\forall g \in G \exists g^{-1}: gfg^{-1} = g^{-1}fg = e$;

GR6 **абелевой группой**, если это группа, и операция f *коммутативна*: $g_1fg_2 = g_2fg_1$.

Схематично определение группы представлено на рис.2.3, где пунктиром показаны взаимозаменяемые стрелки: группа — это моноид, в котором все элементы обратимы, но группа — это еще и моноид, являющийся квазигруппой!

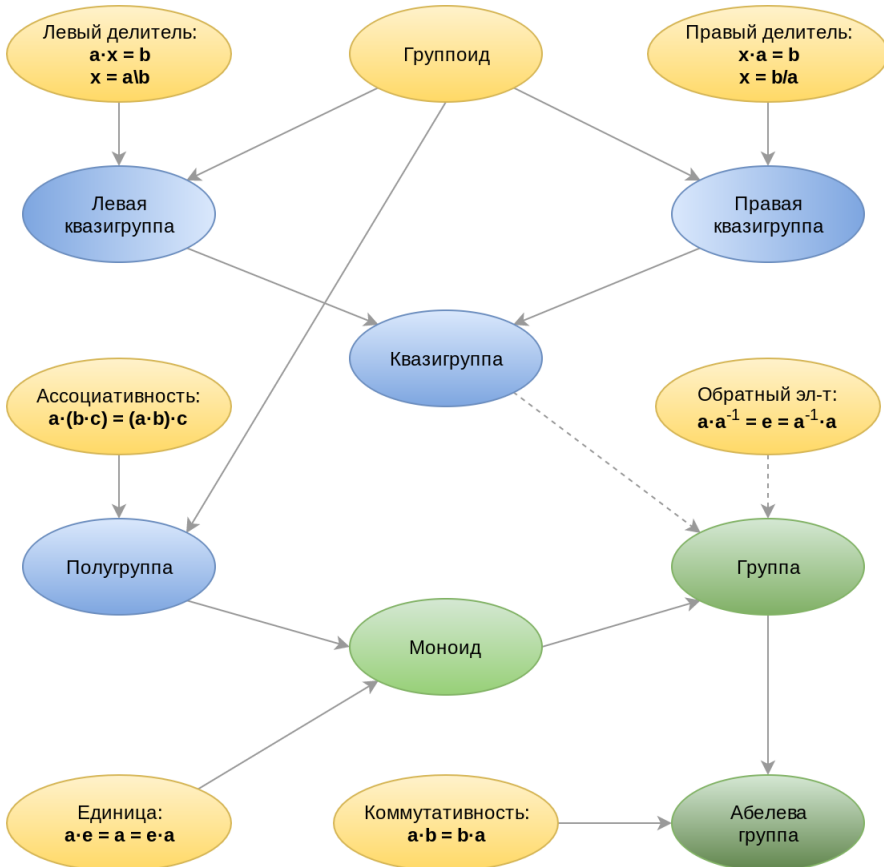


Рис. 2.3: Определение группы (пунктирные стрелки взаимозаменяемы).

Когда говорят о группе, то операцию f на ней обычно называют умножением и обозначают как обычное умножение (точкой или отсутствием символа

операции), если неизвестна природа этой операции. В случае абелевой группы принято считать, что операция называется сложением и обозначается '+'. Кроме того, в случае умножения нейтральный элемент обычно обозначается как 1, а в случае сложения — 0. Обратный элемент в случае сложения принято называть *противоположным*.

Лемма 2.1. *Определения группы как квазигруппы, являющейся моноидом, и как моноида с обратными элементами, эквивалентны.*

Доказательство. Пусть моноид (G, \cdot) является квазигруппой. Тогда уравнения $ax = e$ и $ya = e$ однозначно разрешимы (x является правым обратным элементом, а y — левым). Домножим первое равенство на y слева и воспользуемся вторым равенством, ассоциативностью и свойством единицы в моноиде:

$$(yax = ye) \rightarrow (ex = ye) \rightarrow (x = y),$$

т.е. оба обратных элемента совпали: $x = y = a^{-1}$.

С другой стороны, пусть (G, \cdot) — моноид с обратными элементами. Уравнение $ax = b$ имеет единственное решение $x = a^{-1}b$:

$$ax = a(a^{-1}b) = (aa^{-1})b = eb = b,$$

где мы воспользовались ассоциативностью и свойствами единицы в моноиде. Аналогично, разрешимо и уравнение $xa = b$ ($x = ba^{-1}$). \square

Упражнение
2.15.
Прежде
всего,
проверьте
замкнутость
операций на
этих мно-
жествах!

Несколько простых примеров из пройденного материала:

- множество ω с операцией сложения — коммутативный моноид;
- множество ω с операцией умножения — коммутативный моноид;
- ε_0 — некоммутативный моноид по сложению ординалов;
- ε_0 — некоммутативный моноид по умножению ординалов;
- любое множество кардиналов $< \tau$, где τ — бесконечный кардинал, — коммутативный моноид по сложению кардиналов;
- любое множество кардиналов $< \tau$, где τ — бесконечный кардинал, — коммутативный моноид по умножению кардиналов.

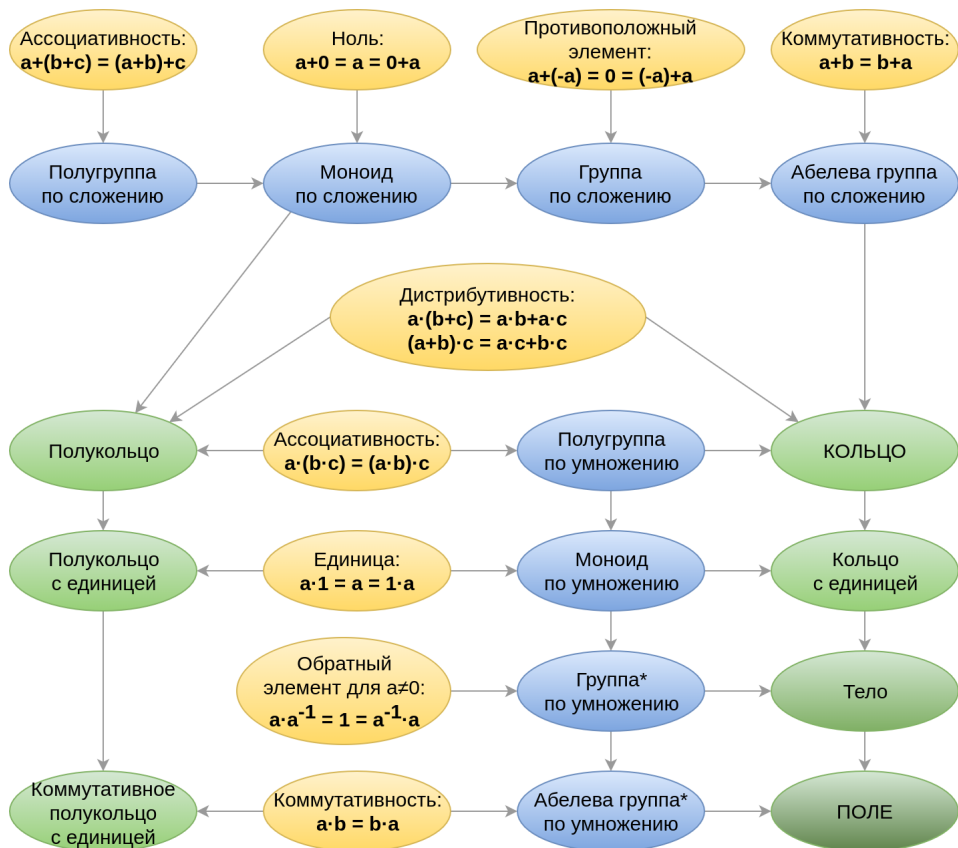


Рис. 2.4: Определение кольца и поля (группа* означает исключение нуля).

Пусть теперь на множестве G заданы две операции, которые мы условно обозначим '+' (которую назовем сложением) и '·' (которую назовем умножением).

Тройка $(G, +, \cdot)$ называется:

А тройку через пары определили или через функции? ;-)

КР1 **кольцом**, если (1) $(G, +)$ — абелева группа, (2) (G, \cdot) — полугруппа, (3) выполнен закон дистрибутивности: $a(b+c) = ab+ac$, $(a+b)c = ac+bc$;

КР2 **кольцом с единицей**, если $(G, +, \cdot)$ — кольцо и (G, \cdot) — моноид;

КР3 **целостным кольцом**, если $(G, +, \cdot)$ — кольцо, в котором нет делителей нуля: $(ab = 0) \rightarrow (a = 0) \vee (b = 0)$;

КР4 **коммутативным кольцом**, если $(G, +, \cdot)$ — кольцо, и операция умножения коммутативна;

КР5 **телом**, если $(G, +, \cdot)$ — кольцо с единицей, в котором ненулевые элементы образуют группу по умножению ((G^*, \cdot) — группа);

КР6 **полем**, если $(G, +, \cdot)$ — коммутативное кольцо, являющееся телом;

КР7 **полукольцом**, если (1) $(G, +)$ — коммутативный моноид, (2) (G, \cdot) — полугруппа, (3) выполнен закон дистрибутивности и (4) $a \cdot 0 = 0 = 0 \cdot a$ (мультипликативное свойство нуля).

Мультипликативное свойство нуля опускается в определении кольца, так как там оно следует из других условий, но в полукольце его приходится добавлять. Отличие полукольца от кольца состоит лишь в том, что по сложению полукольцо образует только коммутативный моноид, а не коммутативную группу.

К полукольцу можно добавлять определения недостающих свойств операции умножения: коммутативное полукольцо, полукольцо с единицей.

Упомянутое выше обозначение G^* используется для обозначения множества $G \setminus D$, где D — множество, включающее ноль и все необратимые (относительно умножения) элементы G . В определении тела прямо требуется, чтобы $G^* = G \setminus \{0\}$, т.е. чтобы все ненулевые элементы были обратимы.

Примеры:

- множество ω с операциями сложения и умножения — коммутативное полукольцо с единицей;
- ординал ε_0 с операциями сложения и умножения ординалов не является полукольцом из-за некоммутативности сложения;
- любое множество кардиналов $< \tau$, где τ — бесконечный кардинал, — коммутативное полукольцо с единицей.

Вспоминая теперь, что на множествах еще можно задавать различные порядки, предположим, что на множестве G , помимо указанных операций, задано отношение порядка \leq . Данный порядок **согласован с операцией сложения** '+', если функция $f(a, b) = a + b$ монотонна по обоим аргументам, т.е. отношение $a \leq c$ влечет $a + b \leq c + b$ (лево-упорядочение сложения) и $b + a \leq b + c$ (право-упорядочение сложения). Если при этом порядок \leq является линейным, то говорят, что группа $(G, +)$ является линейно-упорядоченной группой.

Порядок \leq **согласован с операцией умножения** \cdot , если $0 \leq ab$ при $(0 \leq a) \wedge (0 \leq b)$.

Кольцо $(G, +, \cdot)$, в котором линейный порядок \leq согласован с операциями сложения и умножения, называется **упорядоченным кольцом**. Если при этом кольцо является полем, то оно называется **упорядоченным полем**.

Упорядоченное поле (кольцо) удовлетворяет **аксиоме Архимеда**, если: для любых $a > b > 0$ существует такое натуральное n , что

$$a < \underbrace{b + b + \dots + b}_{n \text{ раз}}$$

В частности, в кольце с единицей достаточно проверить, что для любого числа a найдется такое натуральное n , что $-n < a < n$.

Упорядоченное кольцо (поле) называется **непрерывным**, если заданный на нем линейный порядок непрерывен (см. теорему 1.13)

Теорема 2.12. *Непрерывное упорядоченное поле является архимедовым.*

На самом деле, утверждение этой теоремы можно обратить, если в дополнение к архимедовости упорядоченного поля F потребовать одно из условий:

- (принцип вложенных отрезков) любая последовательность вложенных отрезков $[a_0; b_0] \supseteq [a_1; b_1] \supseteq \dots \supseteq [a_n; b_n] \supseteq \dots$ имеет непустое пересечение (предел);
- (аксиома полноты по Гильберту) F нельзя расширить до архимедова упорядоченного поля F^* так, чтобы индуцированные из F^* в F операции сложения, умножения и линейный порядок совпадали с исходными на F .

Далее мы займемся тем, что определим ряд конкретных числовых структур и рассмотрим их модели (реализации) в ZF.

2.4.2 Целые числа

Определение целых чисел, в общем, не вызывает затруднений при любом подходе к их конструированию. С нашей точки зрения проще всего использовать геометрический подход, основанный на декартовой целочисленной решетке. Смысл его заключается в том, чтобы с помощью уже имеющегося множества ω построить линии параллельных прямых на целочисленной декартовой плоскости (в ее первом квадранте, т. е. в области натуральных чисел) и рассматривать их как реализацию целых чисел.

Итак, рассмотрим базовое множество $\omega \times \omega$, на котором введем отношение эквивалентности $(a, b) \sim (a', b')$, если $a + b' = a' + b$ (если бы мы определили операцию вычитания, это равенство можно было бы переписать так: $a - b = a' - b'$). Класс эквивалентности задается либо уравнением $y + k = x$, либо уравнением $y = x + k$, где $k = |a - b| \in \omega$ является параметром, определяющим класс. Классы с первым типом уравнения мы назовем **неотрицательными** целыми числами, со вторым — **неположительными**. Класс с уравнением

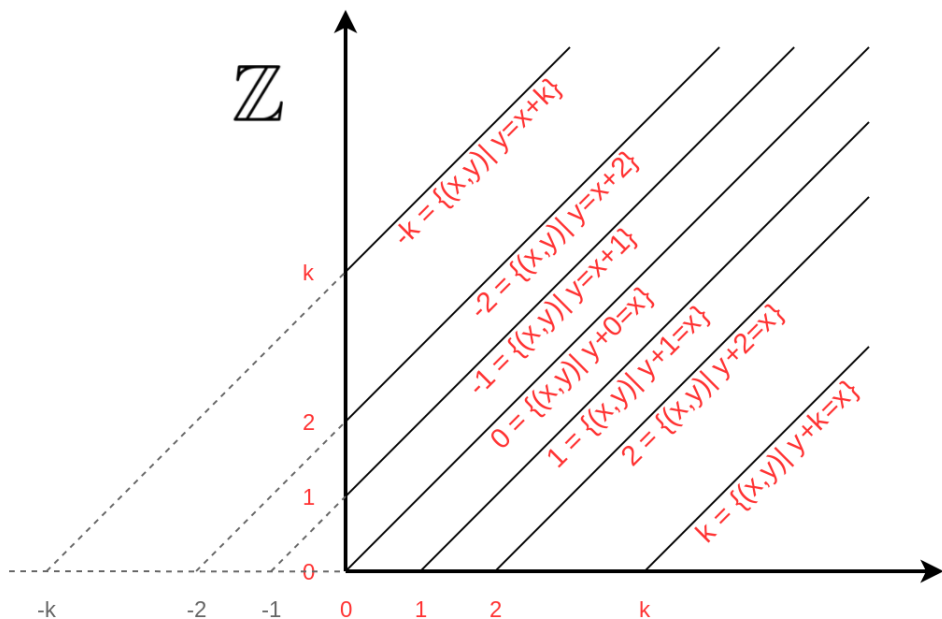


Рис. 2.5: Модель целых чисел.

$y = x$ соответствует нулю (рис. 2.5). Для краткости класс $[(a, b)]_{\sim}$ будем записывать $[(a, b)]$. Легко видеть, что $[(a, b)] = \{(x, y) \mid y + a = x + b\}$.

Заметим, что определенные нами классы эквивалентности на $\omega \times \omega$ представляют собой параллельные прямые под углом $\pi/4$ на целочисленной решетке (т. е. не прямые в полном смысле этого слова, а только их точки с натуральными координатами), причем те, которые расположены ниже прямой $y = x$, соответствуют положительным числам, а те, что выше, — отрицательным. Если мысленно достроить квадрат $\omega \times \omega$ влево и вниз до полной привычной нам плоскости, то эти прямые будут «отсекать» на оси Ox все целые числа в полном соответствии с их определением.

Остается корректно определить операции $+$ и \cdot на факторе $(\omega \times \omega)/_{\sim}$. Положим:

$$\begin{aligned} [(a, b)] + [(a', b')] &= [(a + a', b + b')] \\ [(a, b)] \cdot [(a', b')] &= [(aa' + bb', ab' + ba')] \end{aligned}$$

Упражнение 2.16. | Докажите, что данная операция определена корректно, т. е. равенства сохраняются при замене указанных пар натуральных чисел на эквивалентные.

Докажите, что $((\omega \times \omega)/\sim, +, \cdot)$ является коммутативным | Упражнение
кольцом с 1. 2.17.

Пусть $[(x, y)] < [(x', y')]$, если $[(x', y')] = [(x, y)] + [(n, 0)]$ при некотором $n \in \omega \setminus \{0\}$.

Докажите, что отношение $<$ является отношением линейного | Упражнение
порядка, согласованным с операциями сложения и умножения. 2.18.

Можно придумать еще более простую модель целых чисел, взяв в каждом из построенных классов по одной точке: $(n, 0)$ — положительные числа, $(0, m)$ — отрицательные, $(0, 0)$ — ноль. Это — те самые точки, которые попадают на координатные оси (рис. 2.5). Операции $+$ и \cdot индуцируются на них следующим образом:

$$\begin{aligned}(n, 0) + (m, 0) &= (n + m, 0) \\ (0, n) + (0, m) &= (0, n + m) \\ (n, 0) + (0, m) &= (n - m, 0), \text{ если } n \geq m \\ (n, 0) + (0, m) &= (0, m - n), \text{ если } n \leq m \\ (n, 0) \cdot (m, 0) &= (nm, 0) \\ (0, n) \cdot (0, m) &= (nm, 0) \\ (n, 0) \cdot (0, m) &= (0, nm)\end{aligned}$$

Отношение порядка:

$$\begin{aligned}(n, 0) &< (m, 0), \text{ если } n < m \\ (0, n) &< (0, m), \text{ если } n > m \\ (0, m) &< (n, 0) \text{ при любых } m, n, \text{ одновременно не равных } 0.\end{aligned}$$

Обратите
внимание на
смену знака
у отрица-
тельных
чисел!

Базовым множеством в этом случае будет $\omega \times \{0\} \cup \{0\} \times \omega$ без всякой факторизации, т. е. два «луча» натуральных чисел, «склеенных» общим нулем.

Обе приведенные модели целых чисел изоморфны по обеим | Упражнение
операциям и отношению $<$. 2.19.

В обоих случаях кольцо целых чисел включает в себя натуральные числа (не множество ω , как таковое, а понятие натурального числа, т. е. подмножество, изоморфное ω по операциям и порядку). Нетрудно видеть, что ими являются неотрицательные целые числа. При этом, в кольце целых чисел невозможно выделить собственное подкольцо, содержащее натуральные числа. В общей алгебре данное утверждение принимают за определение кольца целых чисел. Точнее, **кольцом целых чисел** называется минимальное кольцо, содержащее натуральные числа. При этом минимальность означает то, что у данного кольца нет собственных подколец с теми же свойствами.

Покажите
изоморфизм.

Кроме того, кольцо целых чисел может быть определено, как *минимальное нетривиальное линейно упорядоченное кольцо с единицей*. При этом нетривиальность подразумевает, что в этом кольце с единицей более одного элемента (т. е. единица отлична от нуля).

Кольцо целых чисел обозначается \mathbb{Z} .

Построенные нами кольца $((\omega \times \omega) / \sim, +, \cdot)$ и $(\omega \times \{0\} \cup \{0\} \times \omega, +, \cdot)$ являются моделями \mathbb{Z} в ZF, так же, как ω является моделью \mathbb{N} .

Различные соотношения с числами вроде $\mathbb{N} \subset \mathbb{Z}$ следует понимать не в теоретико-множественном смысле, а в алгебраическом. На языке теории множеств это означает, что для конкретной задачи мы выбрали некоторую конкретную модель \mathbb{Z} , и в ней выделили подмножество, изоморфное ω , которое временно стали обозначать \mathbb{N} . Это правило работает и для всех вышестоящих числовых структур: как только мы выходим за пределы \mathbb{Z} , мы начинаем работать в какой-то более обширной модели, содержащей в себе как модель \mathbb{Z} , так и модель \mathbb{N} в качестве подмножеств с индуцированными операциями и отношениями.

Использование конкретной модели сразу предполагает переход от языка алгебры к языку теории множеств, и результаты, полученные для модели на языке теории множеств, в общем случае не являются алгебраическими, т. к. могут не выполняться в других моделях.

Так, все функции, принимающие значения в \mathbb{Z} , являются функциями, действующими в некоторую модель \mathbb{Z} , но при этом знание теоретико-множественной структуры этой модели игнорируется. Например, если мы говорим о функции $f : M \rightarrow \mathbb{Z}$, где $\mathbb{Z} = \omega \times \{0\} \cup \{0\} \times \omega$, то для доказательства алгебраических фактов мы не имеем права пользоваться тем обстоятельством, что $f(x)$ — это некая пара ординалов $(n, 0)$ или $(0, m)$, и брать ее левую или правую проекцию, либо же утверждать, что множество неотрицательных целых чисел является транзитивным или прогрессивным, хотя оно и изоморфно ω . Однако, мы вправе пользоваться аксиоматикой целых чисел, т. е. операциями сложения и умножения, отношением $<$ и теми соотношениями между ними, которые заложены в определении понятия \mathbb{Z} .

В этом смысле теория множеств выглядит сухим юридическим языком на фоне обычного (разговорного) языка общей алгебры. С одной стороны, это может показаться неудобным (поэтому ни алгебра, ни топология, ни какая-либо еще математическая наука никогда не станет частью теории множеств), с другой стороны, приземление на язык теории множеств является мощным формализующим методом, позволяющим строго обосновать непротиворечивость алгебраических структур и «высветить» в них некоторые новые свойства.

Моделирование \mathbb{Z} с помощью двух лучей $\omega \times \{0\} \cup \{0\} \times \omega$ — это взятие клона ω и приписывание его элементам знака «минус», что позволяет легко определить и цифровые обозначения для всех целых чисел. Так, положим,

что если число n записывается цифрами $c_1 \dots c_k$, то целое число $(n, 0)$ в данной модели \mathbb{Z} записывается тем же набором цифр, а в целое число $(0, n)$ записывается как $-c_1 \dots c_k$.

Кроме того, если в схеме с двумя лучами вместо ординала ω подставить какой-либо бесконечный ординал, то, в принципе, ничто не мешает ввести определение отрицательных ординалов. Сложности начнутся тогда, когда мы попытаемся корректно определить операции сложения и умножения с ординалами разных знаков.

Например, пусть $\alpha \geq \beta$, тогда под $\alpha + (-\beta)$ можно понимать порядковый тип $\alpha \setminus \beta$. Однако, такое определение приводит к следующему парадоксу: $(\omega + (-2)) + 2 = \omega + 2 > \omega$. Иначе говоря, операция '+' неассоциативна, и на бесконечных ординалах мы не получим не только группу по сложению, но даже полугруппу.

Однако, мы знаем, что по свойству ординалов OS8 вычитание слева определяется однозначно, т. е. запись вида $(-\beta) + \alpha$, где «отрицательный» ординал всегда стоит слева от положительного, определяется корректно, и можно построить квазигруппу по сложению. Но и в этом случае ассоциативность с участием отрицательных ординалов не выполняется. Кроме того, даже на положительных ординалах операция '+' некоммукативна, а значит, о кольце ординалов говорить не приходится.

Казалось бы, нас может обнадежить коммутативность операций сложения и умножения кардиналов. Положим по определению, что $-\tau$ обладает свойствами: $-\tau + \tau = 0$, $\mu - \tau = \mu$, если $\mu > \tau$, и $\mu - \tau = -\tau$, если $\mu < \tau$.

Пусть теперь $\tau \leq \mu$ и $\mu \geq \omega$. Тогда, с одной стороны, $\tau + (\mu - \mu) = \tau$, с другой стороны, $(\tau + \mu) - \mu = 0$. Следовательно, ассоциативность на «отрицательных» кардиналах также нарушается.

Таким образом, расширение \mathbb{Z} «в стороны» с сохранением нативных операций над ординалами и/или кардиналами становится проблематичным.

Выходом может быть специально сконструированная на основе канторовской нормальной формы арифметика ординалов, меньших ε_0 . Действительно, любой ординал $\alpha < \varepsilon_0$ однозначно представляется в виде многочлена $\omega^{\gamma_1} k_1 + \dots + \omega^{\gamma_n} k_n$ от переменной ω , где ординалы $\gamma_1 > \dots > \gamma_n > 0$, коэффициенты $k_i < \omega$. Получается, что всякий ординал $\alpha < \varepsilon_0$ можно взаимно однозначно представить функцией $f : \varepsilon_0 \rightarrow \omega$ с конечным носителем (соответствие $\gamma_i \mapsto k_i$). Это — аналог векторного пространства, и поэтому арифметику на данных функциях можно определять как арифметику на векторах или многочленах. Чтобы получить отрицательные элементы, нужно всего лишь область значений таких функций расширить до \mathbb{Z} , т. е. позволить коэффициентам k_i принимать отрицательные целые значения. Получаемые таким способом операции сложения и умножения принято называть **натуральной суммой и натуральным произведением**. Их определение восходит еще к работам Гессенберга (1906) и Серпинского (1958) [39].

Пример натуральных операций:

$$\begin{aligned}\omega^3 2 - (\omega^3 3 + \omega^2 3) &= \omega^3 (2 - 3) - \omega^2 3 = -\omega^3 - \omega^2 3 \\ (\omega^\omega 2 - \omega)(\omega^{\omega^2} 3 + \omega^2) &= \omega^{\omega^2 + \omega} 6 + \omega^{\omega + 2} 2 - \omega^{\omega^2 + 1} 3 - \omega^3\end{aligned}$$

Нужно отметить, что при натуральном умножении степени сомножителей складывают в порядке убывания этих степеней, т. е. по правилам натуральной суммы.

Упражнение
2.20.
Проверьте
свойства.

Натуральное сложение и натуральное умножение коммутативны, ассоциативны и подчиняются дистрибутивному закону. С использованием отрицательных коэффициентов мы получаем коммутативное кольцо с единицей. Кроме того, индуцируя естественным способом линейный порядок на отрицательные ординалы, мы, таким образом, получаем линейно упорядоченное кольцо с единицей, в котором порядковым типом множества положительных элементов будет ординал ε_0 .

Натуральные операции отличаются от обычных операций над ординалами, например, обычная сумма $(\omega^2 + \omega + 1) + (\omega^3 + \omega) = \omega^3 + \omega$, в то же время натуральная сумма дает результат $\omega^3 + \omega^2 + \omega 2 + 1$.

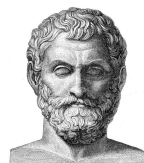
Расширение кольца целых чисел «в стороны» с помощью бесконечных ординалов и специальных (формальных) операций сложения и умножения над ними является иллюстрацией так называемых **сюрреальных чисел**, описанных и изученных Джоном Конвеем [30], Дональдом Кнудом [22] и Мартином Крускалем.

2.4.3 Рациональные числа

Рациональные числа, т. е. дроби вида $\frac{a}{b}$, где a является целым числом, b — положительным целым, могут быть построены аналогичным способом, хоть и более запутанным. Для этого достаточно вспомнить теорему Фалеса, позволяющую отсекавать равные части на одной стороне угла с помощью параллельных прямых, проходящих через точки на другой стороне угла, отложенные с помощью какой-либо единицы длины.

Пусть нам требуется построить число $\frac{1}{p}$, тогда отложим на оси Oy числа $1, 2, \dots, p$, а на оси Ox — единицу. После чего соединим точки $(0, p)$ и $(1, 0)$ прямой, а затем проведем параллельные ей прямые через точки $(0, 1)$, $(0, 2)$ и т. д. Эти прямые отсекут на оси Ox деления с шагом $\frac{1}{p}$ (рис. 2.6).

Таким образом, положительное рациональное число $\frac{k}{p}$ может быть задано как прямая вида $l(k, p) \Rightarrow \{(x, y) \mid y + xp = k\}$, где по-прежнему $x, y \in \omega$. На этот раз мы не задаем отношение эквивалентности на точках



Фалес
Милетский

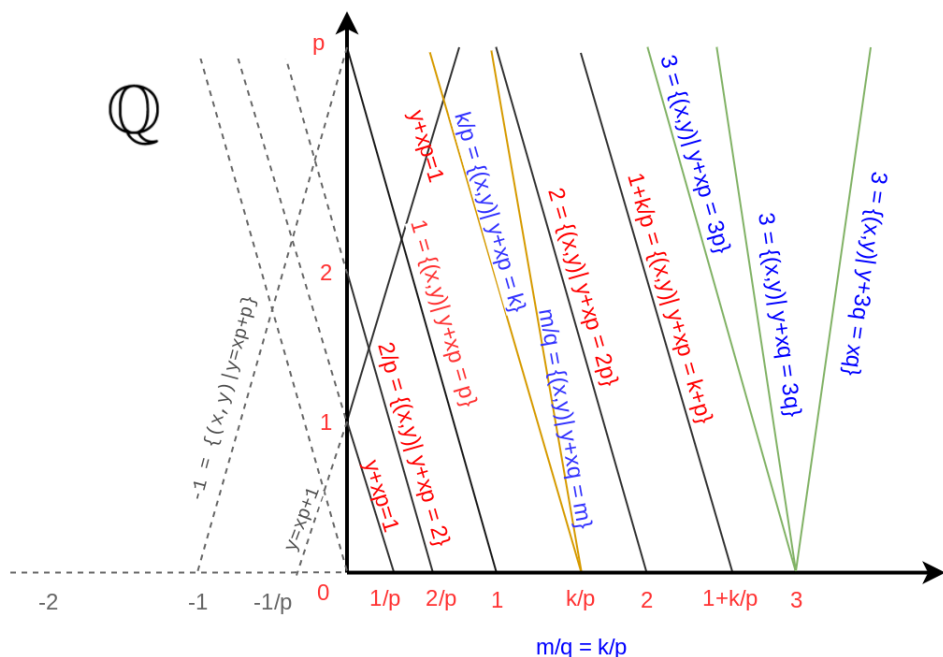


Рис. 2.6: Модель рациональных чисел.

плоскости $\omega \times \omega$, а сразу рассматриваем все возможные прямые $l(k, p)$ как начальные объекты строительства.

Однако, для прямых мы введем отношение эквивалентности следующим образом:

$$l(k, p) \sim l(m, q) \Rightarrow kq = mp, \quad (2.7)$$

т. е. дроби $\frac{k}{p}$ и $\frac{m}{q}$ эквивалентны, когда $kq = mp$.

Данное свойство прямых геометрически выражается следующим образом: они проходят через одну и ту же точку на прямой Ox . На рис. 2.6 эквивалентными являются пара прямых желтого цвета и пара прямых зеленого цвета, отвечающие рациональным числам $\frac{k}{p}$ и $\frac{3}{1}$, соответственно. Таким образом, класс эквивалентных прямых (напомним, что это не обычные прямые, а только лишь точки с целочисленными координатами, лежащие на данных прямых) $[l(k, p)]$ геометрически представляет собой пучок прямых с общей точкой пересечения $(k/p, 0)$. К этим же прямым можно присоединить прямые вида $l'(k, p) = \{(x, y) | y + k = xp\}$, которые симметричны $l(k, p)$ относительно вертикальной оси $x = k/p$ (на рис. 2.6 это третья зеленая прямая). Положим по определению, что $l(k, p) \sim l'(k, p)$. Если считать k, p целыми, а не натуральными числами, то отношение эквивалентности следует просто из того,

что прямая $l'(k, p)$ задается уравнением $y + x(-p) = -k$, а дроби $\frac{-k}{-p}$ и $\frac{k}{p}$, очевидно, эквивалентны.

Итак, мы взяли множество $\omega \times \omega$, выделили в нем подмножества специального вида (прямые $l(k, p)$ и $l'(k, p)$), которые являются элементами $\mathcal{P}(\omega \times \omega)$, т.е. выделили подмножество $L = \{l(k, p), l'(k, p)\} \subset \mathcal{P}(\omega \times \omega)$, и затем ввели отношение эквивалентности на L . Фактор-множество L/\sim представляет собой интерпретацию множества неотрицательных рациональных чисел.

Отрицательные рациональные числа задаются прямыми вида $t'(k, p) = \{(x, y) \mid y = xp + k\}$ и им эквивалентными в смысле соотношения (2.7). Отметим, что эти прямые имеют такой же (положительный) наклон, как и прямые $l'(k, p)$, поэтому они попадают в первый квадрант плоскости, что избавляет нас от необходимости использовать отрицательные целые числа при определении отрицательных рациональных дробей (на рисунке 2.6 это прямые $y = xp + p$ и $y = xp + 1$). Дополнительные к ним эквивалентные прямые вида $t(k, p) = \{(x, y) \mid y + xp + k = 0\}$ уже не пересекают первый квадрант, поэтому мы исключаем их из нашей модели. Обозначим через T множество всех прямых вида $t'(k, p)$, тогда фактор-множество T/\sim представляет собой интерпретацию множества неположительных рациональных чисел.

Упражнение
2.21.
Почему они
только экви-
валентны,
но не равны?
Докажите.

На самом деле, мы всегда можем пользоваться только «штрихованными» прямыми, и в этом случае множество прямых $L \cup T$ будет содержать в себе прямые вида $l'(k, 1)$ и $t'(k, 1)$, а это не что иное, как классы эквивалентности, определяющие целые числа в рассмотренной выше модели целых чисел. В частности, нулю соответствуют прямые $l'(0, 1) = t'(0, 1)$ и им эквивалентные $l'(0, p) = t'(0, q)$. Единице соответствуют прямые $l'(k, k)$ и им эквивалентные.

Таким образом, введенные нами построения сохраняют преемственность архитектуры моделирования при переходе от целых чисел к рациональным.

Операции сложения и умножения для рациональных чисел вводятся следующим образом:

$$\begin{aligned} [l(k, p)] + [l(m, q)] &= [l(kq + mp, pq)] \\ [l(k, p)] + [t(m, q)] &= [l(kq - mp, pq)], \text{ если } kq \geq mp \\ [l(k, p)] + [t(m, q)] &= [t(kq - mp, pq)], \text{ если } kq \leq mp \\ [l(k, p)] \cdot [l(m, q)] &= [l(km, pq)] \\ [l(k, p)] \cdot [t(m, q)] &= [t(km, pq)] \\ [t(k, p)] \cdot [t(m, q)] &= [l(km, pq)] \end{aligned}$$

Упражнение
2.22.

Проверьте, что операции определены корректно, т.е. замена прямых на эквивалентные сохраняет равенства.

Выразите операции сложения и умножения через операции над парами (x, y) , являющимися элементами прямых $l(k, p)$ и $l(m, q)$. Сравните с определением операций для целых чисел. Упражнение 2.23.

Докажите, что $(L/\sim \cup T/\sim, +, \cdot)$ является полем.

Упражнение 2.24.

Данные утверждения, на самом деле, являются нетривиальными фактами в теоретико-множественной (или геометрической) модели рациональных чисел, однако же это теоремы ZF, т.е. для получения всех свойств рациональных чисел нам не требуется привлекать дополнительные аксиомы или пользоваться чем-то, кроме множества ω и стандартных (архетипичных) операций над множествами.

Положим $[l(k, p)] < [l(m, q)]$, если $[l(m, q)] = [l(k, p)] + [l(k', p')]$ при некоторых k', p' .

Докажите, что отношение $<$ является линейным порядком, а также что $(L/\sim \cup T/\sim, <)$ является плотным л.у.м., т.е. между любыми двумя различными рациональными числами есть третье. Упражнение 2.25.

Мы могли бы быть проще и вместо прямых $l(k, p)$ рассматривать точки (k, p) , определяя на них то же самое отношение эквивалентности. В этом случае классом эквивалентности был бы не пучок эквивалентных прямых, а прямая вида $\{yk = xp\}$. Все такие прямые проходят через начало координат, причем нулю соответствует ось Oy , единице — прямая $\{y = x\}$, положительным числам — прямые из первого и третьего квадранта плоскости, отрицательным — прямые из 2-го и 4-го квадрантов, противоположные по сложению прямые симметричны относительно Oy , обратные по умножению прямые симметричны относительно прямой $y = x$ (см. рис. 2.7). Такое представление, безусловно, проще и арифметически, и геометрически, но оно никак не поясняет природу дробей, как равных частей целого. Кроме того, для построения отрицательных рациональных чисел здесь требуются отрицательные целые числа, т.к. в первом квадранте присутствуют только прямые, отвечающие положительным рациональным числам. Наконец, целым числам в такой модели соответствуют прямые разного наклона в отличие от модели целых чисел, представленной выше.

Обе построенные модели рациональных чисел изоморфны как по операциям, так и по упорядочению. Упражнение 2.26.

Обе модели включают в себя модель целых чисел, которая с теоретико-множественной точки зрения будет отличаться от построенных ранее (в одном случае вместо прямых рассматриваются классы эквивалентных прямых, в другом — непараллельные прямые). При этом, в поле рациональных чисел невозможно выделить собственное подполе, которое содержало бы модель целых чисел (и модель натуральных чисел).

Вторая из представленных моделей поля рациональных чисел восходит к методам алгебраической геометрии и по сути представляет собой определение

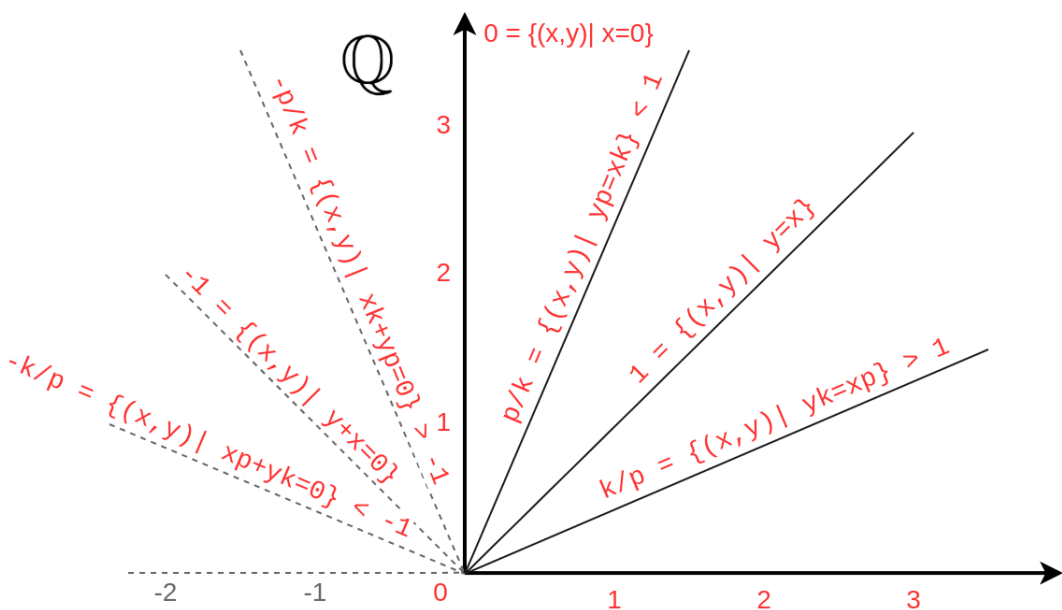


Рис. 2.7: Поле частных кольца целых чисел.

локализации кольца \mathbb{Z} .¹¹ В соответствии с этим можно дать обобщающее определение: **полем рациональных чисел** называется локализация кольца целых чисел.

Другое алгебраическое определение: полем рациональных чисел называется *минимальное упорядоченное поле*.

Поле рациональных чисел обозначается \mathbb{Q} .

Как и ранее, мы отмечаем, что отношения $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q}$ с теоретико-множественной точки зрения следует рассматривать применительно к некоторой модели \mathbb{Q} , в которой естественным (и единственным) способом выделены модели \mathbb{N} и \mathbb{Z} .

С точки зрения компьютерного моделирования рациональные числа можно представлять как конечные дроби, т. е. функции вида $f : \mathbb{Z} \rightarrow 10$ с конечным носителем. При этом цифры с отрицательным аргументом находятся после десятичной точки, а с неотрицательным — перед ней (снова имеем инверсный порядок, как при определении цифровой записи натуральных чисел!). Например, число 43.52 — это функция со значениями: $f(1) = 4$, $f(0) = 3$, $f(-1) = 5$, $f(-2) = 2$. Аргументы f здесь играют роль степеней основания 10: $43.52 = 4 \cdot 10^1 + 3 \cdot 10^0 + 5 \cdot 10^{-1} + 2 \cdot 10^{-2}$.

$$\begin{array}{l} 10 = \\ \{0, 1, 2, 3, \\ 4, 5, 6, 7, 8, 9\} \end{array}$$

¹¹В коммутативной алгебре используется также термин **поле частных** кольца \mathbb{Z} .

Для отрицательных чисел вновь впереди ставим знак «минус» (а точнее, компьютер просто использует специальный знаковый бит).

При таком подходе мы теряем бесконечные периодические дроби вроде $0.(3)$, знаменатели которых не являются делителями степеней 10, что, в общем, легко обойти специальными соглашениями в обозначениях.

На примерах построения целых и рациональных чисел мы подмечаем следующие два архетипа математической архитектуры: **неограниченное расширение** и **безграничное деление**. Первый позволил нам существенным образом выйти за рамки арифметики натуральных чисел, построив их до коммутативного кольца. Второй позволил начать заполнение «пустот» между целыми числами, деля единичный отрезок на произвольное (целое) количество частей. Ниже мы увидим квинтэссенцию этих двух архетипов при построении сюрреальных чисел, которые для ординалов являются тем же, чем для натуральных чисел — действительные числа.

2.4.4 Поле действительных чисел

Действительные числа своим возникновением во многом обязаны неразрешимости уравнений вида $x^2 = c$, где $c \in \mathbb{N}$, в поле рациональных чисел. В самом деле, решение уравнения $x^2 = 2$ не может быть рациональным числом.

Если предположить, что $x = n/m$, причем натуральные числа n и m несократимы, т. е. не имеют общих делителей > 1 , то $n^2 = 2m^2$, откуда по основной теореме арифметики 2.6 следует, что 2 входит в разложение числа n^2 по степеням простых, но тогда 2 входит и в разложение n , а значит, $n^2 = 4k$ при некотором целом $k > 0$. Отсюда имеем $2k^2 = m^2$, т. е. m также кратно 2. Последнее означает, что дробь n/m сократима как минимум на 2, что противоречит ее выбору.

Итак, уравнение $x^2 = 2$ неразрешимо в поле \mathbb{Q} . Отсюда следует и тот факт,

что \mathbb{Q} не является непрерывным полем, поскольку сечение

Упражнение 2.27.

$$(\{x \leq 0\} \cup \{x > 0 \mid x^2 \leq 2\}, \{x > 0 \mid x^2 > 2\})$$

не является дедекиндовым.

Возникает желание пополнить поле \mathbb{Q} некими новыми числами вроде $\sqrt{2}$, чтобы заполнить пустоты между рациональными числами и получить поле, в котором уравнения типа $x^2 = c$ имели бы решение.

Второе (и более сильное) пожелание состоит в том, чтобы построить непрерывное поле, т. е. заполнить пустоты между рациональными числами так, чтобы все сечения в нем были бы дедекиндовыми.¹² В дальнейшем мы увидим, что поле действительных чисел окажется не только непрерывным, но

¹²Иначе говоря, мы хотим иметь упорядоченное поле такое, что его невозможно представить как сумму открытых интервалов.

и покрывает с лихвой все наши «хотелки» по поводу разрешимости уравнений вида $x^2 = c > 0$.

Действительные числа мы построим способом, непривычным для стандартного анализа, но зато прекрасно согласованным с построением сюрреальных чисел (см. раздел 2.4.6). А именно, рассмотрим все пары вида (m, f) , где m — целое число, а f — функция вида $f : \omega \rightarrow 2 = \{0, 1\}$, причем f не имеет «хвоста» единиц, т. е. выполняется условие: $\forall n : (f(n) = 1 \rightarrow \exists m > n : f(m) = 0)$.

Очевидно, что при таком подходе m отвечает за целую часть числа, а двоичные значения функции f — за формирование дробной части с помощью (бесконечных) двоичных дробей. Например, 2.0011 — это число $2\frac{3}{16}$. При этом мы не уточняем, как именно построены целые числа, хотя по умолчанию можем предполагать самую простую модель $\mathbb{Z} : \omega \times \{0\} \cup \{0\} \times \omega$.

Ясно также, что среди пар (m, f) присутствуют коды рациональных дробей, прежде всего, так называемых **двоично-рациональных чисел**, т. е. дробей вида $m + \frac{k}{2^n}$. Им соответствуют такие и только такие коды, в которых $\text{supp}(f)$ конечен.¹³ В этом разделе мы будем отождествлять двоично-рациональное число и соответствующий ему код (m, f) .

Упражнение
2.28.
Докажите,
что эта со-
вокупность
есть
множество.

Совокупность всех пар (m, f) , удовлетворяющих указанным выше условиям, будем называть множеством **действительных чисел** (или **вещественных чисел**) и обозначать \mathbb{R} , а его элементы — действительными (вещественными) числами. Снова обращаем внимание читателя на то, что это лишь одна из многих моделей \mathbb{R} , не следует обозначение \mathbb{R} привязывать к какому-то конкретному объекту теории множеств.

Для конкретной пары $r = (m, f)$ построим два множества $L(r)$ и $R(r)$, где в $L(r)$ входят все рациональные дроби (m, g) , где $f(n) = 1$, $f|_n = g|_n$, $g(k) = 0$ при $k \geq n$, а в $R(r)$ входят все рациональные дроби (m, g) , где $f(n) = 0$, $f|_n = g|_n$, $g(n) = 1$, $g(k) = 0$ при $k > n$. То есть в $L(r)$ находятся двоично-рациональные дроби, двоичная запись которых на некотором начальном отрезке совпадает с записью (m, f) , причем в f дальше следует 1, а в g — 0. В $R(r)$ находятся двоично-рациональные дроби, двоичная запись которых на некотором начальном отрезке совпадает с записью (m, f) , причем в f дальше следует 0, а в g — 1 и затем нули. Для исключительного случая, когда r является целым числом, т. е. $r = (m, f)$, где f — тождественный ноль, в $L(r)$ отнесем все двоично-рациональные дроби вида $(m - 1, g)$. Множества $L(r)$ и $R(r)$ назовем, соответственно, **левым и правым двоично-рациональным множеством** числа r .

¹³Напомним, что supp обозначает носитель функции, т. е. подмножество ее области определения, где функция отлична от нуля. В данном случае это означает, что f принимает значение 1 конечное число раз.

Для двух действительных чисел $r = (m, f)$ и $r' = (m', f')$ положим:

$$r < r' \text{ если } \begin{cases} m < m', \text{ либо} \\ (m = m') \wedge (\exists n : f|_n = f'|_n \wedge f(n) < f'(n)), \end{cases}$$

во втором случае номер $n \geq 0$ называется точкой расхождения функций f и f' . Нетрудно показать неравенства: $L(r) < r < R(r)$.¹⁴

Упражнение 2.29.

Сложение действительных чисел определяется в три этапа:

(1) сложение двоично-рациональных дробей считается известным из арифметики \mathbb{Q} . Например,

$$\begin{array}{r} + \quad 0.011101 = \frac{29}{64} \\ \quad 0.100110 = \frac{38}{64} \\ \hline 1.000011 = 1\frac{3}{64} \end{array}$$

(2) сложение двоично-рациональной дроби и действительного числа (с бесконечным носителем f) сводится к шагу (1) путем представления этого числа как суммы двоично-рациональной дроби и числа с меньшими разрядами. Например,

$$\begin{array}{r} + \quad 0.011101(01)_\omega = 0.011101 + 0.000000(01)_\omega \\ \quad 0.100110 \\ \hline 1.000011 + 0.000000(01)_\omega = 1.000011(01)_\omega \end{array}$$

(3) сложение двух чисел с бесконечным носителем f определяется следующим способом: $r + r' \rightleftharpoons \sigma(L, R)$, где

$$\begin{aligned} L &\rightleftharpoons \{q + r' \mid q \in L(r)\} \cup \{q + r \mid q \in L(r')\}, \\ R &\rightleftharpoons \{q + r' \mid q \in R(r)\} \cup \{q + r \mid q \in R(r')\}, \end{aligned}$$

а число $\sigma(L, R)$ — это пара (s, g) с минимальным носителем g такая, что $L < (s, g) < R$.

Например, нам требуется сложить два числа $r = 0.(01)_\omega$ и $r' = 2.11(10)_\omega$. В данном случае это можно сделать быстро, пользуясь сложением в столбик:

$$\begin{array}{r} + \quad 0.01(01)_\omega \\ \quad 2.11(10)_\omega \\ \hline 3.00(11)_\omega \end{array}$$

¹⁴При сравнении множества и числа предполагается, что все элементы множества удовлетворяют указанному неравенству!

Исключая «хвост» единиц, получаем ответ 3.01. Построим теперь L и R :

$$\begin{aligned} L(r) &\ni 0.(01)_n \\ R(r) &\ni 0.(01)_n 1 \\ L(r') &\ni 2.11(10)_n 1 \\ R(r') &\ni 2.11(10)_n 11 \end{aligned}$$

Упражнение 2.30. Подумать, почему это так? Здесь мы включили в левые множества только максимальные числа, а в правые — только минимальные из возможных. Остальные числа не играют роли при определении суммы $r + r'$. Далее, необходимые для построения L и R суммы включают числа:

$$\begin{aligned} L(r) + r' &\ni 0.(01)_n + 2.11(10)_\omega = 3.00(11)_n(10)_\omega \\ L(r') + r &\ni 0.(01)_\omega + 2.11(10)_n 1 = 3.00(11)_n(01)_\omega \\ R(r) + r' &\ni 0.(01)_n 1 + 2.11(10)_\omega = 3.01(00)_n(10)_\omega \\ R(r') + r &\ni 0.(01)_\omega + 2.11(10)_n 11 = 3.01(00)_n(01)_\omega \end{aligned}$$

где при суммировании мы воспользовались шагом (2).

Ясно, что $L(r') + r < L(r) + r'$, поэтому достаточно полагать, что L состоит только из чисел $3.00(11)_n(10)_\omega$. Аналогично, $R(r') + r < R(r) + r'$, поэтому в R достаточно включить только числа $3.01(00)_n(01)_\omega$. Теперь легко видеть, что

$$3.00(11)_n(10)_\omega < 3.01 < 3.01(00)_n(01)_\omega,$$

причем решение $\sigma = 3.01$ является единственным решением неравенств $L < \sigma < R$ и, следовательно, является суммой $r + r'$.

Таким образом, суммирование чисел сначала сводится к суммированию числа и множества двоично-рациональных чисел, которое, в свою очередь, сводится к сложению двоично-рациональных чисел, после чего между полученными двумя множествами L и R выбирается промежуточное число с самым коротким носителем. В нашем случае длина носителя оказалась равной 2 (два числа после точки-разделителя), однако, легко проверить, что никакое более длинное двоичное число не может удовлетворять этому же неравенству. Если бы мы по методу (3) определяли и сложение двоично-рациональных чисел, сводя их сумму к сумме все более коротких (в смысле носителя f) чисел до тех пор, пока не добрались бы до суммирования целых чисел, нам пришлось бы пользоваться принципом минимальной длины для решения $\sigma(L, R)$. Существование и единственность такого решения будет показана в общем случае в теореме 2.15.

Умножение действительных чисел определяется аналогичным способом.

$r \cdot r' \Rightarrow \sigma(L, R)$, где

$$L \Rightarrow \{ur' + rv - uv \mid u \in L(r), v \in L(r')\} \cup \{ur' + rv - uv \mid u \in R(r), v \in R(r')\};$$

$$R \Rightarrow \{ur' + rv - uv \mid u \in L(r), v \in R(r')\} \cup \{ur' + rv - uv \mid u \in R(r), v \in L(r')\}.$$

Здесь мы снова видим переход от операции над числами к операции над числом и двоично-рациональным числом, которая сводится к конечной сумме чисел. Например,

$$\begin{array}{r} \times \quad 0.(01)_\omega \\ 2.11 = 2 + 0.1 + 0.01 \\ \hline 0.(01)_\omega + 0.(01)_\omega + 0.0(01)_\omega + 0.00(01)_\omega \end{array}$$

Смысл формулы $ur' + rv - uv$ состоит в следующем. Во-первых, от умножения rr' мы переходим к умножению rv и ur' , т. е. к умножению на двоично-рациональное число u или v .

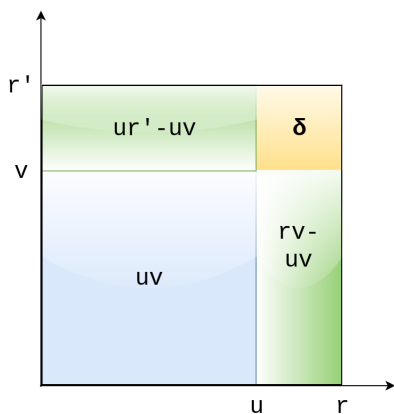
Во-вторых, это выражение является нижней оценкой для rr' , если $r, r' > 0$. Действительно, в случае $u < r, v < r'$ площадь прямоугольника rr' состоит из площадей $(r-u)v$, $(r'-v)u$, uv и $\delta = (r-u)(r'-v)$, а выражение $ur' + rv - uv$ равно сумме трех первых из указанных площадей. Остаток δ делается сколь угодно малым выбором достаточно больших u и v . Итак, $ur' + rv - uv = rr' - \delta$.

Аналогично, $ur' + rv - uv = rr' - \delta$ и в случае $r < u, r' < v$. Таким образом, в L все числа чуть меньше искомого rr' , причем они получены разными способами — как из левых множеств, так и из правых множеств исходных чисел r и r' . Это связано с тем, что в зависимости от знаков r и r' срабатывает либо одно, либо второе множество, как мы видели в примере для сложения.

Такие же рассуждения приводят к обоснованию конструкции R .

Таким образом, мы имеем рекурсивное определение сложения и умножения действительных чисел, сводящееся к операциям над двоично-рациональными числами и поиску числа $\sigma(L, R)$.¹⁵

Более того, как легко видеть, сложение двоично-рациональных чисел вообще сводится к сложению целых, для чего достаточно оба слагаемых умножить на достаточно большое число 2^n , тем самым сдвинув двоичные разряды в область целых чисел. Это означает, на самом деле, что поле действительных чисел строится непосредственно при помощи знания арифметики кольца



¹⁵Его поиск также можно в значительной мере алгоритмизировать, на что указывает доказательство теоремы 2.15.

целых чисел, а знание рациональных чисел для его построения не является необходимым. Скорее, изучая действительные числа, легче построить поле \mathbb{Q} , чем наоборот:

$$\omega \longrightarrow \mathbb{N} \longrightarrow \mathbb{Z} \longrightarrow \mathbb{R} \longrightarrow \mathbb{Q}$$

при этом «спусковым механизмом» для этих построений является аксиома бесконечности, гарантирующая существование ординала ω .

Построенное нами множество \mathbb{R} с отношением порядка $<$ и операциями сложения и умножения является упорядоченным полем. Доказательство этого факта следует из того, что наша конструкция \mathbb{R} в целом повторяет определение действительных чисел как *бесконечных десятичных дробей*, восходящее еще к Вейерштрассу. Мы только вместо основания 10 использовали основание 2. Каждое действительное число (m, f) представляется сходящимся рядом

$$(m, f) = m + \sum_{n=0}^{\infty} \frac{f(n)}{2^{n+1}}.$$

Вместе с тем, определение для всякого действительного числа левого и правого двоично-рациональных множеств отсылает нас к теории *дедекиндовых сечений* — другому способу построения поля \mathbb{R} . В этой теории сложение и умножение действительных чисел определяется через сложение и умножение элементов соответствующего сечения. На самом деле, нетрудно показать, что определенное нами сложение и умножение эквивалентны таковым в теории дедекиндовых сечений для \mathbb{R} . Например, вместо формулы $ur' + rv - uv$ для определения умножения там берется только uv . В случае действительных чисел это вполне оправдано, поскольку «довески» $(r - u)r'$ и $(r' - v)r$ можно сделать сколь угодно малыми выбором достаточно больших $u \in L(r)$, $v \in L(r')$. Это обеспечивается ограниченностью r и r' , т. е. таким свойством \mathbb{R} , которое называется аксиомой Архимеда. Позже мы увидим, что как только мы расширим поле \mathbb{R} «в стороны» с помощью ординалов, переходя к сюрреальным числам, архимедовость такой числовой системы утратит силу, и мы вынуждены будем определять умножение более тонким способом, а именно, как $rr' - \delta$, основываясь на действительно малых отклонениях от искомого числа.

Как видно из теоремы 2.12 свойство упорядоченного поля быть архимедовым является более слабым, чем свойство непрерывности. На самом деле, наша конструкция \mathbb{R} позволяет легко показать и непрерывность линейного порядка на \mathbb{R} . Действительно, имея два множества $L < R$ (а если верно $L \leq R$, то нам достаточно выбросить $L \cap R$, состоящее максимум из 1 элемента, и получить $L < R$), мы можем найти число $\sigma(L, R)$, которое будет разделять L и R : $L \leq \sigma(L, R) \leq R$. А это и означает наличие точных граней у ограниченных множеств.

Возможно, более простым объяснением непрерывности будет тот факт, что любая монотонная ограниченная последовательность будет иметь предел (что равносильно утверждению: всякое непустое ограниченное сверху множество имеет точную верхнюю грань). Для доказательства этого факта нужно от произвольной последовательности аккуратно перейти к последовательности двоично-рациональных дробей, имеющей тот же предел, после чего построение самого предела станет делом техники.

Доказательство — отпад! :)

Наконец, плотность \mathbb{R} прямо следует из арифметических свойств: для любых $x < y$ достаточно вычислить $(x + y)/2$, чтобы найти число в интервале $(x; y)$. Либо можно снова перейти от действительных чисел к двоично-рациональным ($x < d_1 < y$) аналогично построению $\sigma(L, R)$.

А отсюда уже по теореме 1.13 будет следовать непрерывность поля \mathbb{R} . Вся эта техника ювелирно оттачивается на занятиях по математическому анализу, поэтому мы не станем заострять на ней внимание.

Важно то, что построенная модель \mathbb{R} является непрерывным упорядоченным полем. И именно такое свойство принято считать определением поля действительных чисел как алгебраического понятия.

Определение. Полем действительных чисел называется всякое непрерывное упорядоченное поле.

Известно, что такое поле единственно с точностью до изоморфизма, сохраняющего порядок и операции сложения и умножения. А из наших построений, приведенных выше, следует существование такого поля в рамках аксиоматики теории множеств.

Как уже отмечалось (см. комментарий после теоремы 2.12), непрерывность поля эквивалентна выполнению одновременно двух следующих условий: архимедовость и полнота по Гильберту. Поэтому полем действительных чисел также является всякое *максимальное архимедово упорядоченное поле*.

Скажем пару слов об алгебраических числах. **Алгебраическое число** — это корень уравнения $f(x) = 0$, где f — многочлен степени n с целыми коэффициентами. Степенью алгебраического числа называется минимально возможная степень многочлена, который это число обращает в ноль. Так, рациональное число вида p/q есть алгебраическое число первой степени, поскольку зануляет многочлен $p - qx$. С ними мы уже имели дело при построении модели \mathbb{Q} . Число $\sqrt{2}$ зануляет $x^2 - 2$, но не является корнем никакого линейного уравнения с целыми коэффициентами, так что это — алгебраическое число степени 2. И т.д.

Целое число — алгебраическое нулевой степени.

Алгебраические числа тем плотнее заполняют действительную ось, чем выше их степень. Об этом свидетельствует

Теорема 2.13 (Лиувилля). *Если x — алгебраическое число степени $n > 1$,*

то существует $C > 0$ такое, что в окрестности x радиуса C/q^n , где $q \in \mathbb{N} \setminus \{0\}$, нет рациональных чисел со знаменателем q , т. е. всегда выполняется неравенство

$$\left| x - \frac{p}{q} \right| > \frac{C}{q^n}.$$

Существует уточнение теоремы Лиувилля, показывающее, что показатель $n + 1$ можно заменить на $\frac{n}{2} + 1$.

Так, если мы хотим приблизить алгебраическое число степени n двоично-рациональными числами со знаменателем 2^k с точностью ε , необходимо, чтобы выполнялось неравенство: $\varepsilon > 1/2^{k(\frac{n}{2}+1)}$. Например, число $\sqrt{2}$ нельзя приблизить с точностью 0.001 двоично-рациональными числами с основанием 8, т. е. числами вида $k/8$.

Однако действительная ось не исчерпывается алгебраическими числами, больше того, она в основном состоит из чисел **трансцендентных**, т. е. таких, которые не являются алгебраическими. Теорема Лиувилля дает способ построения трансцендентных чисел. Рассмотрим действительное число вида

На самом деле, легко проверить, что и числами вида $k/64$ нельзя, а вот $181/128$ удовлетворяет точности 0.001

$$z = 0.01 \underbrace{00}_{2!} 1 \underbrace{000000}_{3!} 1 \underbrace{000000000000000000000000}_{4!} 1 \dots,$$

где после каждой единицы на n -ом шаге ставится $n!$ нулей. Предлагаем читателю самостоятельно проверить, что данное число z приближается рациональными числами (полученными из z отсечением записи по n -ой единице) намного быстрее, чем того требует теорема Лиувилля для алгебраических чисел. Следовательно, это число трансцендентное.

Упражнение 2.31. Итак, мы видим, что поскольку \mathbb{R} непрерывно, оно включает не только \mathbb{Q} , но и корни любой степени из положительных рациональных чисел, все алгебраические числа (не содержащие мнимой части), а также несоизмеримо большее множество трансцендентных чисел, яркими представителями которых являются константы e и π . Вместе с любой ограниченной последовательностью в \mathbb{R} лежат и все ее частичные пределы, именно в силу свойства непрерывности.

Однако, как мы знаем из анализа, последовательности могут сходиться к одному и тому же числу с различной «скоростью», например, $1/n$ сходится к нулю медленнее, чем $1/n^2$ (относительно параметра n), но это никак не фиксируется в свойствах самого предела, т. е. нуля в данном случае. Поэтому еще со времен Ньютона и Лейбница математики пытались оперировать так называемыми бесконечно малыми и бесконечно большими величинами, считая их тоже некоторыми числами, «живущими» где-то между обычными «стационарными» величинами. Такие нестандартные числа плохо поддавались формализации, и в конце концов математики усилиями, прежде всего,

Больцано, Коши и Вейерштрасса, разработали теорию пределов, которая на время закрыла вопрос о пополнении \mathbb{R} бесконечно малыми числами.

Тем не менее, уже в XX веке, с развитием теории множеств, появились новые идеи по формализации нестандартных величин. А. Робинсон [79, 93] показал, что \mathbb{R} можно непротиворечивым способом дополнить бесконечно малыми и бесконечно большими величинами так, чтобы выполнялся принцип непрерывности Лейбница, т. е. чтобы все утверждения о действительных числах, записанные в логике первого порядка, сохранялись бы и в новой системе чисел.

Интересно, что как только мы начинаем пополнять \mathbb{R} новыми числами (бесконечно малыми и бесконечно большими), вводя нестандартный анализ, мы тут же теряем свойство непрерывности. То есть попытки вставить что-то между действительными числами приводят к тому, что вместе с этим там появляются и «дырки».

Инфините-
зимальный
;))

Например, пусть E — множество всех положительных бесконечно малых чисел, а R — множество всех положительных действительных чисел. Если непрерывность чисел сохраняется, то существует граница между этими множествами — число c , удовлетворяющее неравенствам $E \leq c \leq R$. Ясно, что поскольку $2c > c$, то $2c$ не может быть бесконечно малым, а значит, таковым не может быть и $c = (2c)/2$, в то же время $c/2 < c$ является бесконечно малым, а значит, c — также бесконечно малое, т. к. $c = 2(c/2)$. Получаем противоречие.

На этом же примере легко увидеть нарушение архимедовости расширения \mathbb{R} : действительно, никакое конечное суммирование бесконечно малого ε не превзойдет числа 1.

Учитывая вышеперечисленные свойства, можно сказать, что \mathbb{R} является уникальной числовой структурой (полем), дающей нам фактическое представление о непрерывности, как о некоем локальном свойстве в мире множеств. Как меньшие, так и большие поля таким свойством не обладают. В каком-то смысле можно сделать вывод о том, что *непрерывное есть частный случай дискретного*. Действительные числа, как атомы, — до-
статочно мелкие, чтобы в макромире быть незаметными и обес-
печивать «непрерывность» материи, и достаточно крупные, чтобы в микро-
мире быть «дырявым» решетом для лептонов.

Здравствуй,
Физика!

2.4.5 Гипердействительные числа

Следующее наше построение будет существенно отличаться от предыдущих и носит сугубо теоретико-множественный характер. Здесь мы, пожалуй, впервые существенно прибегнем к строительству математической структуры неалгебраическим методом, а именно: построим некую систему подмножеств с нужными нам свойствами.

Определение. Непустая система $\mathcal{U} \subseteq \mathcal{P}(X)$ подмножеств множества X называется **фильтром на множестве X** , если выполнены следующие условия:

F1 $\forall a, b \in \mathcal{P}(X) : (a \in \mathcal{U}) \wedge (a \subseteq b) \rightarrow (b \in \mathcal{U})$ (монотонность);

F2 $\forall a, b \in \mathcal{U} : a \cap b \in \mathcal{U}$ (замкнутость по \cap);

F3 $\emptyset \notin \mathcal{U}$.

Фильтр \mathcal{U} называется **ультрафильтром** на X , если помимо F1–F3 выполняется условие

F4 $\forall a \in \mathcal{P}(X) : (a \in \mathcal{U}) \vee (X \setminus a \in \mathcal{U})$ (полнота).

Элементы ультрафильтра можно условно называть «большими» множествами: если «большое» множество содержится в каком-то множестве, то это второе множество также «большое»; пересечение двух «больших» множеств — большое, пустое множество — не «большое». Отметим, что одновременно и a , и $X \setminus a$ не могут быть «большими», т. к. мы бы получили, что их пересечение тоже «большое», а это противоречит F3.

Ультрафильтр — это максимальный по вложению фильтр.

Базой фильтра называется непустое семейство множеств, удовлетворяющее свойствам

F2' $\forall a, b \in \mathcal{U} : \exists c \in \mathcal{U} : c \subseteq a \cap b$;

F3 $\emptyset \notin \mathcal{U}$.

Фильтр, порожденный базой, включает все надмножества множеств из базы, и только их, т. е. если \mathcal{B} — база фильтра, то порожденный ею фильтр есть семейство

$$\mathcal{F}(\mathcal{B}) = \{a \in \mathcal{P}(X) \mid \exists b \in \mathcal{B} \ b \subseteq a\}.$$

Базы \mathcal{B}_1 и \mathcal{B}_2 называются эквивалентными, если для всякого $b \in \mathcal{B}_1$ существует $c \in \mathcal{B}_2$ такое, что $b \subseteq c$, и наоборот (т. е. эти базы взаимно вкладываются друг в друга своими множествами). Читатель может самостоятельно Упражнение 2.32. проверить, что эквивалентность баз является отношением эквивалентности на $\mathcal{P}^2(X)$. Эквивалентные базы порождают один и тот же фильтр, который является максимальной базой по вложению в своем классе эквивалентности.

Фильтр *порожден* множеством $a \subseteq X$, если он включает все надмножества a . Если $a = \{x_0\}$ (синглет), то говорят, что фильтр *порожден точкой x_0* . Фильтр, порожденный точкой, называется **главным фильтром** и является ультрафильтром.

Если же ультрафильтр \mathcal{U} на X обладает свойством

$$F5 \quad \forall a \in \mathcal{P}(X) : (||a|| < \omega) \rightarrow (a \notin \mathcal{U}),$$

то он называется **неглавным**. Существование неглавных ультрафильтров следует из теоремы Тарского об ультрафильтрах, которая гласит, что всякий фильтр на X можно вложить в некоторый ультрафильтр. Таким образом, если рассмотреть фильтр, состоящий из дополнений ко всем конечным подмножествам X , то его расширение до ультрафильтра приведет к получению неглавного ультрафильтра. Проблема здесь кроется в том, что теорема Тарского есть следствие аксиомы выбора (а точнее, леммы Цорна), а значит, конструктивное построение неглавного ультрафильтра невозможно. Тем не менее, предполагая, что ультрафильтр существует, можно построить систему чисел, которая является моделью нестандартного анализа, т. е. включает в себя бесконечно малые и бесконечно большие числа.

Пусть далее \mathcal{U} — это какой-нибудь неглавный ультрафильтр на ω . Рассмотрим множество всех функций из ω в \mathbb{R} и введем на них отношение эквивалентности: $f \sim g$, если $\{n \mid f(n) = g(n)\} \in \mathcal{U}$. То есть, последовательности f и g эквивалентны, если они совпадают на «большом» множестве.

Упражнение 2.33. Проверьте, что это отношение эквивалентности

Фактор-множество $(\mathbb{R}^\omega)/\sim$ еще называют **ультрастепенью** \mathbb{R} (относительно ультрафильтра \mathcal{U}). Элементы данного фактор-множества, т. е. классы эквивалентности, как и раньше, будем обозначать $[f]$, где f — произвольный элемент класса эквивалентности $[f]$.

Далее, на фактор-множестве $(\mathbb{R}^\omega)/\sim$ введем операции сложения и умножения по правилам:

$$[f] +_{HR} [g] = [f + g], \quad [f] \cdot_{HR} [g] = [fg],$$

где сложение функций выполняется поточечно. Кроме того, положим:

$$[f] <_{HR} [g], \quad \text{если } \{n \mid f(n) < g(n)\} \in \mathcal{U},$$

Упражнение 2.34. Проверьте корректность определения

что задает линейный порядок на $(\mathbb{R}^\omega)/\sim$.

Структура $(\mathbb{R}^\omega)/\sim$ с определенными выше порядком и операциями сложения и умножения называется **системой гипердействительных чисел** и обозначается \mathbb{HR} . Наконец, для произвольного числа $r \in \mathbb{R}$ обозначим

Отдел кадров?..

$$*r \equiv [\{(n, r) \mid n \in \omega\}],$$

т. е. $*r$ — это класс эквивалентности функции-константы, всюду равной r . Числа $*r$ представляют собой реализацию действительных чисел внутри \mathbb{HR} и называются **стандартными**.

Дадим два определения, которые лексически почти всегда сопровождают термин «гипердействительные числа». **Монадой** числа $*r$ называется множество всех чисел x таких, что $x - *r$ — бесконечно малое. Иначе говоря, монада — это окрестность действительного числа, состоящая из всех бесконечно близких к нему гиперчисел. **Галактикой** числа x называется множество всех чисел y таких, что $x - y$ — конечное. То есть, вся гипердействительная ось распадается на бесконечное множество бесконечных интервалов, каждый из которых порядково изоморфен \mathbb{R} , пополненному монадами (т.е. множеству конечных гипердействительных чисел).

Следующие свойства гипердействительных чисел мы предлагаем читателю для самостоятельного изучения:

Упражнение
2.35.

- HR1 \mathbb{HR} включает изоморфный образ \mathbb{R} ;
- HR2 существует бесконечно большое число в \mathbb{HR} ;
- HR3 существует бесконечно малое число в \mathbb{HR} ;
- HR4 для каждого конечного $x \in \mathbb{HR}$ существует единственное $*r$ ($r \in \mathbb{R}$) такое, что $x - *r$ — бесконечно малое;
- HR5 галактики либо не пересекаются, либо совпадают (и образуют разбиение \mathbb{HR});
- HR6 сумма и произведение бесконечно малых есть бесконечно малое;
- HR7 сумма и произведение бесконечно больших есть бесконечно большое;
- HR8 галактика является плотным ЛУМ без наибольшего и наименьшего элементов;
- HR9 упорядоченное поле \mathbb{HR} не является непрерывным;
- HR10 упорядоченное поле \mathbb{HR} не является архимедовым.

Вновь мы вынуждены акцентировать внимание читателя на том, что построенный на основе произвольно выбранного неглавного ультрафильтра \mathcal{U} объект есть чисто теоретико-множественная конструкция, всего лишь моделирующая гипердействительные числа. Это значит, что при выводе свойств гипердействительных чисел нельзя пользоваться знанием конструкции их модели, т.е. свойствами ультрафильтра. Резонный вопрос — чем же тогда гарантированно можно пользоваться? Ответ: можно пользоваться языком действительных чисел, в котором отсутствует отношение принадлежности.

Иначе говоря, можно пользоваться, как и раньше, только алгебраическими операциями и теоремами о них. В общем смысле полем гипердействительных чисел можно называть любое линейно упорядоченное поле, расширяющее \mathbb{R}

(т. е. содержащее \mathbb{R} как собственное подполе). Естественно, как и раньше, вложение $\mathbb{R} \subset \mathbb{HR}$ мы читаем с точностью до изоморфизма — в \mathbb{HR} имеется изоморфная (по операциям и отношению порядка) копия \mathbb{R} .

Выше мы перенесли операции сложения и умножения на ультрастепень \mathbb{R} естественным способом. Рассмотрим теперь \mathbb{R} как алгебраическую систему, в которой помимо данных арифметических операций и отношения $<$ заданы еще набор операций F и набор отношений R (обычно считается, что это *все возможные* операции и отношения), после чего перенесем и их в ультрастепень \mathbb{R} аналогичным образом: если $F \in F$, $R \in R$, то

$$\begin{aligned} F^{HR}([f]) &= [g], \text{ если } \{n \mid F(f(n)) = g(n)\} \in \mathcal{U}, \\ [f]R^{HR}[g], &\text{ если } \{n \mid f(n)Rg(n)\} \in \mathcal{U} \end{aligned}$$

Иначе говоря, операции и отношения, применяемые в \mathbb{R} , мы переносим сначала на последовательности действительных чисел, а затем переходим к классам эквивалентности с помощью ультрафильтра. Эти определения можно легко обобщить на случай n -местных операций и отношений. Множества всех новых операций и отношений обозначим, соответственно, F^{HR} и R^{HR} .

Таким образом, мы получаем алгебраическую структуру $(\mathbb{HR}, F^{HR}, R^{HR})$, в точности повторяющую исходную структуру операций и отношений, заданную над \mathbb{R} . В теории моделей мы говорим, что система \mathbb{HR} с функциями F^{HR} и отношениями R^{HR} является нормальной моделью исходной сигнатуры (F, R) .

Рассмотрим теперь произвольную формулу $\varphi(x_1, \dots, x_n)$, содержащую в качестве атомарных формул и термов только отношения равенства и порядка действительных чисел, функциональные термы, отвечающие за сложение и умножение действительных чисел, а также формулы отношений из R и термы функций из F , кроме того, предположим, что все переменные в этой формуле пробегают лишь множество \mathbb{R} . Иначе говоря, мы рассматриваем формулу языка действительных чисел.

После чего заменим всюду отношение $<$ на $<_{HR}$, $+$ на $+_{HR}$, \cdot на \cdot_{HR} , отношения R на R^{HR} , функции F на F^{HR} и все переменные x_1, \dots, x_n на $*x_1, \dots, *x_n$, т. е. на их аналоги в \mathbb{HR} . Полученную формулу обозначим φ^{HR} . Существует теорема (Робинсона)¹⁶, выражающая лейбницевский принцип непрерывности и говорящая о том, что

$$\varphi(x_1, \dots, x_n) \leftrightarrow \varphi^{HR}(*x_1, \dots, *x_n), \quad (2.8)$$

т. е. все алгебраические утверждения, верные в аксиоматике действительных чисел, остаются верными в \mathbb{HR} применительно к реализации действительных чисел в \mathbb{HR} , и обратно, доказанное в \mathbb{HR} утверждение указанного вида будет верным и в стандартном анализе. Например, верно равенство

$$(\sin^{HR})^2(x) + (\cos^{HR})^2(x) = 1, \quad (2.9)$$

¹⁶На самом деле, это следствие более общей теоремы Лоса об ультрапроизведениях.

где под единицей мы понимаем ее инкарнацию в гипердействительных числах, т.е. $\ast 1$. «На пальцах» это равенство проверить довольно просто. В самом деле, гипердействительное число x есть класс эквивалентности $[f]$, где f — последовательность обычных действительных чисел. Все, что нам нужно проверить — это удостовериться, что множество тех n , для которых равенство $\sin^2(f(n)) + \cos^2(f(n)) = 1$ верно, является элементом ультрафильтра. Но это равенство справедливо вообще для всех действительных чисел, следовательно, оно справедливо для всех $n \in \omega$, т.е. область его истинности — все ω целиком. В то же время, $\omega \in \mathcal{U}$. Стало быть, равенство переносится в гипердействительные числа. Например, если мы рассмотрим класс $[1/n]$ (т.е. $f(n) = 1/n$), то будем иметь $(\sin^{HR})^2([1/n]) + (\cos^{HR})^2([1/n]) = 1$, где $[1/n]$ олицетворяет бесконечно малое число (одно из). Аналогично можно рассматривать и бесконечно большие числа.

На приведенном примере хорошо видно, что математика — во многом символическая наука. Мы не просто стараемся получить результаты, находясь в «физике» мира теории множеств, а мы сам результат рассматриваем как объект изучения. Мы знаем, что уравнение $\sin^2 x + \cos^2 x = 1$ истинно в некоторой области данных (\mathbb{R}), после чего мы производим изоморфные преобразования (переходим в ультрастепень) самого этого символического равенства и получаем новую символическую запись, которая оказывается верным уравнением в новой области данных (\mathbb{HR}).

Это напоминает нашу работу с «начальными» множествами. Ведь мы их рассматривали как скобочные записи, изучали свойства этих записей и вводили на них некоторые отношения, в конечном итоге приведшие нас к определению равенства множеств и позволившие рассматривать некоторые виртуальные объекты — множества, а скобочные записи остались лишь их наименованиями.

В то же время, все теоремы являются записями на определенном языке. Среди них также могут быть равные записи, являющиеся именами для эквивалентных суждений. И мы можем точно так же изучать записи теорем, придумывать для них интерпретации, производить над ними определенные преобразования (как в примере выше — переход к новым переменным, операциям и отношениям) и делать выводы о корректности таких преобразований и их инвариантах. Иногда это позволяет получать математические теоремы метаматематическими способами. Пример выше иллюстрирует такой подход, когда мы, ничего не зная о свойствах гипердействительных чисел, получаем утверждение (2.9), основываясь на чисто формальной подмене значков и известных свойствах этой подмены (переход к классам с помощью ультрафильтра).

В дальнейшем мы еще вернемся к математической логике (или метаматематике), предметом изучения которой являются как раз языки, формулы, истинность и выполнимость в той или иной модели.

Инфинитезимальные системы

Мы уже упоминали ранее в разделе 2.4.4 о работах А. Робинсона по «узакониванию» нестандартного анализа. Отметим, что развитие его идей привело к появлению иного взгляда на теорию множеств вообще. Было предложено различать стандартные и нестандартные множества. Стандартные множества полностью подчиняются аксиомам ZF, но при этом бесконечные стандартные множества содержат нестандартные элементы. Например, в множестве натуральных чисел существует нестандартный элемент N , который больше любого стандартного натурального числа.¹⁷

В книге Э. Нельсона [108] для иллюстрации свойства стандартности приводится неформальное употребление слова «фиксированное». Например, мы говорим: для любого фиксированного натурального k существует натуральное n такое, что $k < n$. Таким образом, получается, якобы, что все фиксированные натуральные числа лежат в каком-то начальном отрезке натурального ряда.

Чтобы эти идеи формализовать, было предложено несколько расширений стандартной теории множеств (ZF или ZFC) с помощью дополнительных аксиом и новых типов объектов. В чем-то эти расширения схожи с NGB, оперирующей понятиями множества и класса, но по сути это совершенно новые версии теории множеств.

Теория, предложенная Э. Нельсоном и обозначаемая IST (*internal set theory*), расширяет ZF следующим образом: помимо атомарной формулы принадлежности добавляется формула $St(x)$, означающая « x есть стандартное множество», вводя тем самым новый тип объектов теории (точнее, выделяя среди всех объектов стандартные). Очевидно, что при этом язык ZF включен в язык IST. Формулы языка ZF называются *внутренними* формулами, а все остальные формулы (содержащие формулу St явно или неявно) называются *внешними*. Соответственно, область истинности внешней формулы называют *внешним классом*, а внутренней формулы — *внутренним классом*. Множества, которые построены в рамках ZF с помощью аксиом AU, AP, AF, AI и внутренних формул, являются *внутренними множествами*, а пересечения внешних классов с внутренними множествами называются *внешними множествами* (фактически, это множества, построенные по тем же аксиомам, но при этом в определяющей их формуле прямо или косвенно содержится предикат St).

Схематично эти определения можно выписать следующим образом:

¹⁷На самом деле, если все счетные ординалы рассматривать как модель \mathbb{N} (выбросив аксиому Пеано существования $(n - 1)$ -го элемента), то мы увидим сколько угодно возможностей для воплощения бесконечно больших чисел.

	Формула НЕ содержит St	Формула содержит St
Формула	Внутренняя	Внешняя
Область истинности формулы	Внутренний класс	Внешний класс
Определения множеств через терм-квантор и формулу	Внутреннее множество	Внешнее множество
Прогрессивное множество	Внутреннее множество	

Определения множеств с помощью аксиом AU, AP и AF сводятся к терм-квантору с некоторой формулой.

Отметим, что теория IST никак не расширяет объем теории, а лишь разделяет ее объекты на два новых класса — стандартные и нестандартные. Для упрощения записи в языке IST вводятся следующие сокращения:

$$\begin{aligned}\forall^{st}x \varphi(x) &\equiv \forall x (St(x) \rightarrow \varphi(x)) \\ \forall^{stfin}x \varphi(x) &\equiv \forall x ((St(x) \wedge \|x\| < \omega) \rightarrow \varphi(x)) \\ \exists^{st}x \varphi(x) &\equiv \exists x (St(x) \wedge \varphi(x))\end{aligned}$$

Наконец, к аксиомам ZFC добавляются следующие аксиомы:

IST1 **принцип переноса**: $\forall^{st}t_1 \dots \forall^{st}t_n (\forall^{st}x \varphi(x) \leftrightarrow \forall x \varphi(x))$ для каждой внутренней формулы φ со свободными переменными x, t_1, \dots, t_n ;

IST2 **принцип идеализации**: $(\forall^{stfin}z \exists x \forall y \in z \varphi(x, y)) \leftrightarrow (\exists x \forall^{st}y \varphi(x, y))$ для каждой внутренней формулы φ ;

IST3 **принцип стандартизации**: $\forall^{st}x \exists^{st}y (\forall^{st}z : z \in y \leftrightarrow (z \in x) \wedge \varphi(z))$ для каждой (внутренней или внешней) формулы φ .

Эти три аксиомы соответствуют трем неформальным принципам любой инфинитезимальной теории, расширяющей ZF.

Первый — это принцип непрерывности Лейбница, он означает, что все утверждения обычной теории множеств должны сохраняться в неизменном виде в расширенной теории. Кроме того, обычные определения объектов (более точно: определения, задающие единственный объект) дают одновременно определения стандартных объектов в силу этой аксиомы. То есть, стандартных объектов в ZF может быть задано столько, сколько написано формул. Например, пустое множество \emptyset , множество натуральных чисел фон Неймана ω , посторонняя нами выше модель \mathbb{R} — все это стандартные множества.

Второй принцип выглядит довольно запутанно, но он схож с нашей аксиомой равенства AE в варианте 3 (см. раздел 1.2.1) и означает, что подставлять в формулу произвольное стандартное y — это то же самое, что в качестве y использовать произвольный элемент произвольного конечного стандартного

z . На языке «фиксированных» это звучит так: «существует такой x , что для любого фиксированного y выполняется φ » равносильно утверждению, что «каково бы ни было конечное фиксированное z , существует такой y , что φ верно для всех его элементов».

Из аксиомы IST2 следует, например, существование нестандартных объектов. Действительно, полагая $\varphi \equiv (x \neq y)$, мы получаем, что левая часть IST2 истинна, поскольку для любого конечного z существует такой x , что $y \neq x$ для всех $y \in z$ (достаточно взять $x = z$ и воспользоваться аксиомой регулярности), но тогда истинна и правая часть IST2: существует x такой, что ни один стандартный y не равен ему, т. е. существует нестандартный x .

Приведем еще несколько следствий аксиомы IST2 [108]:

1. в любом бесконечном стандартном множестве существует нестандартный элемент;
2. x — стандартное конечное множество тогда и только тогда, когда все его элементы стандартны;
3. существует конечное внутреннее множество, среди элементов которого встречается каждое стандартное множество.

Третий принцип сохраняет возможность с помощью любого (в том числе внешнего) утверждения выделить из стандартного множества стандартное подмножество, состоящее из стандартных элементов.

Существует **теорема Пуэлла**, доказывающая, что теория IST является консервативным расширением ZFC. Это значит, что при доказательстве «стандартных» теорем о множествах мы вправе пользоваться формализмом IST с той же степенью надежности, которую мы имеем в рамках теории ZFC.

Отметим, что хотя теория внутренних множеств носит достаточно общий характер, ее целью, тем не менее, является упрощение языка для описания гипердействительных чисел, и большинство результатов этой теории посвящено подмене теории пределов в математическом анализе. Тем не менее, существует и ряд не совсем обычных свойств, когда теоремы обычного анализа переключаются на нестандартные множества и функции.

Дальнейшее развитие идей работы с нестандартными множествами привели к тому, что потребовалось «выйти» за рамки внутренних множеств Нельсона. Примерами таких расширений являются теория EХТ (К. Хрбачек) и близкая по конструкции теория NST (Т. Каваи). Свойства так называемых внешних множеств в этих теориях приблизительно дают нам ту же свободу, что наивная (доаксиоматическая) теория Кантора, но расплачиваться за нее приходится, и весьма существенно. Так, для внешних множеств не выполняется принцип фундирования.

Язык теории EХТ повторяет язык IST, только к нему добавляется еще один предикат — Int , выражающий свойство быть внутренним множеством.

Таким образом, EXT — это теория ZF, плюс два предиката St и Int , плюс связывающие их аксиомы. Интуитивно считается, что в теории EXT предметной областью переменных является универсум всех внешних множеств $V^{Ext} \equiv \{x \mid x = x\}$, который содержит в себе класс внутренних множеств $V^{Int} \equiv \{x \mid Int(x)\}$, который, в свою очередь, содержит класс всех стандартных множеств $V^{St} \equiv \{x \mid St(x)\}$.

Будем также говорить, что a имеет *стандартный размер* и записывать $a \in V^{size}$, если $a = \text{ran}(f)$, где внешняя функция $f : \circ x \rightarrow a$, x стандартно и $\circ x = \{y \in x \mid St(y)\}$ — стандартное ядро x . Иначе говоря, a можно перечислить внешней функцией со стандартной областью определения, в которой каждый элемент стандартен.

Аналогично определению кванторов $\forall^{St}, \exists^{St}$ в теории INT здесь также определяются кванторы $\forall^{Int}, \exists^{Int}$. Существуют правила для *стандартизации* и *интернализации* формул ZF путем замены всех кванторов на кванторы с указанными значками по тем же правилам. Стандартизованная формула φ обозначается φ^{St} и называется еще *релятивизацией* φ на универсум V^{St} , интернализованная формула φ обозначается φ^{Int} и называется также *релятивизацией* φ на универсум V^{Int} .

В теории EXT принимаются следующие аксиомы:

EXT1 **аксиома объемности**: $(a = b) \leftrightarrow \forall x : (x \in a \leftrightarrow x \in b)$;

EXT2 **аксиома пары**: $\{a, b\} \in V^{Ext}$;

EXT3 **аксиома объединения**: $\cup a \in V^{Ext}$;

EXT4 **аксиома степени**: $\mathcal{P}(a) \in V^{Ext}$;

EXT5 **аксиома свертывания**: $\{x \in a \mid \varphi(x)\} \in V^{Ext}$, где φ — произвольная формула языка EXT;

EXT6 **аксиома полного упорядочения** (теорема Цермело): каждое внешнее множество может быть вполне упорядочено;

EXT7 **принцип моделирования**: выполняются все аксиомы ZFC, релятивизированные в V^{Int} , т. е. V^{Int} — это универсум теории ZFC;

EXT8 **аксиома транзитивности**: $(x \in V^{Int}) \rightarrow (x \subset V^{Int})$;

EXT9 **аксиома вложения**: $V^{St} \subset V^{Int}$;

EXT10 **принцип переноса**: $\forall^{St} x_1 \dots \forall^{St} x_n \varphi^{St}(x_1, \dots, x_n) \leftrightarrow \varphi^{Int}(x_1, \dots, x_n)$ для каждой формулы φ теории ZFC;

EXT11 **принцип идеализации:** $\forall^{Int} x_1 \dots \forall^{Int} x_n \forall a \in V^{size} : (\forall^{fin} z \subset a : \exists^{Int} x \forall y \in z \varphi^{Int}(x, y, x_1, \dots, x_n)) \rightarrow \rightarrow (\exists^{Int} x \forall^{Int} y \in a : \varphi^{Int}(x, y, x_1, \dots, x_n))$ для каждой формулы φ теории ZFC;

EXT12 **принцип стандартизации:** $\forall a \exists^{St} b \forall^{St} x (x \in a \leftrightarrow x \in b)$ — для любого внешнего множества a существует его стандартизация $*a = b$.

Достойным упоминания является такое свойство этой системы:

$$\forall^{Int} x_1 \dots \forall^{Int} x_n \varphi(x_1, \dots, x_n) \leftrightarrow \varphi^{Int}(x_1, \dots, x_n),$$

$$\forall^{St} x_1 \dots \forall^{St} x_n \varphi^{St}(x_1, \dots, x_n) \leftrightarrow \varphi^{Int}(x_1, \dots, x_n) \leftrightarrow \varphi(x_1, \dots, x_n),$$

иначе говоря, любое «ограниченное» свойство стандартных множеств можно без потери истинности и общности выражать как в терминах внутренних, так и внешних элементов. Например,

$$x \subset y \leftrightarrow x \subset^{Int} y \leftrightarrow x \subset^{St} y$$

для стандартных множеств x, y .

Для теории EXT существует теорема Хрбачека, утверждающая, что EXT является консервативным расширением ZFC. Иначе говоря, если формула φ записана на языке ZFC, то

$$\varphi \text{ — теорема ZFC} \leftrightarrow \varphi^{Int} \text{ — теорема EXT} \leftrightarrow \varphi^{St} \text{ — теорема EXT}.$$

В то же время, EXT не является расширением IST, поскольку мир внутренних множеств V^{Int} не является моделью внутренних множеств Нельсона, т. к. принципы стандартизации и идеализации в этих теориях имеют различные формулировки.

Заключительное слово

Мы не станем утруждать читателя нагромождением теорий нестандартного анализа и нестандартной теории множеств, отсылая к упомянутым выше книгам. Скажем только, что приведенная до этого конструкция \mathbb{R} имеет далеко идущие обобщения в топологии, где ультрафильтры играют особенную роль.

Кроме того, нестандартный анализ позволяет намного проще формулировать известные результаты анализа. Платой за это упрощение является внедрение довольно сложных модельных конструкций, о которых вкратце было сказано выше. Возможно, по этой причине «нестандартные» учебники анализа так и не прижились в современной математике.

Наконец, всякий раз при появлении новых теорий логично ожидать и новых результатов (как в случае теории относительности, например). Действительно, такие результаты были получены, но большинство из них впоследствии были передоказаны «стандартными» методами, так что и здесь революция не случилась.

Далее мы рассмотрим еще одну модель гипердействительных чисел, но уже в совсем другой их конструкции, не требующей (для определения) ни аксиомы выбора, ни, тем более, выхода за рамки ZFC. Мало того, это будут числа настолько превосходящие \mathbb{R} , насколько универсум всех множеств превосходит первый бесконечный ординал ω .

2.4.6 Сюрреальные числа

Подобно тому, как любую группу можно представить как подгруппу группы биекций, точно так же все *упорядоченные* поля являются подполями некоторого всеобъемлющего поля-монстра, называемого также универсальным упорядоченным полем.¹⁸

Хм... группу-монстр знаем, а вот поле...

На самом деле это гигантское поле является примерно тем же самым для всех ординалов, чем поле действительных чисел для всех натуральных чисел. Мы уже ранее¹⁹ делали беглый обзор кольца ординалов, основанного на «нормальной» арифметике (в смысле, с операциями «нормальное сложение» и «нормальное умножение», восходящими к формальной записи ординалов в виде канторовской нормальной формы). Здесь мы расширим арифметику как на весь класс ординалов (вширь), так и на бесконечно малые величины, порядок малости которых также будет носить ординальный характер.

Такое поле по понятным причинам не является множеством, поскольку оно включает в себя числа, порядково изоморфные классу ординалов. Это поле является собственно классом в теории множеств Гёделя—Бернайса, а все операции на нем, строго говоря, не являются функциями, однако на каждом выделенном участке этого поля мы вполне можем ограничиваться множествами чисел и функциями над ними.

Идея построения поля сюрреальных чисел восходит к идее построения действительных чисел как дедекиндовых сечений, когда каждое действительное число определяется как сечение поля рациональных чисел. Но в чистой теории множеств это построение стартует не с какого-то готового поля, а с пустого множества. По Конвею [30] их построение производится рекурсивно с одновременным определением порядка на числах. Это построение напоминает определение универсумов множеств (раздел 1.4).

Для легкого и непринужденного ознакомления с сюрреальными числами рекомендуем книгу Д. Кнута [22].

¹⁸В теории Гёделя—Бернайса.

¹⁹См. стр. 143

Генерация и порядок

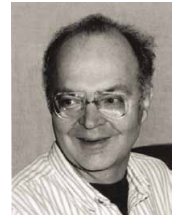
Для начала договоримся о следующем: строчными латинскими буквами мы обозначаем сюрреальные числа, а прописными латинскими буквами — множества сюрреальных чисел (классов тут пока нет). Кроме того, запись $x \not\leq Y$ означает, что $\forall y \in Y : \neg(x \leq y)$. Аналогично определяется $Y \not\leq x$. Заметим, что эти обозначения нельзя путать с $\neg(x \leq Y)$ и $\neg(Y \leq x)$.

Так же, как нельзя путать \forall и \exists .

SUR1 Если L и R — какие-то множества сюрреальных чисел (в том числе пустые), причем $\forall x \in L \forall y \in R : \neg(y \leq x)$, то пара (L, R) является записью сюрреального числа, при этом X_L обозначает множество L , а X_R — множество R , если $x = (L, R)$.

SUR2 $x \leq y$, если $y \not\leq X_L$ и $Y_R \not\leq x$.

Итак, мы видим, что каждое новое сюрреальное число определяется через два множества «ранее созданных» сюрреальных чисел, а их сравнение определяется через сравнение «ранее созданных» чисел. От обычного рекурсивного определения есть только одно отличие — мы не нумеруем вновь создаваемые множества чисел ординалами и, как следствие, в рекурсии отсутствует вариант их построения для предельного ординального числа. На самом деле, мы здесь просто неявно пользуемся аксиомой бесконечности и, следовательно, возможностью формировать бесконечные множества сюрреальных чисел. Впрочем, мы ведь и ординалы определяли не с помощью трансфинитной индукции, что, однако, не помешало нам построить сколь угодно большие бесконечные ординалы.



Дональд
Эрвин Кнут

Процесс генерации сюрреальных чисел от «старых» к «новым» с помощью ограниченного набора правил возвращает нас к истокам — процессу генерации сущностей с помощью правил грамматики, который мы рассматривали в разделе 1.1.1.

Условимся в дальнейшем класс всех сюрреальных чисел обозначать \mathbf{No} , так что запись $s \in \mathbf{No}$ эквивалентна высказыванию « s есть сюрреальное число».

Для сюрреального числа, образованного парой множеств L и R вместо (L, R) принято писать $\{L \mid R\}$, подчеркивая тем самым, что новое число помещается между L и R , а не где-то этажом выше.

Не путаем с термом-квантором $\{x \mid \varphi(x)\}$!

Самое первое сюрреальное число, в соответствии с определением, — это $\{\emptyset \mid \emptyset\}$, причем чаще всего пустое множество вообще опускается в обозначениях, так что мы имеем число $\{ \mid \}$, которое обозначается 0 и в дальнейшем оно играет роль нуля в поле сюрреальных чисел. В силу SUR2 нетрудно ви-

деть, что $0 \leq 0$, поскольку $0_L = \emptyset = 0_R$ и, стало быть, неверно, что $0 \leq x$, где $x \in 0_L$, т. е. $0 \not\leq 0_L$,²⁰ и также неверно, что $y \leq 0$, где $y \in 0_R$, т. е. $0_R \not\leq 0$.

Имея число 0, мы теперь можем создавать новые сюрреальные числа, например, полагая $L = \{0\}$ и $R = \emptyset$. По правилам такое число записывается термом $\{\{0\} \mid \}$, однако и тут мы принимаем некоторые сокращения, а именно, мы опускаем фигурные скобки в записи левого и правого множеств, если они заданы списком. То есть, указанное выше число записывается так: $\{0 \mid \}$. Итак, благодаря нулю, мы уже можем построить два новых числа, сразу введя их обозначения:

$$-1 \doteq \{ \mid 0 \}, \quad 1 \doteq \{0 \mid \}.$$

Заметим, что $\{0 \mid 0\}$ не будет являться корректной записью сюрреального числа, поскольку $0 \leq 0$ (по правилу SUR1 требуется, чтобы элементы правого множества не были больше или равны элементам левого, но таковыми элементами являются нули, а они сравнимы).

Помимо обозначения $x \leq y$ можно пользоваться обозначением равенства $x \doteq y$ ($x \leq y \wedge y \leq x$) и строгого неравенства $x < y$ ($x \leq y \wedge y \not\leq x$), а также их отрицаниями и обращениями. Нужно отметить, что мы здесь использовали новый символ равенства, потому что его определение не опирается на равенство x и y в смысле аксиоматики ZF. Строго говоря, равенство \doteq является разновидностью отношения эквивалентности. Оно соотносится с равенством множеств так же, как равенство множеств соотносится с тождеством записей множеств, которое мы обозначали \equiv . Производным от равенства сюрреальных чисел будет и равенство множеств сюрреальных чисел: $A \doteq B$, если $\forall a \in A \exists b \in B : a \doteq b$ и $\forall b \in B \exists a \in A : a \doteq b$. При этом рефлексивность, симметричность и транзитивность равенств чисел и множеств чисел выполняются, и если числа или множества чисел равны в смысле теоретико-множественного равенства, то они равны и в смысле \doteq .

Аналогично связываются отношениями порядка множества чисел: $A < x$, если $A \leq x \wedge x \not\leq A$, т. е. $\forall a \in A : a < x$. И, кроме того, $A < B$ означает $A \leq B \wedge B \not\leq A$, т. е. $\forall a \in A \forall b \in B : a < b$.

Предлагаем доказать следующие утверждения:

$$-1 \doteq -1, \quad 0 \doteq 0, \quad 1 \doteq 1, \quad -1 < 0, \quad -1 < 1, \quad 0 < 1.$$

*Упражнение
2.36.
Оперируя
только
правилами
SUR1–2!*

²⁰Формально, $(0 \not\leq 0_L) \leftrightarrow (\forall x (x \in 0_L) \rightarrow \neg(0 \leq x))$. Импликация, стоящая под квантором, истинна независимо от x , поскольку высказывание $(x \in 0_L)$ ложно, т. к. $0_L = \emptyset$. Вообще, всегда $x \not\leq \emptyset$ и $\emptyset \not\leq x$.

Для целых чисел и ординалов имеем:

$$\begin{aligned} 2 &\doteq \{1|\}, & 3 &\doteq \{2|\}, & 4 &\doteq \{3|\}, & 5 &\doteq \{4|\}, \\ \omega &\doteq \{0, 1, 2, \dots, n, \dots|\}, & \alpha &\doteq \{\{\lambda| \lambda < \alpha\}|\}, \\ -2 &\doteq \{|\ -1\}, & -3 &\doteq \{|\ -2\}, & -4 &\doteq \{|\ -3\}, & -5 &\doteq \{|\ -4\}, \\ -\omega &\doteq \{0, -1, -2, \dots, -n, \dots\}, & -\alpha &\doteq \{|\{-\lambda| \lambda < \alpha\}\}, \end{aligned}$$

и т.д., причем в данном случае исходные ординалы (в том числе конечные) выступают в роли *обозначений* для их полных аналогов в системе сюрреальных чисел, с теоретико-множественной точки зрения ординалы, конечно же, не являются сюрреальными числами в смысле равенства множеств.

Стоит заметить, что $\{2|\} \doteq \{1, 2|\} \doteq \{0, 1, 2|\}$, а также $\{\mathbb{Z}|\} \doteq \omega$, где под \mathbb{Z} мы понимаем множество сюрреальных чисел $\{0, \pm 1, \pm 2, \dots\}$. Как и при записи множеств и мультимножеств у нас появляются эквивалентные записи, представляющие одно и то же число. Как и в случае с обозначениями $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}$, вложение $\mathbb{Z} \subset \mathbf{No}$ означает, что в качестве \mathbb{Z} была выбрана конкретная реализация понятия \mathbb{Z} в классе сюрреальных чисел.

Как видим, до сих пор одно из множеств-компонент у нас оставалось пустым, хотя мы уже построили немало. Тем не менее, вот еще примеры:

$$\frac{1}{2} \doteq \{0|1\}, \quad -\frac{1}{2} \doteq \{-1|0\}, \quad \frac{1}{4} \doteq \{0|\frac{1}{2}\}, \quad -\frac{1}{4} \doteq \{-\frac{1}{2}|0\}.$$

Можно задать вопрос: почему, собственно, между 0 и 1 мы выбрали число $\frac{1}{2}$, а не $\frac{1}{3}$ или $\frac{1}{128}$.

На самом деле, ответ на этот вопрос дает утверждение $\{0|1\} + \{0|1\} \doteq 1$, которое вытекает из арифметики сюрреальных чисел (они будут даны ниже).

Поэтому за конечное число шагов, отправляясь от нуля, мы можем построить только числа вида $n/2^k$, где n — целое, $k = 0, 1, 2, \dots$. Такие числа называются **двоично-рациональными** и образуют коммутативное кольцо с единицей в поле \mathbb{Q} .

Для сюрреальных чисел вводится понятие «**день рождения**». Под этим понимается (ординальный) номер шага, на котором это число было создано (родилось). Так, число 0 имеет день рождения 0, числа 1 и -1 — первый день рождения, числа $2, \frac{1}{2}, -2, -\frac{1}{2}$ — второй день рождения, все двоично-рациональные числа, и только они, имеют конечный день рождения, ω и $-\omega$ родились в день ω , и т.д. На самом деле, мы вновь находим аналог в построении универсальных множеств и множеств вообще, поскольку каждое множество, как мы уже говорили, имеет ординальный ранг, показывающий глубину его сложности (или «этаж» в башне универсумов). В этом смысле ранг множества можно также называть днем его рождения.

Следующие свойства можно найти в [41] (или попробовать доказать самостоятельно):

*Bon
Anniversaire
à tous!*

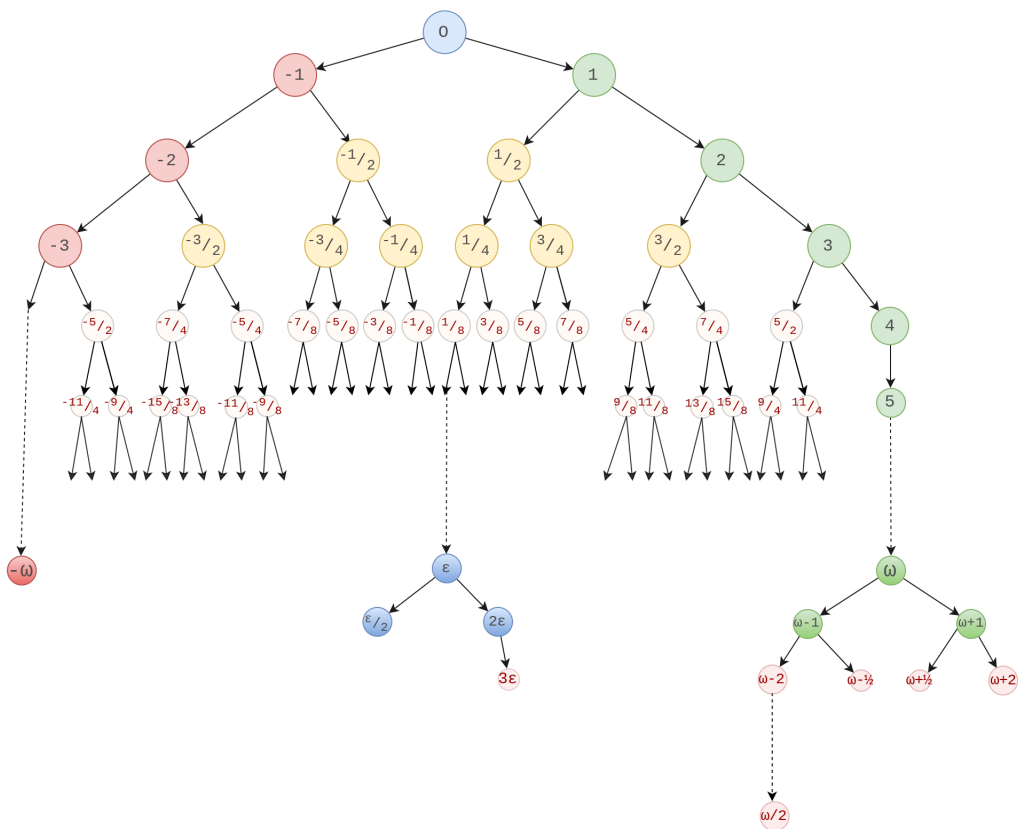


Рис. 2.8: Дерево сюрреальных чисел.

SF1 $x \doteq x$;

SF2 если $X_L \doteq Y_L$ и $X_R \doteq Y_R$, то $x \doteq y$;

SF3 $X_L < x < X_R$;

SF4 если $x \leq y \wedge y \leq z$, то $x \leq z$;

SF5 если $x < y \wedge y < z$, то $x < z$;

SF6 если $x \not\leq y$, то $y \leq x$;

SF7 $x < y$ тогда и только тогда, когда $y \not\leq x$;

SF8 если $x \doteq \{X_L \mid X_R\}$, то удаление из X_L любого элемента, кроме максимального (если он есть), а также удаление из X_R любого элемента, кроме минимального (если он есть), не меняет x в смысле равенства

\doteq , в частности, если $x' \doteq \max X_L$, то $x \doteq \{x' \mid X_R\}$, аналогично, если $x'' \doteq \min X_R$, то $x \doteq \{X_L \mid x''\}$;

SF9 если $A < x < B$, то $x \doteq \{X_L \cup A \mid X_R \cup B\}$;

SF10 если x — самое старое (с минимальным днем рождения) число между числами a и b ($a < x < b$), то $x \doteq \{a \mid b\}$.

Сложение и вычитание

Положим $A + b \rightleftharpoons \{a + b \mid a \in A\}$ и $a + B \rightleftharpoons \{a + b \mid b \in B\}$, $-A \rightleftharpoons \{-a \mid a \in A\}$ и $-a \rightleftharpoons \{-A_L \mid -A_R\}$. Кроме того, сложение с пустым множеством есть пустое множество: $\emptyset + a \rightleftharpoons \emptyset$, $a + \emptyset \rightleftharpoons \emptyset$. Пользуясь данными обозначениями, определим сложение сюрреальных чисел как

$$x + y \rightleftharpoons \{X_L + y \cup x + Y_L \mid X_R + y \cup x + Y_R\},$$

а вычитание как $x - y \rightleftharpoons x + (-y)$.

Приведем ряд примеров сложения:

$$0 + 0 \doteq \{\emptyset + 0 \cup 0 + \emptyset \mid \emptyset + 0 \cup 0 + \emptyset\} \doteq \{\mid\} \doteq 0$$

$$0 + 1 \doteq \{\emptyset + 1 \cup \{0 + 0\} \mid \emptyset + 1 \cup 0 + \emptyset\} \doteq \{0 \mid\} \doteq 1$$

$$0 + \frac{1}{2} \doteq \{\emptyset + \frac{1}{2} \cup \{0 + 0\} \mid \emptyset + \frac{1}{2} \cup \{0 + 1\}\} \doteq \{0 + 0 \mid 0 + 1\} \doteq \{0 \mid 1\} \doteq \frac{1}{2}$$

$$1 + 1 \doteq \{0 + 1, 1 + 0 \mid \emptyset + 0 \cup 0 + \emptyset\} \doteq \{1 \mid\} \doteq 2$$

Как видим, дело это нелегкое и больше смахивает на суровый вычислительный алгоритм для нашего главного героя книги, чем на красивое математическое определение, тем не менее, это сложение обладает очень важными математическими свойствами. А именно, справедлива

Теорема 2.14. *Сюрреальные числа образуют коммутативную группу по сложению.*

Здесь требуется пояснить, что сюрреальные числа не образуют множество, поэтому, строго говоря, назвать их класс группой было бы неверно. Однако, в определении группы мы опираемся только на свойства операции, а они не связаны с неограниченными в классе **No** подклассами. Говоря, что класс является группой, кольцом или полем, мы имеем ввиду формальное выполнение аксиом группы, кольца или поля для заданных операций.

За доказательством данной теоремы мы отсылаем к [41].

Умножение и деление

Более сложным выглядит определение умножения. Положим $A \cdot b \rightleftharpoons \{a \cdot b \mid a \in A\}$ и $a \cdot B \rightleftharpoons \{a \cdot b \mid b \in B\}$, кроме того, $A + B \rightleftharpoons \{a + b \mid a \in A \wedge b \in B\}$ и

$A \cdot B \equiv \{a \cdot b \mid a \in A \wedge b \in B\}$. Наконец, умножение на пустое множество есть пустое множество: $\emptyset \cdot a \equiv \emptyset$, $a \cdot \emptyset \equiv \emptyset$. Пользуясь данными обозначениями, определим умножение сюрреальных чисел как

$$x \cdot y \equiv \{(X_L \cdot y + x \cdot Y_L - X_L \cdot Y_L) \cup (X_R \cdot y + x \cdot Y_R - X_R \cdot Y_R) \mid (X_L \cdot y + x \cdot Y_R - X_L \cdot Y_R) \cup (X_R \cdot y + x \cdot Y_L - X_R \cdot Y_L)\}.$$

Предлагаем читателю самостоятельно доказать следующие простые факты:

Упражнение 2.37.

$$\begin{aligned} x \cdot y &\doteq y \cdot x \\ x \cdot 0 &\doteq 0 \doteq 0 \cdot x \\ x \cdot 1 &\doteq x \doteq 1 \cdot x \\ x \cdot (y + z) &\doteq x \cdot y + x \cdot z \end{aligned}$$

Для определения обратного по умножению элемента $\frac{1}{x}$ используется рекурсивное построение его левого и правого множеств, суть которого сводится к тому, чтобы получить две последовательности, сходящиеся, соответственно, снизу и сверху к числу $\frac{1}{x}$. Подробности можно найти в [41].

Приведем ряд примеров. Поскольку среди сюрреальных чисел с датой рождения $< \omega$ есть только двоично-рациональные числа, для определения дроби $\frac{1}{3}$ нужно указать последовательности двоично-рациональных чисел, сходящиеся к $\frac{1}{3}$:

$$\frac{1}{3} \doteq \left\{ \frac{1}{4}, \frac{5}{16}, \frac{21}{64}, \frac{85}{256}, \dots \mid \frac{1}{2}, \frac{3}{8}, \frac{11}{32}, \frac{43}{128}, \dots \right\},$$

иначе говоря, выбрав знаменатель 2^k , мы находим дробь вида $\frac{j}{2^k}$, ближайшую к $\frac{1}{3}$, и если она меньше $\frac{1}{3}$, то помещаем ее в левое определяющее множество, а если больше, то в правое. Понятно, что мы можем указывать и другие промежуточные числа (свойство SF9), а также стартовать с любого сколь угодно большого k (свойство SF8), от этого результат не зависит.

Еще пример:

$$\pi \doteq \left\{ \frac{3}{1}, \frac{25}{8}, \frac{201}{64}, \dots \mid \frac{13}{4}, \frac{101}{32}, \frac{3217}{1024}, \dots \right\},$$

здесь некоторые дроби со знаменателями 2^k совпадают с уже выписанными, например, $\frac{6}{2}$ совпадает с $\frac{3}{1}$, $\frac{25}{8} = \frac{50}{16}$, $\frac{201}{64} = \frac{402}{128} = \frac{804}{256} = \frac{1608}{512}$, поэтому они не включены в списки.

Из приведенных примеров видно, что любое действительное число можно представить с помощью двух счетных сходящихся к нему последовательностей двоично-рациональных чисел. Это действительно так, поскольку множество двоично-рациональных чисел всюду плотно в \mathbb{R} . Но тогда это означает, что все действительные числа, не являющиеся двоично-рациональными дробями, имеют дату рождения ω .

Некоторые бесконечности

Приведем несколько примеров арифметики бесконечных сюрреальных чисел. Но для начала отметим, что сюрреальные числа, обозначенные ординалами, не обладают свойством транзитивности, т. е. для сюрреальных чисел неверно будет равенство $\alpha \doteq \{\lambda \mid \lambda < \alpha\}$, столь привычное нам из теории порядковых чисел. Поэтому любое обозначение ординала в арифметике и обозначениях сюрреальных чисел нужно воспринимать как число («точку»), но не как множество.

Итак,

$$\begin{array}{ll} \omega \doteq \{\mathbb{Z} \mid \} & -\omega \doteq \{\mid \mathbb{Z}\} \\ \omega + 1 \doteq \{\omega \mid \} & \omega - 1 \doteq \{\mathbb{Z} \mid \omega\} \\ \omega + 2 \doteq \{\omega + 1 \mid \} & \omega - 2 \doteq \{\mathbb{Z} \mid \omega - 1\} \\ \omega + 3 \doteq \{\omega + 2 \mid \} & \omega - 3 \doteq \{\mathbb{Z} \mid \omega - 2\} \\ \omega + \omega \doteq \{\mathbb{Z} + \omega \mid \} & \omega - \omega \doteq 0 \end{array}$$

где $\mathbb{Z} + \omega$ — это множество всех чисел вида $\omega, \omega \pm 1, \omega \pm 2, \dots$. Продолжим:

$$\begin{array}{ll} 2\omega \doteq \{\mathbb{Z} + \omega \mid \} & \omega^2 \doteq \{\omega, 2\omega, 3\omega, \dots \mid \} \doteq \{\mathbb{Z} \cdot \omega \mid \} \\ 3\omega \doteq \{\mathbb{Z} + 2\omega \mid \} & -\omega^2 \doteq \{\mid -\omega, -2\omega, -3\omega, \dots\} \doteq \{\mid \mathbb{Z} \cdot \omega\} \\ -2\omega \doteq \{\mid \mathbb{Z} - \omega\} & \omega^3 \doteq \{\omega^2, 2\omega^2, 3\omega^2, \dots \mid \} \doteq \{\mathbb{Z} \cdot \omega^2 \mid \} \end{array}$$

$$\begin{array}{lll} -3\omega \doteq \{\mid \mathbb{Z} - 2\omega\} & \omega^\omega \doteq \{\omega, \omega^2, \omega^3, \dots \mid \} & \\ \frac{\omega}{2} \doteq \{\mathbb{Z} \mid \omega - \mathbb{Z}\} & \sqrt{\omega} \doteq \{\mathbb{Z} \mid \omega, \frac{\omega}{2}, \frac{\omega}{3}, \dots\} & \frac{1}{\omega} \doteq \{0 \mid 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \dots\} \end{array}$$

Таким образом, мы построили такую систему чисел, которая позволяет производить любые алгебраические операции над ординалами! Стоит отметить также, что арифметические операции над ординалами, как сюрреальными числами, совпадают с натуральными арифметическими операциями, которые

мы определяли, основываясь на канторовской нормальной форме представления ординалов. При этом, конечно же, «родные» операции над ординалами не соответствуют натуральным и сюрреальным операциям.

Число $1/\omega$ еще обозначают ε , что отсылает нас к теории пределов в анализе. Действительно, $\varepsilon \doteq 1/\omega$ больше нуля, но меньше любого положительного действительного числа. Бесконечно малая окрестность произвольного действительного числа x в \mathbf{No} представляется интервалом сюрреальных чисел $(x - \varepsilon; x + \varepsilon)$.²¹ Некоторые арифметические свойства ε :

$$\begin{aligned} 2\varepsilon &\doteq \{\varepsilon \mid 1 + \varepsilon, \frac{1}{2} + \varepsilon, \frac{1}{4} + \varepsilon, \frac{1}{8} + \varepsilon, \frac{1}{16} + \varepsilon, \dots\} \\ \frac{\varepsilon}{2} &\doteq \{0 \mid \varepsilon\} \\ \sqrt{\varepsilon} &\doteq \{\varepsilon, 2\varepsilon, 3\varepsilon, 4\varepsilon, \dots \mid 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \dots\} \end{aligned}$$

Другие подходы

Сюрреальные числа можно построить не только тем способом, который предложил Конвей. Рассмотрим ординальные последовательности знаков:

$$x : \alpha \rightarrow B, \quad B = \{-1, +1\},$$

где α — произвольный ординал.²² Кроме того, мы полагаем, что

$$x|_{\lambda}(\beta) = \begin{cases} x(\beta), & \text{если } \beta < \alpha \\ 0, & \text{если } \alpha \leq \beta < \lambda \end{cases}$$

Иначе говоря, $x|_{\lambda}$ — это либо сужение функции x на ординал λ , если $\lambda < \alpha$, либо продолжение x на λ нулем, если $\lambda > \alpha$. Мы считаем равными функции x и $x|_{\lambda}$ при $\lambda > \alpha$. Такое продолжение нулем делается исключительно для удобства записи арифметических операций.

Ясно, что при фиксированном α все x , определенные на α , образуют множество, в то время как вообще все знаковые последовательности образуют собственно класс.

Для любых двух знаковых последовательностей определим их точку расхождения:

$$\sigma(x, y) \doteq \min\{\beta \mid x(\beta) \neq y(\beta)\},$$

²¹Этот интервал является собственно классом, если мы не зададим ограничение на возраст его элементов. В данном случае можно ограничиться числами с днем рождения до $\varepsilon_0 = \omega \uparrow \omega$.

²²В частности, может быть $\alpha = 0$, тогда мы имеем пустую последовательность.

т. е. это первый ординал, где x и y различаются. В частности, в этой точке одна из функций может быть не определена, в этом случае мы считаем ее значение равным нулю, что позволяет легко ввести сравнение:

$$x < y, \text{ если } x(\sigma(x, y)) < y(\sigma(x, y)).$$

Упражнение 2.38. Предлагаем читателю в качестве упражнения доказать, что такое отношение будет линейным порядком на всех последовательностях.

Пусть L и R — непустые множества знаковых последовательностей, причем $L < R$ (т. е. $\forall x \in L \forall y \in R x < y$).

Теорема 2.15. *Существует единственная последовательность $z = \sigma(L, R)$ такая, что:*

- (1) $L < z < R$ (z разделяет множества),
- (2) для любого w : если $L < w < R$, то $z \subseteq w$ (требование минимальности)

Здесь вложение последовательностей $z \subseteq w$ теоретико-множественное и по сути означает, что z является сужением w на меньший или равный ординал (по отношению к области определения w).

Доказательство. Требуемую последовательность можно построить следующим способом: пусть α — минимальный ординал такой, что $L|_\alpha < R|_\alpha$, где сужение множеств означает сужение всех его элементов. В этом случае при любом $\lambda < \alpha$ существует единственная последовательность $z_\lambda : \lambda \rightarrow B$ такая, что $L|_\lambda \cap R|_\lambda = \{z_\lambda\}$, причем все последовательности z_λ согласованы, т. е. $z_\lambda \subset z_\gamma$, если $\lambda < \gamma < \alpha$.

Далее, если $\alpha = \beta + 1$, то положим $z' = z_\beta$, а в случае предельного α положим $z' = \bigcup_{\lambda < \alpha} z_\lambda$. Последовательность z' будет либо максимумом $L|_\alpha$, либо минимумом $R|_\alpha$, либо $L|_\alpha < z' < R|_\alpha$. Одновременное выполнение этих событий исключено.

В последнем случае мы нашли $z = z'$. Пусть $z' = \max L|_\alpha$. В этом случае нужно продлить z' так, чтобы он на один шаг отошел от $\max L|_\alpha$. Ясно, что в L может быть целый пучок последовательностей, сужение которых на α совпадает с z' , причем некоторые из них могут принимать значение $+1$ на ординалах $\alpha, \alpha + 1, \alpha + 2, \dots$. Пусть γ — наименьший такой ординал $\geq \alpha$, где эти последовательности перестают принимать значение $+1$ (либо уходят в -1 хотя бы на 1 шаг, либо уходят в 0 , т. е. обрываются). Тогда доопределим z' на все точки $\lambda, \alpha \leq \lambda < \gamma$, значением $+1$ (если $\gamma = \alpha$, то такое продление не потребуется) и получим последовательность z'' .

Далее, в точке γ либо все продолжения z'' из L равны -1 , и тогда полагаем $z = z''$, либо есть какое-то продление в L , которое принимает значение 0 , т. е. $z'' \in L$, тогда положим $z(\gamma) = +1$ и $z|_\gamma = z''$. В этом случае z гарантированно будет $> L$ и, кроме того $z < R$, поскольку $z' < R$.

В случае $z' = \min R|_\alpha$ действуем симметричным образом. \square

Пусть далее для последовательности x :

$$\begin{aligned} L(x) &\equiv \{x|_\alpha \mid \alpha < \text{dom}(x) \wedge x(\alpha) = -1\} \\ R(x) &\equiv \{x|_\alpha \mid \alpha < \text{dom}(x) \wedge x(\alpha) = +1\}, \end{aligned}$$

нетрудно показать, что $\sigma(L(x), R(x)) = x$.

Сложение последовательностей определяется рекурсивно следующим образом: $x + y = \sigma(L, R)$, где

$$\begin{aligned} L &\equiv \{u + y \mid u \in L(x)\} \cup \{x + v \mid v \in L(y)\}; \\ R &\equiv \{u + y \mid u \in R(x)\} \cup \{x + v \mid v \in R(y)\}. \end{aligned}$$

Умножение последовательностей определяется рекурсивно следующим образом: $xy = \sigma(L, R)$, где

$$\begin{aligned} L &\equiv \{uy + xv - uv \mid u \in L(x), v \in L(y)\} \cup \{uy + xv - uv \mid u \in R(x), v \in R(y)\}; \\ R &\equiv \{uy + xv - uv \mid u \in L(x), v \in R(y)\} \cup \{uy + xv - uv \mid u \in R(x), v \in L(y)\}. \end{aligned}$$

Рекурсия здесь заключается в том, что все операции определяются через самих себя, но для более коротких последовательностей.

Соответствие f между сюрреальными числами по Конвею и знаковыми последовательностями задается как

$$f(\{L \mid R\}) = \sigma(M, S), \text{ где } M = \{f(x) \mid x \in L\}, S = \{f(x) \mid x \in R\}.$$

Обратное соответствие g :

$$g(x) = \{L \mid R\}, \text{ где } L = \{g(y) \mid y \in L(x)\}, R = \{g(y) \mid y \in R(x)\}.$$

Оба определения также рекурсивны.

Alling [29] предложил аксиоматический подход к определению сюрреальных чисел в рамках теории множеств Гёделя–Бернайса, полагая, что классом сюрреальных чисел называется всякий класс No с заданным на нем отношением-классом $<$ и отображением-классом b сюрреальных чисел в ординалы (называемым функцией даты рождения) такими, что:

$A1 <$ есть линейный порядок на No ;

A2 если A и B — два подкласса \mathbf{No} и $A < B$, то существует $z \in \mathbf{No}$ такой, что $A < z < B$ и если $A < w < B$, то $b(z) \leq b(w)$; кроме того, если ординал $\alpha > b(x)$ для всех $x \in A \cup B$, то $b(z) \leq \alpha$.

Обе модели сюрреальных чисел — Конвея и знаковых последовательностей — удовлетворяют данным аксиомам.

Если сюрреальное число записано знаковой последовательностью, то такая запись имеет две простые интерпретации.

Во-первых, она указывает путь в дереве сюрреальных чисел, начиная от нуля, где $+1$ означает поворот направо, а -1 — поворот налево в бинарном дереве (см. рисунок 2.8). В данном дереве родителем каждого ненулевого числа является ближайшее число предыдущего поколения (для чисел с предельным днем рождения родителем является путь из ближайших к нему чисел предыдущих поколений).

Во-вторых, знаковая интерпретация позволяет довольно просто вычислять значение положительного двоично-рационального числа (правило Берлекампа). А именно, пусть последовательность имеет вид:

$$\underbrace{++ \cdots ++}_{n} (+-) \underbrace{+++ - - - -}_{11100001/2^k} < + > .$$

Тогда мы находим в ней первое вхождение $(+-)$, считая его дробной запятой. Количество плюсов перед разделителем отвечает за целую часть, знаки $+$ и $-$ после разделителя заменяем двоичными цифрами 1 и 0, соответственно, и в конце дописываем 1. Двоичное число (с учетом добавленной 1) ставится в числитель дробной части, а в знаменатель — 2^k , где k — количество двоичных знаков (с учетом добавленной 1).

Например,

$$\begin{aligned} (+-) + - &= 0 + \frac{101_2}{2^3} = \frac{5}{8} \\ ++ (+-) + &= 2 + \frac{11_2}{2^2} = 2\frac{3}{4} \end{aligned}$$

Для отрицательных чисел поступаем следующим образом: заменяем все знаки на противоположные, находим соответствующее положительное число и ставим перед ним минус.

Заключительные свойства

Пожалуй, главным достоинством сюрреальных чисел является следующая

Теорема 2.16. *No с операциями сложения и умножения образует упорядоченное поле.*

Мы вновь вынуждены уточнить, что поскольку \mathbf{No} не является множеством, то было бы правильнее не называть его полем, а просто выписать алгебраическое определение поля через свойства операций и порядка: сложение и умножение коммутативны, ассоциативны, связаны законом дистрибутивности, имеется 0 и 1, имеются противоположные и обратные (кроме как у нуля) числа, операции согласованы с порядком.

В отличие от \mathbb{R} , поле \mathbf{No} не является архимедовым.

За доказательством данной теоремы мы отсылаем к [41] и к [34]. Там же можно найти определение функции Дали $\delta(x)$, которая инъективно вкладывает \mathbb{R} в класс сюрреальных чисел с сохранением порядка и операций сложения и умножения. Таким образом, \mathbb{R} является подполем поля \mathbf{No} .

Стоит отметить, что построение чисел (натуральных, целых, рациональных и т.д.) мы начали с того, что отстранились от конкретной привязки вновь сгенерированных числовых систем к объектам теории множеств, предполагая, что всегда можно взять какую-то конкретную модель изучаемых чисел в \mathbf{ZF} . Однако с введением сюрреальных чисел мы в принципе можем отказаться от идеи определения числа как понятия, поскольку все числа вкладываются в сюрреальные (точнее, все линейно упорядоченные числовые структуры), т.е. имеют совершенно конкретное место в иерархии множеств \mathbf{ZF} . Ну, тут, как говорится, дело вкуса. Если Вы смотрите на математику глазами логика, т.е. из теории множеств, то такой подход Вам будет комфортнее, а если глазами алгебраиста — проще пользоваться аксиоматиками конкретных числовых систем и не искать их место в иерархии \mathbf{ZF} .

Интересны следующие факты, связывающие дни рождения чисел с их алгебраическими свойствами. Пусть α — произвольный ординал, а степени ординалов мы понимаем в их обычном смысле в рамках арифметики ординалов. Обозначим через \mathbf{No}_α множество всех сюрреальных чисел с датой рождения $< \alpha$. Такие множества можно назвать **сюрреальными универсумами**. Тогда:

Упражнение 2.39. Проверьте, что \mathbf{No}_α — действительно множество.

1. $\mathbf{No}_{\omega^\alpha}$ образует коммутативную группу по сложению;
2. $\mathbf{No}_{\omega^\alpha}$ образует коммутативное кольцо;
3. $\mathbf{No}_{\varepsilon_\alpha}$ образует поле.

Напомним, что $\varepsilon_0 = \omega \uparrow \uparrow \omega$, $\varepsilon_{\alpha+1} = \varepsilon_\alpha \uparrow \uparrow \omega$ и $\varepsilon_\alpha = \sup\{\varepsilon_\beta \mid \beta < \alpha\}$, если α — предельный ординал.

В частности, множество $\mathbf{No}_1 = \{0\}$ является тривиальной группой по сложению, множество всех двоично-рациональных чисел (\mathbf{No}_ω) является кольцом, а множество всех сюрреальных чисел с датой рождения меньше ε_0 является полем.

Универсум $\mathbf{No}_{\omega+1}$ не является ни полем, ни кольцом, ни даже группой по сложению, поскольку содержит два бесконечных числа ω и $-\omega$, и сложение с

ними натуральных чисел сразу же выводит за пределы $\mathbf{No}_{\omega+1}$. Однако внутри $\mathbf{No}_{\omega+1}$ «живут» такие поля как \mathbb{Q} и \mathbb{R} , причем \mathbb{R} является максимальным полем в $\mathbf{No}_{\omega+1}$, а их мощности совпадают.

Следующим замечательным фактом является

Теорема 2.17. *Любое вещественно замкнутое поле (с носителем-множеством) является собственным подполем \mathbf{No} .*

В частности, если мы принимаем аксиому выбора, то мы имеем возможность построить поля гипердействительных чисел (см. подраздел 2.4.5), и все они будут вкладываться в \mathbf{No} . Правда, моделями для них будут не любые поля вида $\mathbf{No}_{\varepsilon_\alpha}$, как может показаться сначала, а только поля со специально выбранными ε -числами.

Так, при принятии континуум-гипотезы \mathbf{CH} , единственной (с точностью до изоморфизма) моделью гипердействительных чисел, построенных на ультрастепенях, будет поле \mathbf{No}_{ω_1} , где ω_1 — первый несчетный ординал, который в силу \mathbf{CH} совпадает с $\mathfrak{c} = 2^{\aleph_0}$ (мощностью континуума).

При принятии аксиомы существования первого (строго) недостижимого кардинала (\mathbf{In}_0) моделью гипердействительных чисел в аксиоматике Кейслера [32] будет поле $\mathbf{No}_{\mathbf{In}_0}$. Интересно, что если не требовать от поля гипердействительных чисел, чтобы оно было множеством в \mathbf{ZF} , а могло быть собственным классом в теории Гёделя–Бернайса, то моделью становится все поле сюрреальных чисел. Существование такой модели не требует привлечения «проблематичных» аксиом вроде аксиомы существования недостижимого кардинала.

Достаточно подробно об этом можно прочесть в [32]. Там же дается описание и объяснение аксиоматики Кейслера.

2.4.7 Поле комплексных чисел

Для того чтобы немного отдохнуть от сложных понятий и в то же время подготовиться к переходу от линеаризованных числовых структур (в которых порядок согласован с операциями) к произвольным алгебраическим и геометрическим системам, мы позволим себе вместо философского заключения главы 2 дать здесь краткий обзор поля комплексных чисел — этого сияющего кристалла в короне царицы всех наук.

Необходимость в возникновении комплексных чисел связана, прежде всего, с таким досадным фактом, как невозможность решить уравнение $x^2 + 1 = 0$ в поле \mathbb{R} . Расширение \mathbb{R} с помощью корня этого уравнения, обозначенного i , привело к построению поля \mathbb{C} , удивительным свойствам и при-

*i — от фр.
imaginaire,
мнимый.*

менениям которого нет числа.

Стоит, однако, отметить, что поводом к построению комплексных чисел (как и в случае вещественных) является не только его алгебраическая непол-

нота — невозможность решить некое уравнение. Другим замечательным поводом является желание «закодировать» числами такие важные геометрические преобразования плоскости, как **параллельные переносы**, а также гомотетии (т. е. изменение масштаба) и повороты, называемые общим термином **поворотные гомотетии**. Действительно, ведь на вещественной прямой за параллельный перенос (т. е. попросту *сдвиг*) отвечает прибавление ко всем точкам прямой некоторого вещественного числа (вправо — положительного, а влево — отрицательного); за гомотетии также отвечают вещественные числа: любая гомотетия есть умножение всех точек на какое-то действительное (положительное) число; а за поворот (переворот на 180°) отвечает число -1 . Так что на прямой сдвиги и поворотные гомотетии полностью описываются сложением и умножением вещественных чисел. Хотелось бы получить нечто аналогичное и на плоскости.

Таким образом, возникновение комплексных чисел имеет (в равной степени) как алгебраические, так и геометрические причины. Но, поскольку наша книга носит более алгебраический, чем геометрический характер, вернемся к алгебраическим причинам появления новых чисел.

Итак, мы хотим ввести в употребление некоторое число i так, чтобы оно естественным образом сочеталось с действительными числами во всех алгебраических операциях, причем требуем выполнения равенства $i^2 = -1$. Добавление такого числа к \mathbb{R} аналогично добавлению $\sqrt{2}$ к \mathbb{Q} , в результате чего добавляются все числа вида $r + q\sqrt{2}$ с рациональными коэффициентами.

Дадим строгое определение. Пусть у нас имеется какая-то модель \mathbb{R} (например, та, что рассмотрена в предыдущем разделе). Рассмотрим множество $\mathbb{R} \times \mathbb{R}$, элементы которого мы будем записывать следующим образом:

$$x + iy \Rightarrow (x, y) \in \mathbb{R} \times \mathbb{R},$$

причем $x = \text{pr}_1(x + iy)$ будем обозначать $\Re(x + iy)$ и называть **действительной частью** числа $x + iy$, а $y = \text{pr}_2(x + iy)$ будем обозначать $\Im(x + iy)$ и называть **мнимой частью** числа $x + iy$.

Пусть далее $z = x + iy$, $z' = x' + iy'$. Положим по определению

$$\begin{aligned} z + z' &= (x + x') + i(y + y'), \\ zz' &= (xx' - yy') + i(xy' + x'y). \end{aligned}$$

Множество $\mathbb{R} \times \mathbb{R}$ с заданными на нем таким способом операциями сложения и умножения называется **полем комплексных чисел** и обозначается \mathbb{C} (конкретная модель \mathbb{R} нас при этом мало интересует).

Как и раньше, мы предполагаем, что $\mathbb{R} \subset \mathbb{C}$ означает некоторое естественное изоморфное вложение \mathbb{R} в \mathbb{C} . Точнее, под действительной осью в \mathbb{C} понимается множество $\{x + i0 \mid x \in \mathbb{R}\}$.

То, что поле \mathbb{C} действительно является полем, мы предлагаем в качестве разминки доказать читателю самостоятельно, пользуясь тем фактом, что \mathbb{R} — поле.

Свойства комплексных чисел

1. Формула Эйлера впервые появилась в работах Роджера Котса (помощника И. Ньютона), но в современном виде ее опубликовал Леонард Эйлер:

$$e^{ix} = \cos(x) + i \sin(x),$$

где число e (число Эйлера) определяется, например, как предел $\lim_{n \rightarrow \infty} (1 + \frac{1}{n})^n$, либо как такое основание показательной функции $f(x) = a^x$, при котором производная $f'(x) = f(x)$ (иначе говоря, функция e^x является неподвижной точкой оператора дифференцирования действительных функций).

При $x = \pi$ получаем **тождество Эйлера**, связывающее пять фундаментальных математических констант:

$$e^{i\pi} + 1 = 0.$$



Леонард
Эйлер

В общем виде произвольное комплексное число $z = x + iy$ представляется как:

$$z = re^{it} = r \cos(t) + ir \sin(t),$$

где $r = \sqrt{x^2 + y^2}$ называется **модулем комплексного числа** z и обозначается $|z|$, а угол t , тангенс которого равен²³ $\operatorname{tg}(t) = y/x$, называется **аргументом комплексного числа** и обозначается $\operatorname{arg}(z)$.

Связь комплексных чисел с геометрией плоскости заключается в том, что вектор, соединяющий начало координат и точку $z = x + iy$, имеет длину $|z|$ и угол наклона относительно оси Ox , равный $\operatorname{arg}(z)$.

Операция сопряжения: если $z = x + iy = re^{it}$, то $\bar{z} = x - iy = e^{-it}$. Комплексно сопряженные числа симметричны относительно действительной оси в \mathbb{C} .

На комплексную плоскость естественным образом (с использованием теоремы единственности) продлеваются действительные элементарные функции. Мы не будем здесь углубляться в особенности аналитического продолжения функций и описание ветвей таких функций как $\ln(z)$ или $\sqrt[n]{z}$, а также введением понятия риманова пространства. Приведем только ряд полезных

²³Более точно, при $x > 0$ аргумент определяется как $\operatorname{arctg}(y/x)$, а при $x < 0$ — как $\operatorname{arctg}(y/x) + \pi$, если $y > 0$, и как $\operatorname{arctg}(y/x) - \pi$, если $y \leq 0$.

соотношений:

$$\begin{aligned}
 e^z &= \sum_{n=0}^{\infty} \frac{z^n}{n!} & \ln z &= \ln |z| + i \arg(z) \\
 \cos z &= \frac{e^{iz} + e^{-iz}}{2} & \sin z &= \frac{e^{iz} - e^{-iz}}{2} \\
 \operatorname{ch} z &= \frac{e^z + e^{-z}}{2} & \operatorname{sh} z &= \frac{e^z - e^{-z}}{2} \\
 1 &= \cos^2 z + \sin^2 z & 1 &= \operatorname{ch}^2 z - \operatorname{sh}^2 z
 \end{aligned}$$

2. Основная теорема алгебры. В поле комплексных чисел любой многочлен приводим, т. е. уравнение вида $a_0 z^n + a_1 z^{n-1} + \dots + a_{n-1} z + a_n = 0$, где все $a_i \in \mathbb{C}$ и $a_0 \neq 0$, имеет ровно n комплексных корней z_1, \dots, z_n (некоторые из них могут совпадать, в таком случае они называются кратными) и, следовательно, многочлен можно записать в виде:

Мультимно-
жество
корней.

$$a_0 z^n + a_1 z^{n-1} + \dots + a_{n-1} z + a_n = a_0 (z - z_1) \dots (z - z_n) = a_0 (z - z_i)^{k_i} \dots (z - z_j)^{k_j},$$

где $n = k_i + \dots + k_j$ — кратности соответствующих корней.

Интересно, что этот факт имеет много доказательств, но среди них нет ни одного чисто алгебраического — всякий раз оно существенно опирается на какое-то неалгебраическое свойство \mathbb{C} , прежде всего, — на непрерывность (полноту). Возможно, это унаследованное свойство того уникального положения поля \mathbb{R} среди всех алгебраических полей, о котором мы уже упоминали выше. Красивое тополого–геометрическое доказательство основной теоремы алгебры можно найти в [62].

3. Условия Коши–Римана. Для функции $f : \mathbb{C} \rightarrow \mathbb{C}$, отображающей точку $z = x + iy$ в точку $w = u + iv$, можно рассмотреть естественно возникающие действительные функции двух переменных: $u = u(x, y), v = v(x, y)$. Однако возможность дифференцировать по комплексному аргументу являются более сильной «опцией», чем раздельное дифференцирование функций u и v по своим аргументам. Имеет место критерий

Теорема 2.18 (Условия Коши–Римана/Эйлера–Даламбера). *Функция $w = f(z)$, где $w = u + iv, z = x + iy$, дифференцируема в точке z тогда и только тогда, когда выполнены условия:*

$$u'_x = v'_y, \quad u'_y = -v'_x \quad (f'_x + i f'_y = 0)$$

в точке (x, y) .

Дифференцируемая в открытой области²⁴ D функция $f(z)$ называется **регулярной**²⁵ в D . В отличие от \mathbb{R} , где функция может быть дифференцируема только до n -го порядка, в \mathbb{C} дифференцируемость в области автоматически означает существование всех производных $f^{(n)}(z)$ в этой области, а значит, и разложение функции в ряд Тейлора в окрестности точки $z \in D$. В области, где первая производная регулярной функции не обращается в 0, а сама функция однолистна (т. е. действует инъективно), она осуществляет **конформное отображение**, т. е. такое отображение, которое сохраняет форму фигуры локально (в каждой точке сохраняет углы между кривыми при их преобразовании).

Тригонометрические функции, как и экспонента, являются регулярными на всей комплексной плоскости.

Из условий Коши–Римана следует, что обе компоненты u и v функции f являются гармоническими в смысле вещественного анализа, т. е. для них выполняется уравнение Лапласа:

$$\Delta u = \left(\frac{\partial^2}{\partial^2 x} + \frac{\partial^2}{\partial^2 y} \right) u = 0, \quad \Delta v = \left(\frac{\partial^2}{\partial^2 x} + \frac{\partial^2}{\partial^2 y} \right) v = 0, \quad (2.10)$$

где Δ — дифференциальный **оператор Лапласа** или **лапласиан**.

Если нам известна гармоническая функция u , то можно найти гармоническую функцию v такую, что $u + iv$ будет регулярной функцией комплексного переменного. Нахождение такой v называется восстановлением регулярной функции.

Уравнения вида (2.10) играют важную роль в механике и математической физике. В совокупности с граничными условиями, означающими что функция u совпадает с некоторой заданной функцией на границе области, в которой выполняется уравнение Лапласа, мы имеем так называемую *задачу Дирихле*.

Задача Дирихле — это одна из задач решения дифф.уравнения второй степени с граничными условиями, она имеет приложения для решения уравнения теплопроводности, уравнения Стокса (для поля скоростей), уравнений Максвелла (для электромагнитного поля) и т.д. Таким образом, мы видим изящный переход от чисто физических задач к задачам в комплексных числах.

²⁴Открытой областью называется всякое подмножество D такое, что оно вместе с каждой своей точкой z_0 содержит и некоторую ее круговую окрестность, т. е. множество вида $\{|z - z_0| < r\}$ при достаточно малом $r > 0$.

²⁵Наряду с термином «регулярная функция» используются термины «аналитическая функция» и «голоморфная функция». Эти определения имеют разное происхождение, но при этом эквивалентны в \mathbb{C} . Аналитичность означает возможность разложить функцию в ряд Тейлора, а голоморфность — локальное сохранение формы фигуры.

4. Формула Коши. Не вдаваясь в подробности определения интеграла в комплексном анализе, дадим следующую формулу:

$$f(z) = \frac{1}{2\pi i} \oint_{\Gamma} \frac{f(\zeta)}{\zeta - z} d\zeta,$$

если функция f регулярна в открытой области D , замкнутый контур Γ целиком лежит в этой области, а точка z находится внутри этого контура, причем контур Γ может быть непрерывно стянут в точку z внутри области D .²⁶

Особо подчеркнем, что для интегрирования на плоскости (не обязательно комплексной) важно учитывать направление обхода контура. По умолчанию считается, что контур обходится в положительном направлении, т. е. так, чтобы ограниченная им область оставалась слева, что соответствует вращению вектора e^{it} с возрастанием t .

На первый взгляд кажется, что условий для формулы многовато, но чудо комплексного анализа и состоит в том, что как правило мы здесь работаем именно с такими «хорошими» функциями и областями. Чудом является и сама формула Коши, которая означает, что регулярную функцию можно восстановить в любой точке области, интегрируя по одному известному контуру. Например, таким контуром может быть граница самой этой области, если f регулярна в D и непрерывна на замкнутой области \overline{D} , т. е. сумме D и ее границы. При этих же условиях имеет место формула для любой производной внутри D :

$$f^{(n)}(z) = \frac{n!}{2\pi i} \oint_{\Gamma} \frac{f(\zeta)}{(\zeta - z)^{n+1}} d\zeta.$$

Иначе говоря, интегрируя по границе D , мы получаем все производные f в области D , а значит, и ее аналитическое разложение в произвольной точке области D . Важно, чтобы D была односвязной.

Интересно также, что для регулярной в односвязной области D функции справедлива теорема Коши:

$$\oint_{\Gamma} f(z) dz = 0$$

для любого замкнутого контура Γ , лежащего внутри области D . Если на контуре Γ выделить две различные точки a и b и представить сам контур как два пути из a в b , то в условиях теоремы Коши мы получаем, что интеграл регулярной функции вдоль пути не зависит от выбора пути между точками a и b , если только эти пути можно непрерывно перевести один в другой (использовать односвязность области D).

²⁶ Область, в которой любой замкнутый контур может быть непрерывно стянут в точку, называется односвязной.

...и представлению математиков о положительном вращении.

5. Принцип максимума модуля. Если функция $f(z)$ регулярна в открытой области D и непрерывна на \overline{D} , то справедлив принцип максимума модуля: либо $|f(z)| = \text{const}$, либо максимальные значения модуля $|f(z)|$ достигаются только на границе области D .

Заметим, что если, кроме того, $|f(z)| \neq 0$ в области D , то справедлив принцип минимума модуля — он достигается исключительно на границе области.

6. Теорема Лиувилля. Пусть на всей комплексной плоскости функция $f(z)$ является регулярной²⁷ и ее модуль равномерно ограничен. Тогда $f(z)$ является тождественной константой.

Из этой теоремы следует, например, что $|\sin z|$ не ограничен на \mathbb{C} , т. е. для некоторых z можно получить $|\sin z| = 2$.

Упражнение 2.40. Найдите такие примеры z .

7. Теорема единственности. Пусть функции f и g регулярны в открытой области D , причем существует последовательность $\{z_n\} \subset D$, сходящаяся к точке $a \in D$, на которой эти функции совпадают: $f(z_n) = g(z_n)$. Тогда $f(z) = g(z)$ на всей области D .

У теоремы единственности есть ряд полезных следствий. Например,

- 1° Отличная от тождественного нуля регулярная в открытой области D функция имеет лишь конечное число нулей в любой замкнутой подобласти области D ;
- 2° Регулярная ненулевая функция имеет бесконечное число нулей лишь в открытой или неограниченной области;
- 3° Целая ненулевая функция может иметь лишь счетное число нулей, предельной точкой которых является бесконечно удаленная точка;
- 4° Если две регулярные в открытой области D функции совпадают на некоторой кривой L , лежащей в области D , то они совпадают во всей области D .

Кривая L может быть отрезком действительной оси в \mathbb{C} , так что если на этом отрезке задана обычная действительная функция, раскладываемая в ряд Тейлора, то она может быть единственным способом продолжена на некоторую область D , содержащую отрезок L . Таким способом легко продлеваются в комплексную область все элементарные функции.

8. Ряд Лорана. Мы знаем, что аналитическая (регулярная) функция имеет представление в виде ряда Тейлора:

$$f(z) = \sum_{n=0}^{\infty} a_n \frac{(z - z_0)^n}{n!},$$

²⁷Регулярная на всем \mathbb{C} функция еще называется целой.

причем этот ряд сходится равномерно в своем круге сходимости, радиус которого зависит от последовательности коэффициентов $\{a_n\}$. Но существует возможность рассмотреть более общий ряд

$$\sum_{n=-\infty}^{\infty} c_n(z - z_0)^n$$

с отрицательными степенями, для которого существует кольцо сходимости, внешний радиус этого кольца определяется коэффициентами с положительными номерами, а внутренний — с отрицательными. Иногда это кольцо может совпадать со всей плоскостью \mathbb{C} , иногда вообще оказаться пустым множеством. Чаще всего получается кольцо вида $0 < |z - z_0| < R$, т. е. круг с выколотым центром z_0 . В этом случае говорят, что точка z_0 является **изолированной особой точкой**²⁸, а функция $f(z)$ раскладывается в **ряд Лорана**.

Это не то кольцо, о котором все подумали?

Существует классификация особых точек:

- P1 устранимая особая точка — это точка, в которой $f(z)$ можно доопределить по непрерывности конечным значением;
- P2 полюс порядка m — это точка, в которой $f(z)$ раскладывается в ряд Лорана с $c_{-m} \neq 0$, $c_{-m-1} = c_{-m-2} = \dots = 0$, $m > 0$;
- P3 существенно особая точка — это точка, в которой $f(z)$ представляется рядом Лорана с бесконечным набором отличных от нуля коэффициентов с отрицательными номерами.

Если z_0 — полюс порядка m для $f(z)$, то $f(z) = g(z)/(z - z_0)^m$, где $g(z)$ — регулярная в точке z_0 .

Если z_0 — существенно особая точка $f(z)$, то в любой сколь угодно малой окрестности z_0 найдется значение $f(z)$, сколь угодно близкое к произвольному наперед заданному комплексному числу B .

*Теорема
Сохоцкого–
Вейерштрасса.*

Если мы к плоскости \mathbb{C} добавляем бесконечно удаленную точку (рассматриваем $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$), то к ней также применимо понятие полюса, только в определении нужно поменять местами хвосты ряда Лорана. Действительно, мы можем от $f(z)$ перейти к $g(w) = f(1/w)$ и уже изучать свойства точки 0. Ноль будет полюсом степени m функции g тогда и только тогда, когда ∞ будет полюсом степени m функции f .

Расширенная комплексная плоскость $\overline{\mathbb{C}}$ оказывается весьма удобным и плодотворным инструментом. Например, в такой плоскости удобно изучать дробно-линейные преобразования. Для ознакомления с этой темой рекомендуем обратиться к лекциям [80].

²⁸Изолированная — потому что в ее окрестности нет других особых точек, а особая — потому что значение функции в этой точке не совпадает с суммой ряда Тейлора в ней же.

Всякие интересные штуки

1. Формула Муавра. Это формула, позволяющая легко находить тригонометрические выражения для двойного, тройного и т.д. угла.

$$(\cos(t) + i \sin(t))^n = \cos(nt) + i \sin(nt). \quad (2.11)$$

Например, полагая $n = 2$, имеем:

$$(\cos(t) + i \sin(t))^2 = \cos^2(t) - \sin^2(t) + 2i \cos(t) \sin(t) = \cos(2t) + i \sin(2t),$$

откуда следуют известные школьные тождества:

$$\cos(2t) = \cos^2(t) - \sin^2(t), \quad \sin(2t) = 2 \sin(t) \cos(t).$$

2. Корни из 1. Приравнивая правую часть тождества (2.11) к 1, получаем формулу всех корней степени n из 1.

Действительно, нужно, чтобы $nt = 2\pi k$ ($k \in \mathbb{Z}$), откуда $t = 2\pi k/n$ и

$$\sqrt[n]{1} = z_k = \cos(2\pi k/n) + i \sin(2\pi k/n) = e^{i2\pi k/n}. \quad (2.12)$$

Перебирая все целые k , мы находим ровно n различных корней при $k = 0, 1, \dots, n-1$, которые расположены на единичной окружности с центром в 0, образуют правильный n -угольник, одна из вершин которого совпадает с числом 1.

Упражнение 2.41.
Докажите, что $\{z_k\}$ образуют группу.

Формула (2.12) полностью исчерпывает все корни уравнения $z^n = 1$ в поле \mathbb{C} . Кроме того, эти корни образуют циклическую группу по умножению с порождающим элементом z_1 , т. к. $z_k = z_1^k$.

3. Двойное отношение. Для чисел $z_1, z_2, z_3, z_4 \in \mathbb{C}$ **двойным отношением** называется выражение:

$$[z_1, z_2, z_3, z_4] = \frac{(z_3 - z_1)(z_4 - z_2)}{(z_3 - z_2)(z_4 - z_1)}.$$

Двойное отношение не меняется при сдвигах на константу $z \mapsto z + a$, а всевозможные перестановки (всего их ровно 24) чисел внутри двойного отношения приводят к шести возможным значениям двойного отношения:²⁹

$$\lambda = [z_1, z_2, z_3, z_4], \quad \frac{1}{\lambda}, \quad 1 - \lambda, \quad \frac{1}{1 - \lambda}, \quad \frac{\lambda}{\lambda - 1}, \quad \frac{\lambda - 1}{\lambda}. \quad (2.13)$$

²⁹Это объясняется тем, что группа Клейна V_4 , являющаяся нормальной подгруппой S_4 , действует на двойном отношении тривиально, т. е. сохраняет его. Поэтому существует 6 классов перестановок в S_4 , в каждом из которых находится по 4 перестановки, не различающие двойное отношение. При этом фактор-группа S_4/V_4 , изоморфная S_3 , обеспечивает 6 нетривиальных перестановок двойного отношения.

Упражнение 2.42. | Проверьте, что все 24 перестановки двойного отношения дают именно такие числа.

Для геометрии комплексных чисел (и проективной геометрии, см. раздел 3.6.5) важна следующая

Теорема 2.19. Если $z_1 \neq z_4$ и $z_2 \neq z_3$, то числа z_1, z_2, z_3, z_4 лежат на одной прямой или окружности тогда и только тогда, когда их двойное отношение является действительным числом.

Заметим, что в комплексном анализе прямые часто также относят к окружностям, только проходящим через бесконечно удаленную точку (окружность в $\overline{\mathbb{C}}$), это позволяет упростить формулировки некоторых теорем, в частности, приведенной выше теоремы.

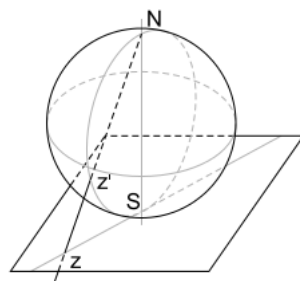
4. Формула вычетов. Если z_0 — изолированная особая точка функции $f(z)$, то коэффициент c_{-1} ряда Лорана этой функции в окрестности z_0 называется вычетом $f(z)$ в точке z_0 и обозначается $\text{Res}(f, z_0)$. Согласно формуле Коши имеем:

Снова не налоговый вычет!

$$\text{Res}(f, z_0) = \frac{1}{2\pi i} \oint_C f(\zeta) d\zeta,$$

где контур C при интегрировании обходится в положительном направлении и не содержит в охваченной области никаких других особых точек.

Если же точка $z_0 = \infty$, т.е. является бесконечно удаленной точкой, то вычет в ней вычисляется аналогично, только ориентация контура меняется на противоположную (поскольку обход ∞ в положительном направлении — это обход нуля в отрицательном направлении, в чем легко убедиться, если представить $\overline{\mathbb{C}}$ в виде глобуса, см. рис. 2.9, где северный полюс соответствует ∞ , а южный — нулю), что соответствует смене знака перед интегралом:



$$\text{Res}(f, \infty) = -\frac{1}{2\pi i} \oint_C f(\zeta) d\zeta,$$

Рис. 2.9: Риманова поверхность $\mathbb{C} \cup \{\infty\}$.

где контур C обходится в положительном направлении, причем функция f вне области, охваченной данным контуром, регулярна и не имеет особых точек, кроме, быть может, ∞ .

Основная теорема теории вычетов дает формулу:

$$\oint_{\Gamma} f(\zeta) d\zeta = \sum_{k=1}^N \text{Res}(f, z_k) = 0,$$

где контур Γ ограничивает такую открытую область D , что функция $f(z)$ регулярна на \overline{D} , за исключением конечного числа изолированных особых точек $z_1, \dots, z_N \in D$, а вне области \overline{D} не имеет особых точек, кроме бесконечно удаленной.

В силу определения вычета в бесконечно удаленной точке мы имеем эквивалентную формулировку:

$$\sum_{k=0}^N \text{Res}(f, z_k) = 0,$$

если z_0, \dots, z_N — полный перечень изолированных особых точек функции $f(z)$, включая точку $z_0 = \infty$.

С помощью этих формул с вычетами удастся достаточно просто получать, например, вещественные интегралы от функций вида $R(\cos \theta, \sin \theta)$, где R — рациональная функция (т.е. отношение полиномов от двух переменных), интегралы по всей действительной оси от функции $f(x)$, которая может быть аналитически продолжена в верхнюю комплексную полуплоскость, а также специальные интегралы вида

$$\int_{-\infty}^{\infty} e^{itx} f(x) dx,$$

имеющие приложения, например, в теории вероятностей, где изучение случайных величин сводится к их комплексно-значным характеристическим функциям.

Здесь мы, пожалуй, находим очередной методологический архетип математики — *выход за рамки изучаемой области в более широкое пространство для того, чтобы в нем сформулировать и решить более простыми методами исходную сложную задачу*. Мы уже видели что-то похожее, когда имели дело с теоремой Гудстейна. Для ее доказательства пришлось выйти за пределы натурального ряда и прибегнуть к помощи бесконечных ординалов. Правда, у этого примера есть и существенное отличие — принципиальная недоказуемость теоремы Гудстейна в своей области формулировки. В случае с переходом от действительных интегралов к комплексным для упрощения задачи мы не утверждаем, что эти действительные интегралы невозможно посчитать, не прибегая к помощи комплексных чисел. Обозначим данный архетип словосочетанием **архетип трансцендентного восприятия**.

Выйдем теперь ненадолго за рамки математики.

5. Законы Кеплера. Покажем применение комплексных чисел при доказательстве закона эллиптичности орбит тел в гравитационном поле.

Ниже мы рассматриваем для определенности только эллипсы, большая полуось которых лежит на вещественной оси в \mathbb{C} .

Теорема 2.20. Если $w \in \mathbb{C}$ пробегает эллипс с центром в 0, то w^2 пробегает эллипс с фокусом в 0. И обратно, всякий эллипс с фокусом в 0 есть квадрат эллипса с центром в 0.

Для доказательства используется **функция Жуковского**: $w = z + 1/z$. При таком преобразовании плоскости \mathbb{C} окружность с центром в 0 и радиусом $r > 1$, которую пробегает z , переходит в эллипс с центром в 0 и полуосями $a = r + r^{-1}$, $b = r - r^{-1}$. При этом, очевидно, что выбором подходящего r можно получить эллипс с любым отношением полуосей a/b .

Кроме того, $w^2 = z^2 + \frac{1}{z^2} + 2$, т.е. w^2 является преобразованием Жуковского для z^2 со сдвигом на 2 вдоль вещественной оси и полуосями $A = r^2 + \frac{1}{r^2}$, $B = r^2 - \frac{1}{r^2}$. При этом квадрат половины расстояния между фокусами равен $A^2 - B^2 = 4$, т.е. фокусы удалены от центра эллипса на 2. Это означает, что w^2 пробегает эллипс с фокусом в 0.

Таким образом, представляя произвольный эллипс как эллипс Жуковского с центром в 0, получаем, что его квадрат есть эллипс Жуковского с фокусом в 0. Обратное очевидно из того, что и для эллипса с полуосями $A = r^2 + \frac{1}{r^2}$, $B = r^2 - \frac{1}{r^2}$ можно задать любое нужное отношение полуосей, т.е. w^2 может описывать любой заданный эллипс (с точностью до масштаба и поворота).

Теорема 2.21. Точка w описывает эллипс по закону Гука тогда и только тогда, когда w^2 описывает эллипс по закону гравитации.

Доказательство. Пусть $w = w(t)$ подчиняется закону Гука, т.е. $w'' = -w$.³⁰

Нетрудно видеть, что решением этого уравнения является $w = ae^{it} + be^{-it}$, и величина $|w'|^2 + |w|^2 = 2(a^2 + b^2)$ является константой. Поскольку для доказательства эллиптичности орбиты нам не важно знать, как прохождение по ней тела зависит от времени, мы вольны выбрать свое время для траектории, описываемой величиной w^2 . Положим $Z = Z(\tau) = w(t)^2$,³¹ причем свяжем переменные времени t и τ дифференциальным равенством:

$$\frac{d\tau}{dt} = |w|^2 = w\bar{w}.$$

Далее,

$$\begin{aligned} Z' &= \frac{d}{dt} w^2 \frac{dt}{d\tau} = \frac{2ww'}{w\bar{w}} = 2\frac{w'}{\bar{w}}, \\ Z'' &= \frac{d}{dt} 2\frac{w'}{\bar{w}} \frac{dt}{d\tau} = -\frac{2}{w\bar{w}^3}(|w'|^2 + |w|^2) = -4(a^2 + b^2) \frac{Z}{|Z|^3}. \end{aligned}$$

На самом деле это свойство закона сохранения энергии вдоль орбиты

³⁰Здесь мы используем штрих как обозначение производной по аргументу функции, указанному в ее определении: для w это t , для Z это τ .

³¹На самом деле это эквивалентно равенству $\tau' = |Z|^2/|w|^2$, которое является следствием законов площадей: $|w|^2 d\phi/dt = \text{const}$, $|Z|^2 d\phi/d\tau = \text{const}$, где ϕ — аргумент соответствующей комплексной переменной.

Как видим, в итоге получается закон гравитации с произвольной константой, зависящей от выбора параметра $a^2 + b^2$ в решении уравнения Гука.

Произвольность выбора константы говорит о том, что для любого гравитирующего заряда можно подобрать соответствующее (двойственное) единственное решение уравнения Гука и, тем самым, показать эллиптичность орбиты заряда в гравитационном поле, что доказывает обратное утверждение теоремы. \square

Более подробно см. у Арнольда [1] «Добавление 1».

6. Аэродинамика. Отойдём немного от того условия, что большая полуось эллипса лежит на действительной оси. Функция Жуковского $w = \frac{1}{2}(z + \frac{1}{z})$ отображает окружность, проходящую через точки 1 и -1 со смещенным вверх центром в фигуру, которая моделирует срез авиационного крыла. При этом горизонтальный аэродинамический поток, обходящий окружность в плоскости z , превращается в аэродинамический поток, обходящий крыло в плоскости w (в силу конформности функции Жуковского). Это позволяет свести сложные вычисления для крыла к более простым для окружности (цилиндра).

Функция $w = \frac{1}{2}(z + \frac{1}{z})$ названа в честь основателя аэродинамики русского ученого Николая Жуковского.

7. Квантовая физика. Наконец, нельзя обойти вниманием использование комплексных чисел в квантовой физике, где псифункция, являясь комплексно-значной, удачно описывает правила взаимодействия на квантовом уровне материи. Почему так происходит, видимо, еще предстоит ответить будущим поколениям матфизиков.

В мир квантовой теории мы заглянем позже, в разделе 5.4.2, когда будем вести речь о классических и квантовых вероятностях.

*и не только
функция, а
целый город
в
Подмосковье
и ряд инсти-
тутов!*

Дальнейшие обобщения чисел

В предыдущей главе мы в основном строили числа с согласованным линейным порядком:

- Вполне упорядоченные:
 - Натуральные числа
 - Ординалы
 - Кардиналы
- Линейно упорядоченные:
 - Целые числа
 - Рациональные числа
 - Действительные числа
 - Гипердействительные числа
 - Сюрреальные числа

Только с построением поля комплексных чисел мы позволили себе нарушить общую канву главы 2 и прибавить одно измерение — выйти в плоскость. Ниже мы увидим, что поле \mathbb{C} предоставляет нам не только инструмент алгебраического пополнения \mathbb{R} и возможность алгебраизации евклидовой плоскости, но и дает возможность расширить решетку целых чисел так, что мы увидим новые нетривиальные обобщения теории чисел и даже коснемся с их помощью доказательства Великой теоремы Ферма.

Кроме того, мы рассмотрим ряд обобщений числовых структур и (алгебраических) методов их «изготовления». Для этого мы снова нырнем в алгебру, изучим некоторые общие конструкции, после чего займемся строительством нелинеаризованных числовых структур, по пути отыскивая новые архетипы.

3.1 Окно в общую алгебру

Построение и изучение общей алгебры напоминает построение теории множеств, которым мы занимались в первой главе. Дело в том, что общая алгебра

вполне может сама претендовать на роль «теории всего» в математике, диктуя исследователю язык и методологию исследования. Тем не менее, в самом ее основании все равно угадываются черты теории множеств по той простой причине, что (архетипичные) логические конструкции, лежащие в основе математического знания, не зависят от выбора модели и методик представления изучаемых объектов, а именно логические конструкции и есть основа теории множеств (что мы видим на примере связи булевой алгебры с логикой).

Итак, рассмотрим некоторое базовое множество M с заданными на нем функциями, отношениями и системами подмножеств. Множество этих функций обозначим F , множество отношений R , множество систем подмножеств P . Тогда четверка (M, F, R, P) называется **математической структурой**. Обычно предполагается, что систем подмножеств задано некоторое конечное множество, а вот отношений и функций может быть сколь угодно много (это зависит, конечно, от M).

Ирония
судьбы:
 n -арные
отношения
задаются
функциями,
а функции —
бинарными
отношениями

Учитывая, что функции — это разновидность отношений, мы можем просто считать, что на множестве M задано несколько отношений, причем отношения могут быть n -арными, т. е. подмножествами M^n , а не квадрата $M \times M$, а n -арные операции будут $(n + 1)$ -арными отношениями. Систему подмножеств при желании тоже можно свести к отношениям, а именно: на $\mathcal{P}(M)$ задать характеристическую функцию со значениями в $\{0, 1\}$, тогда прообраз 1 выделяет в $\mathcal{P}(M)$ некоторую систему подмножеств, а сама характеристическая функция будет отношением на произведении $\mathcal{P}(M) \times \{0, 1\}$.

Функции, которые используют в качестве аргументов не только элементы M , а еще, например, какие-то числа из множества K , также представляются в виде подмножеств в прямых произведениях вида $M^n \times K^s$, и тем самым, любую структуру можно считать набором отношений на наборе базовых множеств $(M, K, \mathcal{P}(M))$ и, возможно, еще каких-то производных от этих).

Правда, такое понятие структуры несколько не богаче априори по своему содержанию, чем какой-нибудь достаточно большой универсум или, в конце концов, сама теория множеств. Так что разделение структурных элементов на функции, отношения и системы подмножеств — вполне компромиссный вариант. К тому же, это разделение соответствует классификации Бурбаки для порождающих структур (функции отвечают за алгебру, отношения — за порядки, системы подмножеств — за топологические конструкции).

3.1.1 Архетип переноса свойств на базовое множество

Математическая структура (M, F, R, P) с кучей заданных в ней отношений и функций часто неявно отождествляется с базовым множеством, на которое негласно переносятся все ее свойства. Так, при определении изоморфизма отношений мы уже упоминали, что вместо $(A, R_A) \sim (B, R_B)$ пишут $A \sim B$,

хотя по сути своей изоморфизм опирается на отношения, а не на базовые множества.

Аналогичное употребление понятия изоморфизма можно встретить и применительно к любым математическим структурам, в частности, изоморфизм групп G и G' означает на самом деле изоморфизм их операций: $g_1 \circ g_2 \leftrightarrow f(g_1) \circ' f(g_2)$, где f — биекция между G и G' .

Перенос свойств касается не только изоморфизмов. Например, свойство плотности порядка линейно упорядоченного множества $(A, <)$ может быть перенесено на базовое множество A . Так, принято говорить, что \mathbb{Q} плотно, имея ввиду стандартный порядок на множестве рациональных чисел.

Аналогично, взятие мощности вполне упорядоченного множества $|(A, <)|$ вовсе не означает, что мы должны получить ответе 2, поскольку, формально говоря, $(A, <) = \{\{A\}, \{A, < \}\}$, вместо этого ответом будет кардинал τ , биективно сопоставленный с базовым множеством A .

Тем самым, напрашивается очередной архетип математики — *перенос свойств формального объекта на его базовое множество*, а иногда и на некоторое понятие (например, на множество рациональных чисел) вне контекста его конкретного воплощения в иерархии множеств теории ZF. Назовем это **архетипом базового множества**.

В этом смысле было бы удобно пользоваться более «продвинутым» формальным языком (похожим, например, на `phr`), который применяет высказывание к объекту теории, автоматически заменяя объект на его содержательную часть — базовое множество, отношение, систему подмножеств, операцию или функцию. Это легко понимает человек-математик, но это крайне затруднительно объяснить главному герою нашей книги — компьютеру.

3.1.2 Алгебраические структуры

Под алгеброй в зависимости от контекста понимают и науку о свойствах чисел и порождаемых ими симметрий, и некоторую специальную математическую структуру. Мы можем условиться в рамках данной книги, что наука Алгебра занимается изучением алгебраических структур, к определению которых мы сейчас перейдем, а термин алгебра будет означать некоторый специальный вид такой структуры.

Как мы уже заметили, Алгебра изучает свойства чисел и симметрий, причем под числами мы понимаем не только натуральные и действительные числа, а вообще все возможные системы с операциями, похожими на привычные нам сложение и умножение. Симметрии, которые возникают при воздействии на числа различными операциями (функциями) являются неизбежным следствием устройства чисел и потому неотделимы от Алгебры.

Таким образом, мы выходим на общее определение **алгебраической структуры** как некоего базового множества с операциями и отношениями

Арабское
al-jabr

(что, как мы выше видели, есть практически одно и то же), т. е. алгебраическая структура — это разновидность математической структуры, где исключены из рассмотрения системы подмножеств. Исключены — в том смысле, что в рамках изучения предмета Алгебры мы не опираемся на какие-либо свойства систем подмножеств (прежде всего топологию), но это не значит, что мы не можем использовать системы подмножеств при построении алгебраических структур.

Возможно, алгебраические геометры с нами не согласятся.

Сигнатурой алгебраической структуры называется символьный список функций и отношений с указанием количества аргументов этих функций и отношений. Например, сигнатурой \mathbb{Z} будет $(_ + _ ; _ \cdot _ | _ < _)$, где вертикальной чертой отделены функции от отношений, а подчеркивания указывают на расположение аргументов и их количество.

На сигнатуру можно «натянуть» счетный набор формул точно так же, как мы строили формулы теории множеств, пользуясь рекурсивными определениями грамматики. При этом функциональные символы будут играть роль функциональных термов и поставлять нам все новые и новые многоэтажные конструкции термов, а отношения в купе с логическими связками и подстановками на аргументные места термов будут поставлять все новые и новые термы. Таким образом, мы построим формальный язык, присущий конкретной сигнатуре. В этом языке мы можем задать аксиомы, т. е. выбрать формулы, которые будем считать истинными (например, аксиомы линейного порядка), после чего можно будет выводить из них теоремы по правилам вывода. В итоге мы получим то, что называется **теорией**, причем эта теория будет построена исключительно в заданной сигнатуре и к теории множеств будет иметь весьма слабое отношение. Алгебраическая структура (как объект теории множеств) с данной сигнатурой будет моделью этой теории, если в ней выполняются аксиомы теории. Например, ω не может быть моделью \mathbb{Z} , несмотря на идентичность сигнатур, поскольку аксиомы кольца не выполняются в ω .

А теперь вопрос: что если рассмотреть алгебраическую структуру с сигнатурой $(| _ \varepsilon _ ; _ = _)$, состоящей только из двух отношений, одно из которых есть равенство, и построить теорию, в точности повторяющую теорию с аксиоматикой \mathbf{ZF} , где вместо принадлежности используется символ ε ? На каком базовом множестве с каким отношением ε нам удастся смоделировать не что иное, как саму теорию \mathbf{ZF} ? Этот вопрос мы оставим до раздела 4.2 (конкретно — метатеорема 4.12).

Не путать с алгеброй над кольцом и алгеброй множеств!

Алгебраическая структура с пустым набором отношений (т. е. в ней есть только операции) называется **алгеброй**. Строго говоря, нужно иметь в виду, что алгебра может включать не только функции из M^n в M , но, как говорилось выше, и функции из $M^n \times K^s$ в M , т. е. к операции над элементами M примешиваются некоторые вспомо-

гательные (числовые) параметры, операции с которыми подчиняются ряду аксиом. В этом случае говорят об алгебре над K (полем, кольцом и т.п.).

То, что мы так легко переходим от операций на M к операциям, включающим некие дополнительные параметры из K , легко объясняется следующими соображениями. Всякую функцию $f : M^n \times K^s \rightarrow M$ «без потери качества» можно заменить на множество функций $\{f_k : M^n \rightarrow M \mid k \in K^s\}$, полагая

$$f(m_1, \dots, m_n, k_1, \dots, k_s) = f_k(m_1, \dots, m_n),$$

где $k = (k_1, \dots, k_s)$. Тем самым, мы просто увеличим множество функций на M^n , параметризовав их наборами чисел из K . Этим приемом часто пользуются, например, в анализе и аналитической геометрии.

Великое разнообразие алгебраических структур диктуется выбором системы аксиом (правил), которым должны подчиняться операции и отношения. В разделе 2.4.1 мы уже ввели ряд определений алгебраических структур, такие как: *группоид*, *полугруппа*, *моноид*, *группа*, *кольцо*, *поле*. Например, алгебра с одной бинарной операцией называется **группоидом**. К алгебрам относятся и все остальные перечисленные только что структуры. Однако, если мы вспомним такой термин, как упорядоченное поле, то мы здесь увидим как операции, так и отношение, поэтому такая структура не будет алгеброй, но будет алгебраической структурой.

В Алгебре при определении структур проще всего использовать созданный Д. Гильбертом аксиоматический подход. Сами аксиомы, собственно говоря, и являются определениями тех операций и отношений, которые использованы в данной конкретной алгебраической структуре. При этом бонусом является полное абстрагирование от теоретико-множественной базы, т. е. от базового множества M . Его природа нам совершенно безразлична, а все свойства структуры выводятся исключительно из определяющих ее аксиом. Как уже отмечалось (см. начало раздела 2.4), это тонкое место, где Алгебра и теория множеств как бы «мешают» друг другу. С одной стороны, аксиоматический подход избавляет нас от привязки к конкретным множествам, с другой — мы по-прежнему хотим пользоваться теоретико-множественным языком, давая те или иные определения или теоремы. Возможно, эту небольшую проблему можно разрешить модификацией языка математики.



Давид
Гильберт

3.1.3 Алгебра множеств

Рассмотрим один простой пример как раз на стыке алгебры и теории множеств.

Пусть дано произвольное множество X и его булеан $\mathcal{P}(X)$, в котором выделим подмножество A со следующими свойствами:

т. е. не что иное как систему подмножеств!

A1 $\emptyset \in A$;

A2 если $a \in A$, то $X \setminus a \in A$;

A3 если $a, b \in A$, то $a \cup b \in A$.

Такое множество A называется **алгеброй множеств** на X . Минимальная алгебра множеств на X : $\{\emptyset, X\}$. Максимальная алгебра множеств на X — весь булеан $\mathcal{P}(X)$.

Казалось бы — при чем тут алгебра, если речь идет о системе подмножеств множества X ? На самом деле в качестве основного множества алгебры множеств мы берем именно A , после чего рассматриваем на нем теоретико-множественные операции:

$$a + b = (a \setminus b) \cup (b \setminus a) \text{ (симметрическая разность)}$$

$$ab = a \cap b$$

Упражнение 3.1. Проверьте эти свойства!

Эти операции ассоциативны, коммутативны, имеют свои нейтральные элементы (для умножения таковым является само X , для суммы — \emptyset), т. е. по каждой из этих операций A является коммутативным моноидом, кроме того, $a + a = 0$, т. е. каждый элемент сам себе противоположен, а также имеют место дистрибутивные законы, так что A с указанными операциями является коммутативным кольцом с единицей.

Возможно, было бы интересно поизучать кольцо *мультимножеств* над числовой структурой, отвечающей за кратность элементов. В качестве такой числовой структуры по определению мультимножеств можно выбирать любой ординал. Однако, даже самый хороший из ординалов — ω — не обладает свойствами кольца в силу необратимости сложения натуральных чисел, а отрицательные целые числа в качестве кратностей элементов мы не используем.

Упражнение 3.2. Постройте алгебру мультимножеств.

Таким образом, задать алгебру мультимножеств над кратностями некоторым естественным способом не получается, а выстраивать какие-то сложные арифметики на \mathbb{N} без особой на то причины — занятие сомнительное. Более перспективным полем деятельности нам кажется изучение свойств того, что есть, т. е. кольца мультимножеств с коэффициентами — натуральными числами. Возможно, это выведет исследователя на какие-то слабо изученные алгебраические структуры или поможет иначе взглянуть на хорошо известные.

Усилением понятия алгебры множеств является понятие **σ -алгебры**. Это — тоже алгебра множеств, но с требованием счетной замкнутости по объединению, т. е. если $\{a_n \mid n \in \omega\} \subseteq A$, то объединение $\bigcup_{n=0}^{\infty} a_n \in A$. Сигма-

алгебры играют важную роль при изучении пространств с мерами, в частности, в теории вероятностей.

Минимальная сигма-алгебра, содержащая семейство множеств \mathcal{B} , называется *порожденной* этим семейством и обозначается $\sigma(\mathcal{B})$.

Комментарий 3.

Нужно отметить, что в математике вообще замкнутость (или иное свойство) по дизъюнкции (объединению) принято называть «сигма»-замкнутостью (иным свойством), а замкнутость (иное свойство) по конъюнкции (пересечению) — «дельта»- или «пи»-замкнутостью (иным свойством). Так, мы можем вспомнить про \sum -, Δ - и \prod -формулы, речь о которых пойдет в разделе 4.3.3.

\sum -формулы получаются присоединением кванторов существования (аналог дизъюнкции), а \prod -формулы — присоединением квантора всеобщности (конъюнкция). По всей видимости, такие обозначения связаны с арифметическим смыслом: \sum — сумма, \prod — произведение, которые «архетипично» соответствуют сумме и пересечению множеств.

На этом фоне использование слова «дельта» для замкнутых по пересечению систем множеств в книге [81] выглядит не совсем корректно.

Поскольку σ -алгебра замкнута еще и по дополнению множеств, для нее сигма-замкнутость равносильна пи-(дельта)-замкнутости.

Нетривиальным примером алгебры множеств является следующая конструкция, также перекидывающая мостик между теорией множеств и Алгеброй. Рассмотрим множество X и биекцию $f : X \leftrightarrow X$ на нем. Назовем множество $D \subseteq X$ **неподвижным** (относительно f), если $fD = D$ (т. е. f действует биективно на D). Как минимум все множество X и пустое множество являются неподвижными относительно любой биекции. Множество всех неподвижных множеств, которое мы далее будем обозначать $\mathcal{A}(X; f)$, образует алгебру множеств, причем не просто алгебру, а сигма-алгебру, и даже более того, в этой алгебре любое пересечение и сумма элементов не выводят за ее пределы. По аналогии с обозначениями в Алгебре через $\mathcal{A}(X; f)^*$ обозначим $\mathcal{A}(X; f) \setminus \{\emptyset\}$, т. е. удалим аддитивный ноль. | Упражнение 3.3.

Примером тривиальной алгебры неподвижных множеств является алгебра, порожденная биекцией на окружности S^1 , осуществляющая ее поворот на ненулевой угол $< 2\pi$. Она состоит из двух элементов: S^1 и \emptyset . В том случае, когда биекция на X является тождественной функцией, неподвижными множествами будут все подмножества X , и в этом случае алгебра неподвижных множеств совпадет с $\mathcal{P}(X)$.

Еще один пример алгебры связан с разбиением или фактор-множеством. Пусть Z есть разбиение X . Включим в \mathcal{A} все такие множества, которые получаются как конечные суммы элементов разбиения Z и/или их дополнений.

Множество \emptyset получается как пустая сумма элементов Z . Поскольку Z есть разбиение X , любой элемент \mathcal{A} есть либо конечная сумма элементов Z , либо сумма всех элементов Z за исключением конечного набора. Ясно, что и дополнения, и объединения элементов \mathcal{A} имеют такую же структуру, так что \mathcal{A} есть алгебра множеств, причем это минимальная алгебра, содержащая Z .

Следующая конструкция позволяет получить сигма-алгебру из любого счетного семейства подмножеств некоторого непустого множества X . Пусть $\{A_\lambda\}$ — семейство множеств, где $\lambda \in \omega$ (т. е. $A : \omega \rightarrow \mathcal{P}(X)$). Определим

$$A_\lambda^\delta = \begin{cases} A_\lambda, & \text{если } \delta = 1, \\ X \setminus A_\lambda, & \text{если } \delta = 0. \end{cases}$$

Далее, для произвольной функции $\beta : \omega \rightarrow \{0, 1\}$ положим

$$B(\beta) = \bigcap_{\lambda} A_\lambda^{\beta(\lambda)}.$$

Легко проверить, что разным функциям β соответствуют разные множества $B(\beta)$. Совокупность $\{B(\beta) \mid \beta \in 2^\omega\}$ представляет собой сигма-алгебру множеств, минимальную над семейством $\{A_\lambda\}$.

Константы — тоже операции, только тривиальные. Алгебра множеств является иллюстрацией еще одной алгебры — булевой. Пусть на множестве A заданы три операции $a \vee b$, $a \wedge b$, $\neg a$ и две константы 1 и 0. В полной аналогии с логическими связками, имеющими такое же обозначение, эти операции подчиняются правилам:

В1 $a \vee (b \vee c) = (a \vee b) \vee c$; $a \wedge (b \wedge c) = (a \wedge b) \wedge c$ (ассоциативность)

В2 $a \vee b = b \vee a$; $a \wedge b = b \wedge a$ (коммутативность)

В3 $a \vee (a \wedge b) = a$; $a \wedge (a \vee b) = a$ (законы поглощения)

В4 $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$; $a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$ (дистрибутивность)

В5 $a \vee \neg a = 1$; $a \wedge \neg a = 0$ (дополнительность)

Тогда структура $(A, \vee, \wedge, \neg, 0, 1)$ называется **булевой алгеброй**.

Если в алгебре множеств объединение обозначить как \vee , пересечение как \wedge , дополнение $X \setminus a$ как $\neg a$, в качестве 1 взять X , а в качестве нуля — пустое множество, то алгебра множеств будет замкнута относительно данных операций и будет удовлетворять аксиомам булевой алгебры. В этом свойстве кроется тесная связь между логикой и теорией множеств, позволяющая часто подменять логические соотношения теоретико-множественными и наоборот (особенно ярко это видно на примере алгебры событий в теории вероятностей).

Наконец, еще одно понятие, связанное с системой множеств. Алгебра (A, \wedge, \vee) называется **решёткой**, если

L1 $a \vee a = a$; $a \wedge a = a$ (идемпотентность);

L2 $a \vee (b \vee c) = (a \vee b) \vee c$; $a \wedge (b \wedge c) = (a \wedge b) \wedge c$ (ассоциативность);

L3 $a \vee b = b \vee a$; $a \wedge b = b \wedge a$ (коммутативность);

L4 $a \vee (a \wedge b) = a$; $a \wedge (a \vee b) = a$ (законы поглощения).

На решетке можно определить отношение $a \leq b$ по правилу:

$$a \leq b \leftrightarrow a \wedge b = a,$$

это определение эквивалентно:

$$a \leq b \leftrightarrow a \vee b = b.$$

Такое отношение на A будет частичным порядком. Таким образом, решетка порождает частично упорядоченное множество. При этом оно обладает следующим свойством: каждая пара элементов в нем имеет верхнюю и нижнюю точные грани, которые можно определить следующим образом:

$$\sup\{a, b\} = a \vee b; \quad \inf\{a, b\} = a \wedge b. \quad (3.1)$$

Иногда решетку определяют именно как *упорядоченное множество, в котором любое двухэлементное множество имеет точную верхнюю и точную нижнюю грани*. При этом переход от отношения к операциям осуществляется по указанным в (3.1) формулам.

Хорошим примером решетки с дополнительно заданным соответствующим отношением порядка является алгебра множеств, где

$$a \wedge b \equiv a \cap b; \quad a \vee b \equiv a \cup b,$$

при этом получится, что $a \leq b$, если $a \subseteq b$.

Соответственно, решеткой является и всякая алгебра неподвижных множеств $\mathcal{A}(X; f)$ относительно биекции $f : X \leftrightarrow X$. При этом, отношение порядка \subseteq на этой алгебре множеств обладает одним замечательным свойством: минимальные элементы $\mathcal{A}(X; f)^*$ образуют разбиение множества X .¹ Действительно, во-первых, любые два непустых минимальных неподвижных множества не пересекаются (иначе бы их пересечение было меньше их самих), во-вторых, для любой точки $x \in X$ пересечение всех содержащих ее

Упражнение 3.4.
Докажите эквивалентность.

Упражнение 3.5.
Проверьте выполнение аксиом порядка.

Упражнение 3.6.
Проверьте, что это так.

¹Заметим, что не для всякой алгебры множеств это верно, например, алгебра открытых интервалов на вещественной прямой не обладает этим свойством. С другой стороны, борелевская сигма-алгебра порождает тривиальное разбиение, состоящее из одноточечных множеств.

неподвижных множеств является неподвижным множеством, минимальным в $\mathcal{A}(X; f)^*$. Это значит, что минимальные множества в сумме дают все X , так что их совокупность представляет собой разбиение X .

Разбиение X на минимальные неподвижные относительно биекции f множества единственное, причем, для всякого минимального множества D сужение $f|_D$ является биекцией на D , а алгебра $\mathcal{A}(D; f|_D)$ тривиальна. Таким образом, биекцию f мы также можем единственным способом разложить в сумму *простых (неразложимых)* биекций, соответствующих минимальным неподвижным множествам. Такие неразложимые биекции называются **циклами** биекции f . Обратно, имея произвольное разбиение X и неразложимые биекции на компонентах этого разбиения, мы получаем единственную биекцию f на множестве X .

Упражнение
3.7.
Проверьте
выполнение
L1 для
булевой
алгебры

Аналогично, решеткой будет любая булева алгебра, в которой обозначения операций в точности совпадают с таковыми для решетки.

Другой пример решетки — линейно-упорядоченное множество, в котором $a \vee b = \max\{a, b\}$, $a \wedge b = \min\{a, b\}$.

Наконец, если на \mathbb{N}^+ (т. е. \mathbb{N} без нуля) задать отношение частичного порядка $a \leq b$ как « a делит b », то \sup будет наименьшее общее кратное, а \inf — наибольший общий делитель, соответственно, (\mathbb{N}^+, \leq) будет решеткой в смысле нашего второго определения. При этом мы можем задать операции \wedge и \vee как НОД и НОК, и увидеть, что здесь законы поглощения выполняются в силу арифметических тождеств

$$\text{НОК}(a, \text{НОД}(a, b)) = a, \quad \text{НОД}(a, \text{НОК}(a, b)) = a.$$

Однако аксиома B5 булевой алгебры не выполняется, поскольку из B3 и B5 легко вывести свойства $a \wedge 1 = a$ и $a \vee 0 = a$, т. е. $\text{НОД}(a, 1) = a$ и $\text{НОК}(a, 0) = a$. Но если второе равенство удовлетворяется, если в качестве нуля взять натуральное число 1, то первое невозможно при замене константы 1 на любое положительное целое число.²

3.1.4 Немного теории групп

Вернемся к одному из важнейших понятий современной Алгебры — группам. Итак, **группа** — это алгебраическая структура с одной бинарной операцией, подчиняющейся аксиомам ассоциативности, существования нейтрального и обратных элементов. Абелева группа подчиняется также аксиоме коммутативности операции. **Порядком** группы называется мощность ее базового множества.

²Если бы мы рассматривали \mathbb{N} с нулем, то именно целое число 0 подошло бы в качестве 1 в данном случае. При этом ноль был бы максимальным элементом \mathbb{N} в смысле отношения делимости.

Простейший и самый понятный пример группы из рассмотренных нами числовых систем — это целые числа с операцией сложения: $(\mathbb{Z}, +)$.

Еще один очень важный пример — это группа вычетов по модулю n . Рассмотрим множество остатков от деления (вычетов) по модулю n :

$$G = \{0, 1, \dots, n-1\}, \quad x \circ y = x + y \pmod n.$$

Легко проверить, что G с такой операцией \circ будет абелевой группой.

Важное замечание: далее во всех случаях, если не оговорено обратное, групповую операцию \circ мы считаем умножением и соответствующим образом записываем, а нейтральный элемент обозначаем e .³

Далее мы рассмотрим еще несколько примеров групп и посмотрим, как они связаны между собой.

Пусть (G, \cdot) — группоид. Для любого множества $H \subseteq G$ и элемента $g \in G$ определим

$$gH = \{gh \mid h \in H\}, \quad Hg = \{hg \mid h \in H\}.$$

H есть **подгруппа** G , если $H \subseteq G$ и $(H, \cdot|_{H \times H})$ — группа.⁴

Например, $2\mathbb{Z}$ является подгруппой \mathbb{Z} по сложению и содержит все четные числа.

Пусть G — группа, H — ее подгруппа и выполнено условие:

$$\forall g \in G : gH = Hg.$$

В этом случае говорят, что H является **нормальной подгруппой** группы G , что записывается как $H \triangleleft G$. Очевидно, что в абелевой группе все подгруппы нормальные.

В каждой группе существует две тривиальных нормальных подгруппы — единичная $\{e\}$ и вся группа G . Если в группе G нет нетривиальных нормальных подгрупп, то такую группу называют **простой**. Группа $(\mathbb{Z}, +)$ — не простая, поскольку любая подгруппа вида $n\mathbb{Z}$ будет нормальной.

Если H — подгруппа G , то множества gH и Hg называются левыми и правыми **классами смежности** H , порожденными элементом g . Классы смежности обладают такими свойствами:

$$1^\circ \quad |gH| = |H|$$

2° G есть прямая сумма классов смежности по подгруппе H (т.е. $G = \bigcup_{g \in G} gH$ и все классы gH попарно либо не пересекаются, либо совпадают)

3° $g_1H = g_2H$ тогда и только тогда, когда $g_1^{-1}g_2 \in H$

³Мы вводим здесь отдельное обозначение e для нейтрального элемента группы, поскольку в комплексном анализе уже заняли букву e .

⁴ $\cdot|_{H \times H}$ обозначает сужение операции \cdot на H , в дальнейшем будем пропускать подобные уточнения.

здесь
 $2\mathbb{Z} = gH$ в
моноиде
 (\mathbb{Z}, \cdot) .

Снова
архетип
порождающего
элемента

Упражнение
3.8.
Докажите
свойства
1°–3°.

Такие же свойства справедливы для правых классов смежности. Заметим, что первые два свойства означают, что *порядок подгруппы делит нацело порядок группы* (в смысле арифметики кардиналов и натуральных чисел). Данное утверждение называется Теоремой Лагранжа. Так, если группа вычетов G имеет порядок n , то она может иметь подгруппы только порядка k , делящего n . В частности, при простом числе n группа вычетов также будет простой, т. е. будет иметь только тривиальные подгруппы.

Возвращаясь к целым числам, видим, что $2\mathbb{Z} \triangleleft \mathbb{Z}$, а классами смежности являются множества вида $2\mathbb{Z} + k$, причем если k четное, то мы получаем класс $2\mathbb{Z}$, а если нечетное, то $2\mathbb{Z} + 1$, и других классов по подгруппе $2\mathbb{Z}$ нет. Критерий совпадения классов $2\mathbb{Z} + k$ и $2\mathbb{Z} + m$ в этом случае принимает вид $k - m \in 2\mathbb{Z}$, т. е. числа k и m имеют одинаковую четность.

Будем умножать подмножества группы так же, как элемент на подмножество: пусть $K, L \subseteq G$, тогда

$$KL = \{kl \mid k \in K \wedge l \in L\}$$

Если подгруппа H — нормальная, то $(g_1H)(g_2H) = (g_1g_2)H$, т. е. операция на смежных классах этой подгруппы сводится к операции над элементами группы.

В случае $H \triangleleft G$ множество $\{gH \mid g \in G\}$ является группой относительно данной операции умножения подмножеств, причем единицей в этой группе будет H , а обратным к gH элементом будет $g^{-1}H$. Такая группа, заданная на подмножестве $\mathcal{P}(G)$, называется **фактор-группой** по нормальной подгруппе H и обозначается G/H .

Например, $\mathbb{Z}/2\mathbb{Z} = \{2\mathbb{Z}, 2\mathbb{Z} + 1\}$ изоморфна группе вычетов по модулю 2, в которой операцией является сложение по модулю 2. Аналогично, $\mathbb{Z}/n\mathbb{Z}$ изоморфна группе вычетов по модулю n . Принято отождествлять указанные фактор-группы и изоморфные им группы вычетов. Часто вместо $\mathbb{Z}/n\mathbb{Z}$ пишут \mathbb{Z}_n .

$\mathbb{Z}_1 = \mathbb{Z}/\mathbb{Z}$ — тривиальная группа с 1 элементом.

Если $n = km$, то в группе \mathbb{Z}_n существует нормальная подгруппа H порядка k , изоморфная \mathbb{Z}_k . В эту подгруппу входят элементы $\{0, m, 2m, \dots, n - m\} = m\{0, 1, \dots, k - 1\}$. Сложение по модулю n в этой подгруппе равносильно сложению по модулю k в \mathbb{Z}_k с последующим умножением на m , что вытекает из свойства сравнений:

$$(am \equiv bm \pmod{n}) \leftrightarrow (a \equiv b \pmod{k})$$

Умножение на m устанавливает указанный изоморфизм между группой \mathbb{Z}_k и подгруппой H .



Жозеф Луи
Лагранж

Напомним, что в поле комплексных чисел существует группа корней из 1. Это комплексные числа вида $e^{2i\pi k/n}$, $k = 0, \dots, n-1$. Все корни можно получить, возводя в степень k корень $e^{2i\pi/n}$, причем k достаточно брать по модулю n (сложение в группе \mathbb{Z}_n изоморфно умножению в группе корней). Таким образом, группа корней из 1 имеет структуру

$$G = \{a, a^2, a^3, \dots\}.$$

Группы, состоящие из степеней одного элемента a и обратных к ним (т.е. в качестве степени нужно использовать целые числа), называются **циклическими** с образующим элементом a . В циклической группе может быть несколько образующих группу элементов.

Заметим, что в группе корней из 1 существует несколько образующих элементов. Если $k \perp n$, то корень $e^{2i\pi k/n}$ также является образующим, поскольку все произведения $k, 2k, \dots, nk$ различны по модулю n в силу теоремы 2.8. У образующего элемента **порядок** (т.е. минимальная степень, в которой он равен e) совпадает с порядком группы. Для других элементов он меньше. Например, в группе корней из 1 степени 4 корень $e^{i\pi/2}$ является образующим и имеет порядок 4, а корень $e^{i\pi}$ имеет порядок 2.

Упражнение 3.9. Покажите, что в конечной группе $g^n = e$ при некотором n , зависящем от g .

Множество всех степеней одного элемента группы составляют ее подгруппу. Из теоремы Лагранжа следует, что порядок элемента делит порядок группы. Отсюда, в частности, следует, что $g^{|G|} = e$ для любого $g \in G$.

Если элементы группы G представляют собой все возможные произведения целых степеней элементов некоторого множества $T \subseteq G$, то T называется **системой образующих** группы G и этот факт записывается следующим образом: $G = \langle T \rangle$. В частности, для циклической группы имеем $G = \langle a \rangle$ (фигурные скобки у синглтона $T = \{a\}$ мы опускаем для упрощения записи). Множество T образующих называется **минимальной системой образующих** или **базисом**, если любое собственное подмножество $T' \subset T$ не является системой образующих. Базисов может быть несколько, например, у группы \mathbb{Z}_9 имеется 6 одноэлементных базисов $\{1\}, \{2\}, \{4\}, \{5\}, \{7\}, \{8\}$.

Заметим, что если операцию группы обозначить как $+$, то все элементы группы будут линейными комбинациями элементов базиса с целыми коэффициентами (вместо $g_1^{s_1} \dots g_k^{s_k}$ мы запишем $g_1 s_1 + \dots + g_k s_k$, где g_i — элементы базиса, $s_i \in \mathbb{Z}$), и в данном случае мы видим здесь полную аналогию с векторным пространством.

Комментарий 4.

На самом деле, базис — это все тот же архетип порождающего элемента. В математике мы стремимся упаковать данные так, чтобы максимальная нагрузка ложилась на вычисление сущностей, и минимальная — на их основу: базис,

систему порождающих, аксиоматику, теоремы, правила генерации. В этом смысле деятельность математиков можно назвать квинтэссенцией борьбы человека с энтропией — мы упаковываем знание так конструктивно и плотно, как никто другой.

Если сравнить эту деятельность с обработкой BigData, то в математике мы все стараемся свести к вычислению агрегированных данных, оставляя невычисленными только исходные данные (и то, чаще всего мы и в них ищем закономерности и уstraняем зависимые данные), в то время как в областях, далеких от математики, стараются оперировать готовыми, вычисленными на все случаи жизни массивами, срезами данных, минимизируя вычисления и повышая энтропию.

- Рассмотрим некоторые свойства циклических групп:

 - 1° Циклические группы являются абелевыми
 - 2° Конечная циклическая группа порядка n изоморфна \mathbb{Z}_n
 - 3° Бесконечная циклическая группа изоморфна $(\mathbb{Z}, +)$
 - 4° Каждая подгруппа циклической группы циклична
 - 5° У циклической группы порядка n существует ровно $\varphi(n)$ порождающих элементов, где φ — функция Эйлера
 - 6° Если p — простое число, то любая группа G порядка p циклическая и единственна с точностью до изоморфизма

*Упражнение
3.10.
Докажите
1°–5°*

Последнее свойство доказывается следующим образом. Возьмем любой элемент $g \neq e$ (при простом p это всегда возможно), тогда в группе G имеется циклическая подгруппа $\langle g \rangle$, порядок которой делит p . Но p — простое, а группа $\langle g \rangle$ нетривиальная, следовательно, $G = \langle g \rangle$.

*Упражнение
3.11.
Докажите,
что \mathbb{Z}_n^* —
группа.*

Через \mathbb{Z}_n^* обозначим множество всех $k \in \mathbb{Z}_n$, взаимно простых с n . Очевидно, что $|\mathbb{Z}_n^*| = \varphi(n)$. Множество \mathbb{Z}_n^* образует группу с операцией умножения по модулю n . Как мы уже видели выше, любой элемент группы, будучи возведенным в степень, равную порядку группы, дает единицу. Стало быть, $k^{\varphi(n)} \equiv 1 \pmod n$ для $k \in \mathbb{Z}_n^*$. А это — в точности теорема Эйлера 2.10.

Например, пусть $n = 9$, тогда таблица умножения группы \mathbb{Z}_9^* и таблица сложения изоморфной ей группы \mathbb{Z}_6 выглядят так:

\mathbb{Z}_9^*	1	2	4	5	7	8		\mathbb{Z}_6	0	1	2	5	4	3
1	1	2	4	5	7	8		0	0	1	2	5	4	3
2	2	4	8	1	5	7		1	1	2	3	0	5	4
4	4	8	7	2	1	5		2	2	3	4	1	0	5
5	5	1	2	7	8	4		5	5	0	1	4	3	2
7	7	5	1	8	4	2		4	4	5	0	3	2	1
8	8	7	5	4	2	1		3	3	4	5	2	1	0

Во второй таблице мы специально перемешали порядок элементов таким образом, чтобы показать изоморфизм групп, при котором умножение в \mathbb{Z}_9^* соответствует сложению в \mathbb{Z}_6 , а соответствие элементов можно установить по правилу: $2^a \equiv b \pmod 9$, где $a \in \mathbb{Z}_6$, $b \in \mathbb{Z}_9^*$, поскольку $\mathbb{Z}_9^* = \langle 2 \rangle$. Аналогичное соответствие можно построить, опираясь на степени любого другого элемента \mathbb{Z}_9^* (кроме единицы).

Заметим, что не любая группа \mathbb{Z}_m^* изоморфна некоторой группе \mathbb{Z}_n . Например, в группе \mathbb{Z}_8^* содержится 4 элемента, но ни один из них не является образующим, группа \mathbb{Z}_8^* не является циклической, а значит, она не может быть изоморфна \mathbb{Z}_4 .

*Упражнение 3.12.
Постройте таблицу умножения \mathbb{Z}_8^**

Интересный частный случай группы \mathbb{Z}_n получается при простом $n = p$. В этом случае \mathbb{Z}_p^* содержит все ненулевые элементы \mathbb{Z}_p , а значит, \mathbb{Z}_p является полем (поскольку это абелева группа по сложению по модулю p , а без нуля — абелева группа по умножению по модулю p). Это — пример конечного поля.

Следующий простой пример: группа биекций произвольного множества. Рассмотрим множество X и все биекции вида $f : X \rightarrow X$. В качестве операции над ними возьмем композицию, т. е. $(f \circ g)(x) = f(g(x))$. Нетрудно видеть, что множество всех биекций с операцией композиции образует группу. Ее мы обозначим $S(X)$. В группе биекций $S(X)$ можно выделять различные интересные подгруппы, связывая биекции некоторыми дополнительными условиями, например, линейностью или сохранением отношений.

В том случае, если мы рассматриваем только биекции, сохраняющие отношения на X и/или операции на X , т. е. являющиеся автоморфизмами, то мы имеем дело с **группой автоморфизмов**. В частности, если X — группа с операцией \circ , то можно рассмотреть группу автоморфизмов, сохраняющих операцию \circ . Такая группа обозначается $\text{Aut}(X, \circ)$ или, в соответствии с архетипом базового множества, $\text{Aut}(X)$.

В случае конечных групп определенный интерес представляют собой группы автоморфизмов $\text{Aut}(\mathbb{Z}_m)$. Например, группа $\text{Aut}(\mathbb{Z}_9)$ изоморфна \mathbb{Z}_6 , поскольку в ней ровно 6 автоморфизмов, имеющих вид

$$a_1(x) = 2x, \quad a_2(x) = 4x, \quad a_3(x) = 8x, \quad a_4(x) = 7x, \quad a_5(x) = 5x, \quad a_6(x) = x,$$

где коэффициенты перед x — это степени числа 2 по модулю 9, которые, как мы видели выше, образуют циклическую группу \mathbb{Z}_6^* . В то же время, это

означает, что автоморфизм a_k есть k -ая степень композиции автоморфизма a_1 с самим собой, т.е. группа $\text{Aut}(\mathbb{Z}_9) = \langle a_1 \rangle$ является циклической группой порядка 6.

Если теперь подняться еще на одну ступень и рассмотреть группу $\text{AutAut}(\mathbb{Z}_9)$, изоморфную $\text{Aut}(\mathbb{Z}_6)$, то мы получим группу, изоморфную \mathbb{Z}_2 , поскольку $\varphi(6) = 2$.

Верно ли,
что те, кто
считает 0
натураль-
ным числом,
говорят «пе-
рестановка»,
а все
остальные —
«подстано-
вка»?

Пусть множество X_n имеет мощность $n < \omega$, тогда группа всех биекций (без доп.условий на эти биекции) X_n называется **груп-пой подстановок** (перестановок) и обозначается S_n . Порядок такой группы равен $n!$. В качестве множества X_n обычно выбирается $\{1, 2, \dots, n\}$, т.е. рассматриваются все возможные перестановки начального отрезка положительных целых чисел.

Теория групп в XIX в. начиналась именно с изучения групп подстановок, и лишь позже понятие группы было обобщено Артуром Кэли. Он же сделал первый важный шаг на пути классификации групп.

Теорема 3.1 (Кэли). *Любая конечная группа порядка n изоморфна некото-рой подгруппе S_n .*

Для доказательства достаточно заметить, что каждый элемент g исходной группы G порождает биекцию на G по правилу $h \mapsto gh$ («правые» биекции), а эти биекции образуют изоморфную G подгруппу внутри группы биекций на G (см., например, [66]). На самом деле, вообще любая группа мощности τ вкладывается изоморфно в группу биекций на кардинале τ как подгруппа.

В группе S_n , как и в любой другой, можно построить циклическую подгруппу, отправляясь от произвольно взятого элемента, т.е. биекции на X_n . Например, пусть $s \in S_n$, тогда можно рассмотреть циклическую подгруппу $G(s) = \{s, s^2, s^3, \dots\}$, где под степенью понимается многократная композиция биекции s с самою собой. Ясно, что эта подгруппа не может быть бесконечной, т.к. она входит в конечную группу, поэтому при некотором k имеем $s^k = e$, где e — тождественная биекция.

Рассмотрим некоторую перестановку $s \in S_n$. Ее можно записать в виде таблицы аргумент–значение:

$$s = \begin{pmatrix} 1 & 2 & \dots & n-1 & n \\ s_1 & s_2 & \dots & s_{n-1} & s_n \end{pmatrix}$$

т.е. $s(i) = s_i$. При этом $\{1, 2, \dots, n\} = \{s_1, s_2, \dots, s_n\}$ в полном соответствии с определением равенства множеств.

Возьмем теперь элемент 1 и начнем «раскручивать» его так же, как мы «раскручивали» степени элемента в циклической группе:

$$1 \mapsto s(1) \mapsto s(s(1)) \mapsto \dots \mapsto s^k(1)$$

Мы получим то, что называется **орбитой** элемента 1 при действии группы $G(s)$ на множестве X_n . Действительно, все элементы данной цепочки составляют множество $G(s)1 = \{g(1) \mid g \in G(s)\}$. Кроме того, если $s^k = e$, то $s^k(1) = 1$, и мы получаем **цикл**:

$$(1 \ s(1) \ s(s(1)) \ \dots \ s^{k-1}(1))$$

(единицу в конце мы не пишем, подразумевая, что последний элемент цикла переходит в первый).

Действие группы $G(s)$ на множестве X_n позволяет разбить это множество на несколько попарно непересекающихся орбит или циклов. Отсюда мы получаем представление самой перестановки s как набора независимых циклов. Поэтому перестановки принято записывать в виде последовательности циклов. Например, пусть

$$s = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 1 & 3 & 5 \end{pmatrix}$$

В этой перестановке мы наблюдаем два цикла: (1243) и тривиальный (5). Тогда

$$s = (1243)(5),$$

причем, тривиальные циклы принято пропускать в такой «циклической» записи, т.к. они однозначно восстанавливаются по всем остальным циклам и по параметру n (в нашем случае $n = 5$).

Отметим, что поскольку перестановка s является биекцией на X_n , ей соответствует алгебра неподвижных множеств $\mathcal{A}(X_n; s)$, которая содержит разбиение X_n на минимальные неподвижные непустые подмножества, которые соответствуют неразложимым биекциям, в сумме дающим перестановку s . Эти неразложимые биекции мы называли ранее циклами биекции s (см. стр. 204). Так вот, эти циклы и есть в точности циклы перестановки s .

Рассмотрим более сложный пример:

$$s = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 4 & 3 & 1 & 7 & 5 & 6 \end{pmatrix} = (124)(3)(576) = (124)(576)$$

Предположим теперь, что у нас имеется три перестановки:

$$s_1 = \begin{pmatrix} 1 & 2 & 4 \\ 2 & 4 & 1 \end{pmatrix} \quad s_2 = \begin{pmatrix} 5 & 6 & 7 \\ 7 & 5 & 6 \end{pmatrix} \quad s_3 = e$$

Тогда исходная перестановка s получается как последовательное применение этих новых перестановок:

$$s = s_1 s_2 s_3,$$

причем порядок перестановок в композиции неважен, т.к. две их них «работают» на разных орбитах, а третья тождественна и коммутирует с любой перестановкой.

Жаль, что
неверно
 $\min = m \cdot i \cdot n$
:(

Таким образом, каждую перестановку из S_n можно единственным образом (с точностью до порядка) представить как композицию циклов, и, таким образом, запись перестановки в виде набора ее циклов является не только удобным соглашением, но еще и функционально верной.

Наконец, введем такое понятие как **транспозиция**. Это — микроцикл, состоящий из двух элементов, например, (12) или (59) и т.п. Транспозиция меняет местами два элемента X_n , а остальные оставляет на месте. Любой цикл длины k можно представить как композицию $k - 1$ транспозиций. Например,

$$(1234) = (14)(13)(12),$$

причем, это представление неоднозначное, поскольку:

$$(2341) = (21)(24)(23),$$

тем не менее, любая перестановка (не только цикл) имеет *инвариант* по разложению в транспозиции.

Теорема 3.2. Если перестановка $g \in S_n$ имеет два представления транспозициями

$$t_1 \dots t_k = g = \tau_1 \dots \tau_m,$$

то $k \equiv m \pmod{2}$.

Иначе говоря, четность перестановки, определяемая количеством входящих в ее разложение транспозиций, не зависит от способа этого разложения. Величина $\text{sgn}(g) = (-1)^k = (-1)^m$ называется **знаком перестановки** g .

Упражнение 3.13. Доказательство этой теоремы можно также найти в [66] или доказать самостоятельно индукцией по сложности перестановки.

Функция sgn , определенная на элементах группы S_n и принимающая значения из множества $B = \{-1, 1\}$, является гомоморфизмом групп S_n и B (проверьте, что (B, \cdot) есть группа по умножению, изоморфная \mathbb{Z}_2).

В связи с этим дадим общее определение: **гомоморфизмом групп** (G, \cdot) и (G', \circ) называется всякая функция $h : G \rightarrow G'$, сохраняющая групповую операцию, т. е.

$$h(g_1 \cdot g_2) = h(g_1) \circ h(g_2).$$

Изоморфизм — частный случай гомоморфизма. Ядром гомоморфизма h называется прообраз единицы:

$$\text{Ker}(h) = h^{-1}\{e'\} = \{g \in G \mid h(g) = e'\},$$

где e' — единица группы G' .

Гомоморфизм обладает слеующими свойствами

- 1° Ядро гомоморфизма есть нормальная подгруппа: $\text{Ker}(h) \triangleleft G$.
- 2° Если $H \triangleleft G$, то существует гомоморфизм $h : G \rightarrow G/H$ такой, что $H = \text{Ker}(h)$.
- 3° Фактор-группа $G/\text{Ker}(h)$ изоморфна образу hG в группе G' .

Нетрудно видеть, что функция sgn на группе S_n действует как гомоморфизм в группу B , поэтому прообраз 1 в группе S_n относительно данного гомоморфизма, а именно, *все четные перестановки* образуют нормальную подгруппу в группе подстановок S_n . Эта нормальная подгруппа обозначается A_n и называется **знакопеременной группой** порядка n . Следует не путать употребленное здесь слово «порядок» с порядком группы, означающем мощность группы как множества, поскольку в A_n находится ровно половина элементов группы S_n , т. е. $n!/2$, что значительно больше n .

*Упражнение 3.14.
Докажите это.*

Функция sgn является инвариантом на подгруппе A_n , а также на ее смежном классе в S_n .

Здесь мы, пожалуй, впервые явно сталкиваемся с еще одним важным архетипом математики — инвариантом. **Инвариант** — это некоторая функция, заданная для однородных объектов и имеющая постоянное значение, т. е. константа. Инвариант хорош обычно тем, что выявляет нетривиальные свойства объектов, на которых он задан.

Комментарий 5.

С конца XIX века известна игра «пятнадцать», суть которой в следующем. Имеем поле 4×4 , в котором расставлены одинаковые по размеру фишки размером 1×1 . Всего фишек 15, и они пронумерованы числами от 1 до 15. Одно место на поле пустое, что позволяет производить следующие простые манипуляции: занимать данное место фишкой с любого смежного места, т. е. передвигать ее на это место, освобождая соседнее. При этом нельзя совершать никакие другие действия, например, вынимать фишки с поля и расставлять их произвольным образом.

В результате таких действий порядок номеров у фишек меняется, т. е. мы осуществляем перестановку из группы S_{15} .

«Фишка» этой игры в том, что все разрешенные манипуляции не меняют четности исходной перестановки номеров. А это значит, что никакую нечетную изначальную расстановку невозможно привести (разрешенными действиями) к четной перестановке, и наоборот. Например, две расстановки фишек, отличающиеся лишь одной транспозицией (обменом двух соседних фишек местами), не могут быть переведены одна в другую.

Создатель игры (никто еще не знал тогда алгебраического решения задачи) даже обещал приз 100 долларов тому, кто приведет расстановку

1	2	3	4
5	6	7	8
5	6	7	8
9	10	11	12
13	15	14	

к виду

1	2	3	4
5	6	7	8
5	6	7	8
9	10	11	12
13	14	15	

(они отличаются транспозицией (15 14)).

С тех пор прошло больше 100 лет, и до сих пор многие пытаются это сделать, но алгебра дает нам беспощадный ответ: это сделать невозможно! Потому что четность перестановки инвариантна относительно действий с фишками!

Другой замечательный пример инварианта — теорема Эйлера о числе выпуклого многогранника: величина $V-P+G=2$ для всех выпуклых многогранников (V — число вершин, P — ребер, G — граней). Отсюда, в частности, следует, что на футбольном мяче, сшитом только из правильных 5- и 6-угольников, может быть только 12 пятиугольников, никакое другое число не удовлетворяет этому инварианту.

Факторизацию группы можно воспринимать как делимость групп, и в этом смысле группы становятся подобны числам. Есть простые группы — они ни на что не делятся, а есть сложные — они делятся на нормальные подгруппы.

Естественно ввести и умножение групп. **Внешним прямым произведением** групп (G, \cdot) и (H, \circ) называется группа $(G \times H, \star)$ такая, что:

$$(g, h) \star (g', h') = (g \cdot g', h \circ h'). \quad (3.2)$$

Упражнение 3.15. Проверьте, что это действительно группа, причем если исходные группы — абелевы, то и их прямое произведение также абелево.

Если в исходных группах операция интерпретируется как сложение, то прямое произведение называют прямой суммой групп. Но, учитывая, что это может привести к путанице понятий, мы в любом случае будем пользоваться мультипликативной терминологией и символикой. Тем более, что она согласуется с теоретико-множественным прямым произведением.

Мы уже видели выше, что группа \mathbb{Z}_8^* имеет порядок 4, но не изоморфна группе вычетов \mathbb{Z}_4 . Теперь мы можем легко увидеть из таблиц умножения, что $\mathbb{Z}_8^* \cong (\mathbb{Z}_2 \times \mathbb{Z}_2)^5$

⁵Мы опять редуцировали группу до ее базового множества и обозначили как прямое произведение множеств. По умолчанию предполагается, что операция на этом множестве подчинена (3.2)

\mathbb{Z}_8^*	1	3	5	7	$\mathbb{Z}_2 \times \mathbb{Z}_2$	00	01	10	11
1	1	3	5	7	00	00	01	10	11
3	3	1	7	5	01	01	00	11	10
5	5	7	1	3	10	10	11	00	01
7	7	5	3	1	11	11	10	01	00

Здесь мы использовали упрощенную запись для пар вида $(1, 0)$, обозначив их двоичными числами. Операция сложения в группе $\mathbb{Z}_2 \times \mathbb{Z}_2$ является побитовым сложением по модулю 2, когда сложение осуществляется независимо в каждом разряде числа. Как видим, с точностью до переобозначений перед нами одна и та же группа.

Этот пример показывает нам, что порядок группы не определяют однозначно ее структуру (как нам того бы хотелось, памятуя об основной теореме арифметики) — группа вычетов \mathbb{Z}_4 имеет единственную нетривиальную подгруппу $\{0, 2\}$, а $\mathbb{Z}_2 \times \mathbb{Z}_2$ имеет три независимых подгруппы, изоморфных \mathbb{Z}_2 : $\{00, 01\}$, $\{00, 10\}$, $\{00, 11\}$.

Группа $\mathbb{Z}_2 \times \mathbb{Z}_2$ имеет изоморфный клон среди подгрупп группы перестановок S_4 . Этот клон носит свое собственное название «**четверная группа Клейна**» и обозначается V_4 . В таблице В.4 (в конце книги) помещена полная таблица умножения группы S_4 с использованием кратких обозначений перестановок как произведений циклов. Там же выделены две подтаблицы, отвечающие группам A_4 и V_4 , а также отмечены (жёлтым) элементы (и их произведения) подгруппы 8-го порядка, которая не является ни нормальной, ни абелевой.

А также группа диэдра D_2 .

Выделенная подгруппа 8-го порядка:

$$\{e, (12)(34), (13)(24), (14)(23), (12), (34), (1324), (1423)\}.$$

Все подгруппы 8-го порядка изоморфны. Аналогичная ситуация с подгруппами 6-го порядка, вот одна из них: $\{e, (123), (132), (12), (13), (23)\}$, которая совпадает с S_3 .

Комментарий 6.

Вычисление группы A_4 вручную оказалось столь утомительным, что для построения полной таблицы умножения S_4 я решил написать скрипт в Google Spreadsheet. Покорпев субботним утром примерно 2 часа, я получил скрипт, умеющий умножать две перестановки, заданные циклами (без верификации на корректность написания циклов и отсутствие у них пересечений). После чего построение полной таблицы не заняло и 5 минут. Результат был скопирован в эту книгу в таблицу В.4.

Кроме того, в листинге С.3 приведен алгоритм умножения перестановок на языке Python.

Группа V_4 примечательна также тем, что это группа симметрий вытянутого ромба, состоящая из 4 преобразований: тождественного, двух симметрий относительно диагоналей и поворота на 180° относительно центра, а также V_4 является группой поворотов сферы вокруг трех ортогональных осей на 180°

В группе S_4 существует только 2 нетривиальные нормальные подгруппы: A_4 и V_4 . Все подгруппы порядков 2, 3, 6, 8 не являются нормальными.

Ненормальные?
:)

Говорят, что группа G имеет **субнормальный ряд** (называемый также **субнормальной башней**, **субинвариантным рядом**, **субнормальной матрёшкой** или просто **рядом**) длины n , если имеют место вложения:

$$\{e\} = G_0 \triangleleft G_1 \triangleleft \cdots \triangleleft G_{n-1} \triangleleft G_n = G,$$

где G_i — собственная нормальная подгруппа в G_{i+1} . Ряд называется **нормальным**, если все G_i нормальны также в исходной группе G . Факторгруппы G_{i+1}/G_i называются **факторами** (факторгруппами) **ряда**.

Для простых групп (например, \mathbb{Z}_p) тривиальный субнормальный ряд длины 1 является единственно возможным: $\{e\} \triangleleft G$.

Для группы S_4 имеем:

$$\{e\} \triangleleft V_4 \triangleleft A_4 \triangleleft S_4, \quad \{e\} \triangleleft V_4 \triangleleft S_4$$

Эти утверждения можно извлечь непосредственно из таблицы В.4. Например, нормальность V_4 в A_4 следует из того, что симметричные столбец и строка в зеленой области напротив и под группой V_4 совпадают с точностью до перестановки элементов (т. е. выполняется условие $gH = Hg$).

Если для группы G существует такой субнормальный ряд, что все его факторы — абелевы группы, то группа G называется **разрешимой**.

Так как

- а) $S_4/A_4 \cong \mathbb{Z}_2$, т. е. является циклической и, тем более, абелевой группой,
- б) $A_4/V_4 \cong \mathbb{Z}_3$, т. е. является циклической и, тем более, абелевой,
- с) $V_4/\{e\}$ — абелева группа (см. таблицу В.4),

то S_4 разрешима. Заметим, что ряд $\{e\} \triangleleft V_4 \triangleleft S_4$ не годится для установления разрешимости, поскольку фактор $S_4/V_4 \cong S_3$ не является абелевой группой.

Для $n = 3$ имеем $\{e\} \triangleleft A_3 \cong \mathbb{Z}_3 \triangleleft S_3$ и, таким образом, S_3 также разрешима. Тем более разрешима и $S_2 \cong \mathbb{Z}_2$.

Известно, что все S_n порядка $n \geq 5$ неразрешимы. Именно на этом замечательном факте построено доказательство знаменитой теоремы Галуа о неразрешимости в радикалах уравнений степени 5 и выше.

Для более плотного знакомства с устройством симметрической группы S_4 рекомендуем обратиться к [ресурсы](https://en.wikiversity.org/wiki/Symmetric_group_S4) (https://en.wikiversity.org/wiki/Symmetric_group_S4).

Итак, мы видим, что Алгебру и теорию множеств снова связывает некая глубинная связь, на этот раз мы увидели в биекциях богатую алгебраическую структуру, очень похожую на числа. При этом, группами биекций и их подгруппами мы закрываем вообще все возможные группы (с точностью до изоморфизма). Получается, что если рассматривать группы вида $S(\aleph_\alpha)$, то их можно считать некими универсумами групп,⁶ по аналогии с универсумами множеств и мультимножеств. Отличие от привычных универсумов здесь состоит в том, что нижестоящие (по шкале ординалов) универсумы не вкладываются в вышестоящие как подмножества. Но, на самом деле, и это поправимо. Вместо $S(X)$ рассмотрим «усеченную» версию:

$$\bar{S}(X) = \{f \setminus \text{id}_X \mid f \in S(X)\},$$

где $\text{id}_X = \{(x, x) \mid x \in X\}$ — тождественная функция на X . Иначе говоря, $\bar{S}(X)$ содержит в себе все те же биекции, что и $S(X)$, но с «выброшенной» тождественной частью (там, где $f(x) = x$). Это в точности соответствует нашим соглашениям о том, что конечную перестановку мы записываем как произведение только нетривиальных циклов: $(1234)(5) = (1234) = (1234)(6)(7)$, etc.

При этом композиция $f \circ g$ для $f, g \in \bar{S}(X)$ определяется следующей формулой:

$$f \circ g = \{(x, z) \mid (\exists y \, z = f(y) \wedge y = g(x)) \vee (x \notin \text{dom}(g) \wedge z = f(x)) \vee (z \notin \text{dom}(f) \wedge z = g(x))\} \setminus \text{id}_{\text{dom}(g)},$$

т. е. композиция $f \circ g$ в точке x вычисляется как $g(x)$, если f не определено в точке $g(x)$, как $f(x)$, если g не определено в точке x , и стандартным образом, если определены $g(x)$ и $f(g(x))$, а для прочих x композиция не вычисляется вовсе. Кроме того, если в результате композиции получаются неподвижные точки, то вычитание $\text{id}_{\text{dom}(g)}$ убирает их. По сути это означает, что мы каждую усеченную биекцию доопределяем тождественной функцией на весь X и далее получаем композицию, у которой снова выкидываем тождественную часть.

Для так определенных групп биекций имеет место теоретико-множественное вложение $\bar{S}(X) \subseteq \bar{S}(Y)$, если $X \subseteq Y$, поскольку все усеченные биекции на X являются также и усеченными биекциями на Y . Тогда универсумы $\bar{S}(\aleph_\alpha)$ будут монотонны по α в смысле вложения. Тем

⁶Следует помнить о том, что свести все множества к кардиналам можно только в том случае, если мы принимаем аксиому выбора, в противном случае, по-видимому, придется рассматривать группы $S(V_\alpha)$ на универсумах множеств.

самым, мы построили шкалу «эталонных» групп, проиндексированную ординалами и построенную на кардиналах. Группы S_n конечных перестановок изоморфны эталонным группам $\bar{S}(\{0, \dots, n-1\})$.

Если мы не хотим связываться с аксиомой выбора, то можем рассмотреть группы $\bar{S}(V_\alpha)$, где V_α — универсумы множеств. Они также образуют монотонную последовательность. При этом следует помнить, что $\|V_n\| = 2 \uparrow n$, и, стало быть, $\bar{S}(V_n)$ изоморфна $S_{2 \uparrow n}$, т.е. эти группы будут пропускать многие S_n , и поэтому уже сами S_n будут изоморфно вкладываться в эталонные группы $\bar{S}(V_\alpha)$.

Упражнение 3.16. Придумайте такое определение; Наконец, можно предположить, что существует рекурсивное определение «эталонных» групп, аналогичное определению универсумов множеств и мультимножеств, не зависящее от аксиомы выбора и покрывающее в смысле изоморфизмов групп все возможные группы (как мы уже знаем, для этого достаточно накрыть все группы биекций).

3.1.5 Идеалы, модули, базис

Итак, как мы уже знаем, кольцо — это некое расширение группы путем добавления второй операции, согласованной с первой через закон дистрибутивности. При этом групповая операция у нас переходит в статус операции сложения (с соответствующей атрибутикой: единица становится нулем, а обратный элемент противоположным), а новая операция нарекается умножением со своей единицей и обратными элементами. Известными нам примерами колец являются $\mathbb{Z}, \mathbb{Z}_n, \mathbb{Q}, \mathbb{R}, \mathbb{C}, \text{No}_\omega, \text{No}_{\varepsilon_0}$, среди которых $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \text{No}_\omega, \text{No}_{\varepsilon_0}$ являются упорядоченными кольцами, а $\mathbb{Z}_p, \mathbb{Q}, \mathbb{R}, \mathbb{C}, \text{No}_{\varepsilon_0}$ — полями.

Важное место в теории колец занимает понятие идеала кольца. Это — прямой аналог нормальной подгруппы. Дадим определение: подмножество $I \subseteq K$ кольца K называется левым (правым) **идеалом** этого кольца, если (i) $a \pm b \in I$ при $a, b \in I$, (ii) $ka \in I$ ($ak \in I$) при $a \in I, k \in K$. Иначе говоря, идеал замкнут по сложению, а по умножению схлопывает кольцо в себя. Если идеал одновременно является левым и правым (например, в коммутативном кольце), то он называется двусторонним или просто идеалом. Идеал является подкольцом.

Идеал может быть порожден одним или несколькими элементами кольца путем составления всевозможных линейных комбинаций из этих элементов и коэффициентов из кольца. Так, если для некоторого $a \in K$ множество $I = \{ak \mid k \in K\} = aK$ ⁷ является идеалом, то он называется главным правым идеалом (порожденным элементом a), а если $I = Ka$, то главным левым

⁷Следует отметить, что такое представление годится только для колец с единицей. Если в K нет единицы, то порожденный из a главный правый идеал имеет вид $aK + \mathbb{Z}a$, где под операцией na подразумевается сумма n слагаемых, равных a (или $-a$, если $n < 0$).

идеалом (порожденным элементом a). В коммутативном кольце $Ra = aR$, и в этом случае идеал I называется просто **главным идеалом** (порожденным элементом a) и обозначается $I = (a)$.

Сравните с циклической подгруппой $\langle a \rangle$.

Простейшим примером идеала является подкольцо $n\mathbb{Z}$ кольца \mathbb{Z} . Оно же является главным идеалом, порожденным числом n , т. е. $n\mathbb{Z} = (n)$.

Забегая вперед, отметим также, что в **евклидовом кольце** (т. е. таком, где возможен алгоритм Евклида деления с остатком (см. процедуру (2.4), описанную на стр. 114), причем остаток имеет меньшую евклидову норму,⁸ чем делитель) все идеалы главные. Действительно, если в I есть элемент с нулевой нормой, то это ноль (и тогда $I = (0)$), иначе выберем внутри идеала I какое-нибудь число a с минимальной евклидовой нормой. Тогда для произвольного числа $b \in I$, используя деление с остатком, получаем, что $b = ak + r$, и норма r меньше нормы a (чего не может быть, т. к. $r = b - ak \in I$). Значит, $r = 0$. Но тогда b делится на a . А это означает, что $I \subseteq (a)$. С другой стороны, легко видеть, что $(a) \subseteq I$, т. е. $I = (a)$.



Евклид

Пример евклидова кольца — это снова кольцо целых чисел. В дальнейшем мы рассмотрим еще один замечательный пример евклидовых колец: *кольцо многочленов*.

В каждом кольце существует два тривиальных идеала — все кольцо и множество $\{0\}$. Если кольцо является телом или полем, то в нем нет нетривиальных идеалов. Вообще, всякое кольцо, в котором нет нетривиальных двусторонних идеалов, называется **простым**. \mathbb{Z} не является простым кольцом, в то же время кольцо \mathbb{Z}_p является простым при простом p (и является полем). Если в кольце нет нетривиальных идеалов (не обязательно двухсторонних), то оно является целостным кольцом, т. е. в нем нет делителей нуля. Пример тела (и одновременно целостного кольца), не являющегося полем, доставляют кватернионы. Их мы еще разберем более подробно (см. раздел 3.6.2).

Тело — это поле минус коммутативность умножения

Ясно, что поскольку кольцо является абелевой группой по сложению, то любая аддитивная подгруппа кольца, в том числе идеал, является нормальной аддитивной подгруппой. Поэтому по идеалу I можно факторизовать кольцо K точно так же, как мы факторизуем группу по нормальной подгруппе. В итоге получается новое кольцо, обозначаемое K/I и называемое **фактор-кольцом**.

Таким образом, $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$ является не только фактор-группой, но и фактор-кольцом.

Упражнение 3.17. Дайте определение фактор-кольца и умножения в нем!

⁸ **Евклидова норма** — это функция на кольце $N : K \rightarrow \omega$ такая, что для любых $a, b \in K, b \neq 0$ имеет место равенство $a = bq + r$, где $N(r) < N(b)$, причем минимум норма достигает на нулевом элементе, и только на нем. Обычно считается $N(0) = 0$, но бывают исключения.

Если $f : K \rightarrow E$ — гомоморфизм колец (т. е. сохраняет обе операции), то прообраз $f^{-1}(0)$ является двусторонним идеалом. Для любого идеала $I \subseteq K$ существует естественный гомоморфизм $f : K \rightarrow K/I$, который каждому элементу k кольца ставит в соответствие его класс эквивалентности.

Идеал I называется **простым**, если из $a \notin I, b \notin I$ следует, что и $ab \notin I$. Собственный идеал называется **максимальным**, если он не содержится ни в каком другом собственном идеале. В частности, все кольцо не может быть максимальным идеалом, хотя и является идеалом.

Известная следующая

Теорема 3.3. Пусть I — идеал коммутативного кольца K с единицей. Тогда:

(1) I — простой идеал $\Leftrightarrow K/I$ — целостное кольцо;

(2) I — максимальный идеал $\Leftrightarrow K/I$ — поле.

В частности, если идеал максимальный, то он и простой.

Из арифметики целых чисел мы также можем показать, что верна

Теорема 3.4. Пусть $0 \neq p \in \mathbb{Z}$. Тогда следующие утверждения равносильны:

(1) p — простое число;

(1) (p) — простой идеал в \mathbb{Z} ;

(2) (p) — максимальный идеал в \mathbb{Z} .

В частности, отсюда мы видим, что $\mathbb{Z}_p = \mathbb{Z}/(p)$ является полем при простом p .

Если не считать комплексные числа. До сих пор мы имели дело только со структурами, в которых операции вводились на базовом множестве (и его подмножествах), но при определении алгебраической структуры мы предполагали разрешать операции, определенные как на элементах базового множества, так и на некоторых «вспомогательных» числах — элементах кольца, поля и т.п.

Пусть далее $(K, +, \cdot)$ — коммутативное кольцо с 1 или поле, а $(M, +)$ — аддитивная абелева группа. И пусть определена совместная операция $*$ над элементами этих кольца и группы: $*$: $K \times M \rightarrow M$, т. е. $k * m \in M$, где $k \in K$ и $m \in M$. Если данная операция подчиняется аксиомам

MOD1: $1 * m = m$,

MOD2: $(k_1 + k_2) * m = k_1 * m + k_2 * m$ (дистрибутивность по $+$ в K),

MOD3: $(k_1 k_2) m = k_1 * (k_2 * m)$ (дистрибутивность по \cdot в K),

MOD4: $k * (m_1 + m_2) = k * m_1 + k * m_2$ (дистрибутивность по $+$ в M),

то структура $(M, +, \{*_k\})$ ⁹ называется **модулем над кольцом K** . Если элементы K мы воспринимаем как числа, то элементы множества M удобно интерпретировать как векторы.

Следующий шаг — добавить операцию умножения на элементах M , т. е. рассмотреть структуру $(M, +, \cdot, \{*_k\})$. Пусть при этом $(M, +, \cdot)$ — коммутативное кольцо с единицей (обозначаемой e). Добавим к предыдущим аксиомам еще один закон дистрибутивности:

MOD5: $k * (m_1 m_2) = (k * m_1) * m_2 = m_1 (k * m_2)$ (дистрибутивность по \cdot в M).

Тогда структура $(M, +, \cdot, \{*_k\})$ называется **алгеброй над кольцом K** . Схематично данные два определения представлены на рис. 3.1.

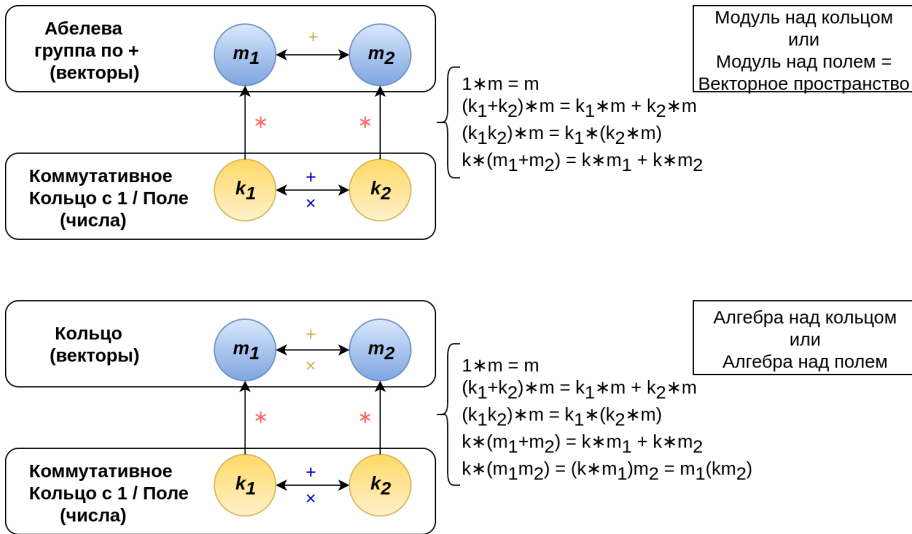


Рис. 3.1: Алгебра над кольцом.

Заметим, что мы не стали вводить различные обозначения для сложения и умножения в кольце K и в базовом множестве M . Обычно из контекста (т. е. по природе стоящих около этих операций термов) ясно, что за операция имеется ввиду. В целях экономии места мы не станем здесь увлекаться трудностями синтаксического разбора формул.

Понятие модуля тесно связано с понятием идеала кольца. Действительно, само кольцо можно рассматривать как модуль над собой, и тогда всякий (левый/правый) подмодуль (т. е. аддитивная подгруппа модуля, замкнутая относительно умножения на элементы базового кольца) такого модуля является (левым/правым) идеалом данного кольца.

⁹Напомним, что K у нас выступает как параметризующее множество для операции $*$, т. е. на самом деле мы рассматриваем много унарных операций вида $*_k : M \rightarrow M$, заданных для каждого $k \in K$.

Если $f : M \rightarrow M'$ — гомоморфизм модулей, то ядро $f^{-1}(0)$ является подмодулем модуля M . И обратно, всякий подмодуль является ядром некоторого гомоморфизма модулей.

В модуле и алгебре, как и в группе, можно ввести понятие базиса. Пусть подмножество $S \subseteq M$ таково, что все конечные линейные комбинации $k_1 e_1 + \dots + k_n e_n$ ($k_1, \dots, k_n \in K$, $e_1, \dots, e_n \in S$) порождают весь модуль M и при этом S линейно независимо (т. е. никакая нетривиальная конечная линейная комбинация элементов S не равна 0). Тогда множество S называется **базисом (Гамеля) модуля M** ,¹⁰ а его мощность — размерностью этого базиса. В некоторых случаях¹¹ может быть ситуация, что в модуле M существуют базисы различной размерности (n и m) или базисов не существует вовсе. Модуль называется **свободным**, если он либо нулевой, либо обладает базисом.

Стоит отметить, что размерность модуля вводится на основе операции сложения и умножения на коэффициенты из базового кольца. В то же время, если модуль является алгеброй, т. е. в нем задано умножение векторов, то можно говорить о базисе на основе операции умножения, и в общем случае размерность такого базиса может отличаться от размерности базиса Гамеля, определенного выше.

В которых кроется дьявол. Не углубляясь в детали, отметим, что поскольку мы сразу же предположили, что кольцо K коммутативно, то размерность базиса модуля (при его наличии) постоянна (не зависит от выбора базиса). Докажем это с помощью следующих утверждений, типичных для линейной Алгебры (чем они и ценны в данной книге).

Лемма 3.1. *В коммутативном кольце линейная система уравнений*

$$\begin{cases} a_{11}x_1 + \dots + a_{1n}x_n = 0 \\ \dots \\ a_{m1}x_1 + \dots + a_{mn}x_n = 0 \end{cases}$$

имеет нетривиальное решение x_1^, \dots, x_n^* , если $m < n$.*

Доказательство. Проведем доказательство индукцией по n . Для $n = 2$ имеем единственный вариант системы: $a_{11}x_1 + a_{12}x_2 = 0$.

Случай 1. $a_{11} = a_{12} = 0$. Тогда решение $x_1^* = 1, x_2^* = 0$ является нетривиальным решением (как и вообще любое другое).

¹⁰Мы снова апеллируем к архетипу базового множества, отождествляя M со структурой $(M, +, \{*_k\})$.

¹¹Например, когда над кольцом K существуют прямоугольные матрицы $A \in K^{n \times m}$ и $B \in K^{m \times n}$ такие, что AB и BA есть единичные матрицы рангов n и m , соответственно

Теорема 3.5. Если в модуле над коммутативным кольцом существует конечный базис, то все базисы имеют такую же размерность.

Доказательство. Пусть e_1, \dots, e_m и e'_1, \dots, e'_n — базисы и $m < n$. Тогда в силу определения базиса имеем:

$$\begin{aligned} e'_1 &= a_{11}e_1 + \dots + a_{m1}e_m \\ &\dots\dots\dots \\ e'_n &= a_{1n}e_1 + \dots + a_{mn}e_m \end{aligned}$$

и рассмотрим линейное уравнение $x_1e'_1 + \dots + x_ne'_n = 0$. Если e'_1, \dots, e'_n — базис, то это уравнение может иметь только тривиальное решение $x_1 = \dots = x_n = 0$. Подставим сюда разложения базисных векторов и получим:

$$\begin{aligned} 0 = & (a_{11}x_1 + \dots + a_{1n}x_n)e_1 + \\ & \dots\dots\dots \\ & (a_{m1}x_1 + \dots + a_{mn}x_n)e_m \end{aligned}$$

откуда в силу того, что e_1, \dots, e_m — базис, все коэффициенты должны быть равны нулю, т. е.

$$\begin{cases} a_{11}x_1 + \dots + a_{1n}x_n &= 0 \\ \dots\dots\dots \\ a_{m1}x_1 + \dots + a_{mn}x_n &= 0 \end{cases}$$

Но в силу предыдущей леммы эта система имеет нетривиальное решение x_1^*, \dots, x_n^* , а это означает, что и уравнение $x_1e'_1 + \dots + x_ne'_n = 0$ имеет нетривиальное решение. Следовательно, e'_1, \dots, e'_n не может быть базисом при $n > m$.

Таким образом, все конечные базисы модуля (над коммутативным кольцом) равномощны, если они существуют. \square

Комментарий 7.

Аналогичное утверждение справедливо и для бесконечных базисов Гамеля в модуле M над кольцом, но только с применением аксиомы выбора. Действительно, пусть у нас есть базисы $S \subset M$ и $S' \subset M$, тогда каждый элемент базиса S можно разложить по конечному набору элементов базиса S' . Имеем функцию: $f : S \rightarrow \mathcal{P}(S')$, где $\|f(s)\| < \omega$. Эта функция определяется единственным образом (если бы s раскладывался по-разному, мы получили бы нетривиальную линейную комбинацию базисных векторов). Тогда можно составить множество $S'' = \cup\{f(s) \mid s \in S\}$. Ясно, что $S'' \subseteq S'$, с другой стороны, $S' \subseteq S''$, т. к. в противном случае, взяв элемент $s' \in S' \setminus S''$, мы бы разложили его по элементам из S'' и, тем самым, построили бы нетривиальную линейную комбинацию элементов базиса S' . Следовательно, $S' = S''$. Остается показать, что S и S' равномощны. Для этого рассмотрим дизъюнктивную сумму $\Sigma \equiv \cup\{s \times f(s)\}$,

которая инъективно вкладывается в квадрат $S \times S$, т. к. S бесконечно (здесь уже требуется счетная форма аксиомы выбора). Но S'' инъективно вкладывается в Σ (для построения инъекции нужна аксиома выбора), а значит, и в квадрат $S \times S$. Наконец, поскольку $S \times S$ равносильно S (теорема о квадрате, требующая аксиомы выбора), получаем инъективное вложение S' в S . Рассуждая симметричным образом, доказываем, что имеет место и обратное вложение S в S' . Теперь, по тереме Кантора—Бернштейна—Шредера 1.22 получаем равносильность S и S' .

Данное рассуждение нельзя применить для случая конечного базиса, поэтому оно не влечет предыдущую теорему. Требование аксиомы выбора можно немного ослабить, если пользоваться выбором только для мощностей не выше мощности базиса. Таким образом, если в модуле M имеется один счетный базис Гамеля (как в алгебре многочленов) и если мы принимаем счетную аксиому выбора, то любой базис Гамеля в модуле M будет счетным.

Поэтому корректно определение: размерностью модуля (алгебры) над K называется размерность его базиса. Соответственно, если размерность модуля (алгебры) конечная, то он называется конечномерным модулем (алгеброй). Работа с бесконечномерными модулями без определения хотя бы счетных линейных комбинаций (что требует дополнительной структуры вроде метрики или топологии) не представляется возможной, а работа с конечными базисами Гамеля в случае бесконечномерных пространств дает мало полезной информации. Поэтому в «чисто алгебраическом» виде, как правило, рассматриваются конечномерные модули и алгебры.

Рассмотрим несколько примеров реализации понятия модуля/алгебры над кольцом.

Пример 1.

Любая абелева группа — модуль над кольцом целых чисел. При этом базис группы будет и базисом модуля. | Упражнение 3.18.

Пример 2.

Если положить $M = K$, то все аксиомы MOD1–MOD5 выполняются автоматически в силу аксиом коммутативного кольца с единицей, т. е. K является алгеброй над самим собой с размерностью 1.

Пример 3.

\mathbb{C} — алгебра над \mathbb{R} . Размерность \mathbb{C} , как алгебры над \mathbb{R} , равна 2.

Пример 4.

Поле \mathbb{R} является континуум-мерным векторным пространством над полем \mathbb{Q} .

Пример 5. Векторное пространство.

Ранее мы уже определили понятие кортежа или вектора (см. стр. 91) как функции, заданной на натуральном числе, а точнее — на конечном ординале $n = \{0, 1, \dots, n-1\}$. Пусть элементами вектора (значениями функции) являются только числа из кольца K . Тогда, полагая, что K — поле, $M = K^n$, а также определяя покомпонентно сумму векторов с помощью суммы в K , а умножение вектора на скаляр как покомпонентное умножение на этот скаляр, мы получаем не что иное как модуль над полем K или **векторное пространство над полем K** . В частности \mathbb{R}^n — векторное пространство размерности n над полем действительных чисел.

Пример 6. Матрицы.

Теперь вместо множества $n = \{0, \dots, n-1\}$ возьмем его квадрат: $n \times n$ и определим на нем функцию f со значениями в кольце K . Такую функцию называют **квадратной матрицей** над K и записывают в виде:

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix}$$

где $a_{ij} \Leftarrow f(i-1, j-1)$.

Замечание 1. Не обязательно ограничиваться квадратными матрицами, можно определить f и на прямоугольнике $n \times m$.

Замечание 2. Матрица — это элемент степени $K^{n \times m}$, но ни-

что не мешает нам сначала взять векторы длины n , а затем из них соорудить еще более сложные векторы (векторы из векторов), рассматривая следующую башню степеней: $(K^n)^m$. Как множества, это будут разные объекты, но установить между ними естественный изоморфизм не составит никакого труда, поэтому матрицы можно считать векторами из векторов.

Замечание 3. Можно не ограничиваться натуральными числами, а перейти к ординалам. В частности, K^ω дает нам множество всех последовательностей над полем K , а $K^{\omega \times \omega}$ — бесконечные матрицы.

Замечание 4. Можно также не ограничиваться числом измерений матрицы, рассматривая в общем виде множество функций $K^{\alpha_1 \times \dots \times \alpha_k}$, где α_i — какие-то ненулевые ординалы. Если все они — конечные, то такие многомерные матрицы можно рассматривать как представление **тензоров**.

Привязка к нулю все-таки порой избыточна...

Отсылая читателя к любой книжке по линейной алгебре, приведем без доказательства тот факт, что квадратные матрицы образуют алгебру над кольцом, т. е. их можно складывать и перемножать друг с другом, а также умножать на число из базового кольца (из которого взяты элементы матриц).

Упражнение 3.19. Проверьте это.

Отметим одно интересное свойство матриц. Будем обозначать через $M_k^{n \times t}$ матрицу, в которой n строк и t столбцов. По определению произведения матриц [58] (скалярное произведение соответствующих строки и столбца) для существования произведения матриц $M_1^{n,m} \times M_2^{k,j}$ необходимо и достаточно, чтобы было $m = k$. Представим теперь, что мы составили матрицы не из чисел какого-нибудь кольца, а тоже из матриц, и попробуем перемножить их:

$$\begin{pmatrix} M_{11}^{n_1 \times m_1} & M_{12}^{n_1 \times m_2} & M_{13}^{n_1 \times m_3} \\ M_{21}^{n_2 \times m_1} & M_{22}^{n_2 \times m_2} & M_{23}^{n_2 \times m_3} \end{pmatrix} \begin{pmatrix} \tilde{M}_{11}^{m_1 \times k_1} & \tilde{M}_{12}^{m_1 \times k_2} \\ \tilde{M}_{21}^{m_2 \times k_1} & \tilde{M}_{22}^{m_2 \times k_2} \\ \tilde{M}_{31}^{m_3 \times k_1} & \tilde{M}_{32}^{m_3 \times k_2} \end{pmatrix} = \begin{pmatrix} \hat{M}_{11}^{n_1 \times k_1} & \hat{M}_{12}^{n_1 \times k_2} & \hat{M}_{13}^{n_1 \times k_3} \\ \hat{M}_{21}^{n_2 \times k_1} & \hat{M}_{22}^{n_2 \times k_2} & \hat{M}_{23}^{n_2 \times k_3} \\ \hat{M}_{31}^{n_3 \times k_1} & \hat{M}_{32}^{n_3 \times k_2} & \hat{M}_{33}^{n_3 \times k_3} \end{pmatrix}$$

где $n_1 + n_2 = n$, $m_1 + m_2 + m_3 = m$, $k_1 + k_2 + k_3 = k$, причем результирующие матрицы вычисляются по правилу:

$$\hat{M}_{ij}^{n_i \times k_i} = M_{i1}^{n_i \times m_1} \tilde{M}_{1j}^{m_1 \times k_j} + M_{i2}^{n_i \times m_2} \tilde{M}_{2j}^{m_2 \times k_j} + M_{i3}^{n_i \times m_3} \tilde{M}_{3j}^{m_3 \times k_j}.$$

Такое *блочное умножение матриц* позволяет свести умножение больших матриц $n \times t$ на $t \times k$ к умножению блоков более низкой размерности. В результате мы все равно получим правильную матрицу размерности $n \times k$. Что это дает? Во-первых, эстетическое наслаждение — ведь можно очень сильно компактифицировать работу с большими матрицами (например, 8×8 представлять как 2×2 с блоками 4×4), если у них есть какое-то естественное блочное представление. Во-вторых, если у вас есть какой-то достаточно шустрый алгоритм умножения матриц низкой размерности (а такие есть в разных языках программирования), то для перемножения больших матриц выгоднее разбить их на небольшие блоки и воспользоваться блочным умножением (например, алгоритмами Фокса или Кэннона с использованием параллельных вычислений, см., например, [43, 49]).

Больше того, процесс дробления может быть многоступенчатым: ничто не мешает блоки подразбивать на еще более мелкие подблоки, и т.д. Здесь мы сталкиваемся с таким архетипом математики, который можно назвать **архетипом редукции**. С этим понятием мы уже сталкивались, когда редуцировали записи «начальных» множеств, переходя к записям их компонентов. Это примерно то же самое, что разбить матрицу на блоки. В целом же, архетип редуцирования предполагает сведение исходной задачи (или объекта, формулы) к более простой, ранее хорошо изученной.

Старый анекдот про «выльем воду из чайника»...

Само по себе множество квадратных матриц порядка n со стандартными операциями сложения и умножения образуют (некоммутативное!) кольцо с единицей, размерность которого равна n^2 относительно исходного кольца (и относительно операции сложения матриц). Кольцо квадратных матриц не является телом, т. к. содержит *делители нуля*. Последнее свойство нам особенно интересно, т. к. подчеркивает существенное отличие алгебры квадратных матриц от обычных числовых систем.

Упражнение
3.20.
Проверьте!

Приведем простой пример:

$$\begin{pmatrix} 1 & x \\ y & xy \end{pmatrix} \begin{pmatrix} 1 & y' \\ x' & x'y' \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad (3.4)$$

если и только если $xx' = -1$. С точностью до линейного коэффициента мы привели все возможные нетривиальные примеры пар матриц 2-го порядка, произведение которых дает нулевую матрицу. Таким образом, алгебра матриц является примером кольца с делителями нуля.

Пример 7. Многочлены.

Особое место в Алгебре занимают многочлены. В смысле стандартного анализа это функции вида $f(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$, где переменная x и коэффициенты a_i суть числа из какого-либо кольца или поля. Многочлен степени n однозначно задается набором коэффициентов, т. е. вектором $(a_0, \dots, a_n, 0, \dots) \in K^\omega$ с условием, что начиная с некоторого номера все компоненты равны нулю (т. е. функция, определяющая данный вектор, имеет конечный носитель). Напомним, что примерно так же мы определяли множество \mathbb{R} , только элементы вектора брали из \mathbb{Z}_2 и допускали бесконечный набор единиц, при этом вместо x можно было подставить 2 и получить сходящийся к действительному числу ряд. В этом смысле между числами и многочленами много общего.

Упражнение
3.21. | Многочлены образуют счетно-мерную алгебру над кольцом K , и одновременно — **кольцо многочленов**.

Если все-таки рассматривать все возможные векторы из \mathbb{R}^ω , т. е. включая те, которые содержат бесконечно много ненулевых элементов, то можно говорить о кольце формальных степенных рядов вида $\sum_{k=0}^{\infty} a_k x^k$, в котором операции определяются аналогично операциям над многочленами:

$$\sum_{k=0}^{\infty} a_k x^k + \sum_{k=0}^{\infty} b_k x^k = \sum_{k=0}^{\infty} (a_k + b_k) x^k, \quad \sum_{k=0}^{\infty} a_k x^k \cdot \sum_{k=0}^{\infty} b_k x^k = \sum_{k=0}^{\infty} \left(\sum_{i+j=k} a_i a_j \right) x^k.$$

Если мы теперь вспомним про ряд Лорана, то придется рассмотреть базовое множество $\mathbb{C}^{\mathbb{Z}}$, т. е. все функции из \mathbb{Z} со значениями в \mathbb{C} . Эти функции будут определять коэффициенты при целых степенях переменной z .

Пример 8. Групповое кольцо.

Пусть у нас есть некая группа (G, \cdot) и кольцо $(K, +, \cdot)$. Пусть

$$M \rightleftharpoons \{f \mid (f : G \rightarrow K) \wedge (\|\text{supp}(f)\| < \omega)\},$$

т. е. все функции из группы в кольцо, которые отличны от нуля (кольца) лишь на конечном множестве элементов группы. Определим операции в M следующим образом:

$$(f + g)(x) = f(x) + g(x); \quad (fg)(x) = \sum_{uv=x} f(u)g(v)$$

Предлагаем читателю самостоятельно проверить, что такие операции на M задают структуру кольца, причем если K и G коммутативны, то M коммутативно; если K — кольцо с единицей, то M — кольцо с единицей.

Добавим теперь операцию умножения элементов M на число из K по формуле: $(k \cdot f)(x) = kf(x)$, пользуясь умножением в кольце K . Легко проверить, что эта операция «умножения на скаляр» согласована с операциями в кольцах K и M . Таким образом, мы имеем алгебру над кольцом K . Эта алгебра обозначается $K[G]$ и называется **групповым кольцом**.

Группа G вкладывается в $K[G]$ естественным образом с помощью функций-индикаторов:

$$\tilde{G} \rightleftharpoons \{1_g \mid g \in G\},$$

где $1_g(x) = 1_K$ если и только если $x = g$, в противном случае 0. При этом любой элемент $K[G]$ является линейной комбинацией этих индикаторов:

$$f = \sum_{g \in G} 1_g \cdot f(g), \quad (3.5)$$

откуда следует, что все элементы $K[G]$ порождаются (составляют линейную оболочку) множеством таких индикаторов. Кроме того, все индикаторы линейно независимы в $K[G]$. Иначе говоря, вложение G в $K[G]$, состоящее из индикаторов, является базисом алгебры $K[G]$.

Кроме того, множество \tilde{G} является образом группы G в кольце $K[G]$ относительно естественного изоморфизма, при котором элементу $g \in G$ сопоставляется элемент $1_g \in K[G]$, причем $1_g \cdot 1_h = 1_{gh}$ (т. е. изоморфизм мультипликативный).

Формула (3.5) позволяет рассматривать элементы $K[G]$ как линейные комбинации самих элементов G с коэффициентами из K и соответствующим образом производить все арифметические действия над ними. В частности, если группа G — циклическая, то все ее элементы можно записать в виде

$g, g^{-1}, g^2, g^{-2}, g^3, g^{-3}, \dots$, и тогда элементы $K[G]$ будут очень похожи на запись ряда Лорана с тем ограничением, что сумма всегда будет конечной:

$$f = \sum_{n \in S} k_n g^n,$$

где S — конечное подмножество \mathbb{Z} . При этом сложение и умножение таких «рядов» производится точно так же, как сложение и умножение многочленов (с поправкой на отрицательные степени). В этой связи групповое кольцо $K[\mathbb{Z}]$ даже носит название **полиномы Лорана**.

В случае конечной циклической группы использование отрицательных степеней не требуется, поэтому аналогия с алгеброй многочленов еще больше усиливается, но при этом нужно помнить, что сложение степеней у порождающего элемента g будет происходить по модулю некоторого натурального числа. Например, пусть $G = \{e, g, g^2\} = \langle g \rangle$, тогда

$$\mathbb{C}[G] = \{k_0 + k_1g + k_2g^2 \mid k_0, k_1, k_2 \in \mathbb{C}\}.$$

Сложение:

$$k_0 + k_1g + k_2g^2 + k'_0 + k'_1g + k'_2g^2 = (k_0 + k'_0) + (k_1 + k'_1)g + (k_2 + k'_2)g^2.$$

Умножение:

$$\begin{aligned} (k_0 + k_1g + k_2g^2)(k'_0 + k'_1g + k'_2g^2) = \\ = (k_0k'_0 + k_1k'_2 + k_2k'_1) + (k_0k'_1 + k_1k'_0 + k_2k'_2)g + (k_0k'_2 + k_2k'_0 + k_1k'_1)g^2, \end{aligned}$$

поскольку $gg^2 = g^3 \bmod 3 = e$, $g^2g^2 = g^4 \bmod 3 = g$.

Наконец, если мы в качестве группы возьмем группу корней из 1 в поле \mathbb{C} , то на ее основе также можно построить групповое кольцо. Например, группа корней 4 степени $G_2 = \{1, i, -1, -i\}$ порождает групповое кольцо

$$\mathbb{R}[G_2] = \{x_1 + ix_2 + (-1)x_3 + (-i)x_4\},$$

однако арифметика такого кольца никак не связывает между собой 1 и -1 , i и $-i$, поэтому перед нами — четырехмерное вещественное пространство, а не комплексная плоскость.

Пара общих слов

Как видим, понятие модуля и алгебры над кольцом (полем) обладает весьма широким спектром примеров, позволяя строить новые алгебраические структуры, оставаясь при этом в связке со старыми, хорошо известными числовыми структурами. Так, вводя в свой обиход векторы, мы все равно хотим умножать их на число и получать какие-то их числовые характеристики (например, норму или скалярное произведение), вводя матрицы, мы также умножаем их на число и считаем определитель, а многочлены так и вовсе являются прямыми арифметическими продуктами обычных чисел.

При этом мы конструируем некое двухэтажное здание, каждый этаж которого является сам по себе алгебраической структурой, а между этажами действуют связывающие их операции-лифты. Кроме того, сам по себе этот шаг является инструментом к неограниченному возведению этажей. Действительно, взяв поле действительных чисел, мы строим над ним кольцо многочленов, после чего над этим кольцом многочленов можем выстроить кольцо операторов, а затем еще рассмотреть над ним какие-нибудь полезные группы.

Эти алгебраические методы строительства объектов математики переключаются с теоретико-множественными, о которых мы говорили ранее: переходу от множества к системе подмножеств, либо к прямому произведению множеств (такт 1), а затем к выделению некоторого подмножества с определенными свойствами (такт 2).

Как алгебраические, так и теоретико-множественные способы строительства используются в таких интересных разделах математики, как топология, теория вероятностей и многих других. По сути, в математике невозможно выделить отдельные науки — настолько она едина. Можно выделить направления и методы, но в любой момент вы можете взять и придти с инструментами тополога, например, в геометрию и поставить ряд неожиданных вопросов. В этом и состоит то архетипическое единство математики, которое мы здесь пытаемся показать читателю.

Чем мячик от бублика отличается? Что общего у сферы и тетраэдра? Где живут руконожки?

Примечательно, однако, то, что в реальной математической практике такое бесконечное алгоритмизованное навёрчивание подмножеств и отношений, групп и колец друг над другом в автоматическом режиме не дает практически ничего интересного. Чаще получается так, что как только математика придумала некий машинный путь строительства своих объектов, так сразу она об этом забывает и уходит в поиск новых, ранее невиданных конструкций. Такое свойство постоянного ухода от машинного пути развития, наверное, тоже стоило бы отнести к **архетипам математики**, но это больше присуще психологии самих математиков, чем математическому знанию как таковому.

Но вернемся к алгебре над кольцом. Как мы видим, в ней задействовано целых 5 операций: 2 операции в кольце, 2 — в самой алгебре, и одна связующая операция умножения на скаляр. Мы специально не стали вводить разные обозначения для пар операций, действующих внутри каждого этажа, поскольку зачастую они очень тесно связаны (нижние порождают верхние), и только межэтажную операцию выделили звездочкой.

Если теперь окинуть взором структуры, с которыми нам довелось столкнуться и попытаться их отклассифицировать по количеству используемых операций и отношений, т. е. по сигнатуре, то возникает небольшая таблица.

Поясним, при чем тут универсум множеств. Дело в том, что универсум не только является хорошим примером алгебры множеств и, значит, решетки, он также существенным образом (через аксиому регулярности) опирается на

Таблица 3.1: Структуры и их сложность.

	Порядки	Числа
1	ч.у.м.	группа
2	решетка	кольцо
3	ординалы	упорядоченное кольцо
4	универсум множеств	векторное пространство
5		алгебра над кольцом

отношение принадлежности, фундаментальное в теории множеств. Которое, к тому же, порождает и отношение вложения множеств, обычно выступающее в роли примера частичного порядка. В ординалах эти два отношения склеиваются в одно (из-за транзитивности принадлежности), и поэтому ординалы являют пример системы с двумя операциями и одним отношением. Универсум множеств имеет два разных отношения (\in и \subset) и две операции (объединение и дополнение, либо пересечение и дополнение). При желании его можно отнести и к строке 5, включив в систему операций одновременно \cup, \cap, \setminus (что мы и наблюдаем в аксиоматике булевой алгебры), но помня о законах де Моргана,¹² нет смысла включать одновременно эти три операции.

3.2 Матричное представление чисел

Выше мы требовали, чтобы все элементы матрицы были числами из кольца, но в ряде случаев можно обходиться и полукольцом. Рассмотрим, например, все квадратные матрицы 2×2 над натуральными числами с обычными операциями сложения и умножения матриц. И выделим среди них матрицы следующего вида:

$$M(n, m) = \begin{pmatrix} n & m \\ m & n \end{pmatrix}$$

Легко видеть, что $M(n, m) + M(n', m') = M(n + n', m + m')$, $M(n, m)M(n', m') = M(nn' + mm', nm' + n'm)$. Предполагая, что $n \geq m$, мы можем рассмотреть разность $n - m$. Эта разность следует судьбе операций над матрицами, т. е. сумма переходит в сумму разностей, а произведение — в произведение: $(n - m)(n' - m') = (nn' + mm') - (nm' + n'm)$.

¹²В логике: $\neg(\varphi \vee \psi) = \neg\varphi \wedge \neg\psi$; $\neg(\varphi \wedge \psi) = \neg\varphi \vee \neg\psi$. В множествах: $X \setminus (A \cup B) = (X \setminus A) \cap (X \setminus B)$; $X \setminus (A \cap B) = (X \setminus A) \cup (X \setminus B)$.

Положим теперь матрицы $M(n, m)$ и $M(n', m')$ эквивалентными ($M(n, m) \sim M(n', m')$), если $n + m' = n' + m$.

Нетрудно показать, что это — отношение эквивалентности.

Для каждой пары классов эквивалентности $[M(n, m)]_{\sim}$ и $[M(n', m')]_{\sim}$ можно корректно¹³ ввести операции сложения и умножения классов:

$$\begin{aligned} [M(n, m)]_{\sim} + [M(n', m')]_{\sim} &= [M(n, m) + M(n', m')]_{\sim}, \\ [M(n, m)]_{\sim} \cdot [M(n', m')]_{\sim} &= [M(n, m)M(n', m')]_{\sim} \end{aligned}$$

В итоге мы получим структуру из классов с двумя операциями над ними, изоморфную \mathbb{Z} . Каждый такой класс эквивалентности будет представлять целое число, которое можно легко вычислить — классу $[M(n, m)]_{\sim}$ соответствует число $n - m$ (теперь уже неважно, какое из них больше).

В каждом классе находится минимальный представитель вида $M(k, 0)$ или $M(0, k)$ (если $n \geq m$, то $k = n - m$ и представитель равен $M(k, 0)$, а если $n \leq m$, то $k = m - n$ и представитель равен $M(0, k)$). Таким образом, целому числу $k \geq 0$ соответствует класс $[M(k, 0)]_{\sim}$, а целому числу $-k \leq 0$ — класс $[M(0, k)]_{\sim}$.

Например, легко видеть, что $M(k, 0) + M(0, k) = M(k, k) \sim M(0, 0)$.

Конечно, модель \mathbb{Z} , основанная на двух лучах $\omega \times \{0\} \cup \{0\} \times \omega$, все равно проще, но и такой матричный подход с факторизацией имеет право на существование. Тем более что он богат своими приложениями к другим числам, и это мы увидим ниже на многочисленных примерах.

Отметим, что попытка аналогичным образом продлить арифметику бесконечных ординалов в отрицательную область упирается в некоммутативность операций над ними. Действительно, пусть имеется три ординальных матрицы

$$A = \begin{pmatrix} \alpha & \beta \\ \beta & \alpha \end{pmatrix}, \quad B = \begin{pmatrix} \gamma & \delta \\ \delta & \gamma \end{pmatrix}, \quad C = \begin{pmatrix} \chi & \sigma \\ \sigma & \chi \end{pmatrix}.$$

Зададим такое же отношение эквивалентности: $\alpha + \delta = \gamma + \beta$ — в этом случае пишем $A \sim B$. Данное отношение рефлексивно и симметрично, однако оно не будет транзитивным. Приведем конкретный пример:

$$\begin{aligned} \alpha &= \omega + n & \gamma &= \omega & \chi &= \omega^2 + n \\ \beta &= n & \delta &= 0 & \sigma &= \omega^2 + n \end{aligned}$$

где $0 < n < \omega$. При таких значениях имеем: $A \sim B$, $B \sim C$, но неверно, что $A \sim C$, поскольку $\alpha + \sigma = \omega^2 + n$, $\chi + \beta = \omega^2 + 2n$. Более того, в данном

¹³Корректность означает, что результат не зависит от выбора представителей этих классов.

Упражнение
3.23.
Докажите!

Упражнение
3.24.
Проверьте
корректность
операций.

случае C должно быть нейтральным элементом по сложению, т. к. $\chi = \sigma$, но очевидно, что $A + C = C$, т. к. высокие степени ординала ω , добавляемые справа, поглощают все низкие степени.

Поэтому такой способ продления \mathbb{Z} в *инфинитезимальную* область не является корректным.

Более симпатичным примером матричного расширения чисел является аналогичное представление комплексных чисел через действительные. Пусть $z = x + iy$ и рассмотрим матрицу

$$C(x, y) = \begin{pmatrix} x & -y \\ y & x \end{pmatrix}$$

Упражнение 3.25. Проверьте, что множество таких матриц при указанном взаимно однозначном соответствии с комплексными числами изоморфно \mathbb{C} , т. е. сложение и умножение матриц в точности соответствует сложению и умножению соответствующих комплексных чисел. Иначе говоря, алгебра таких матриц над полем \mathbb{R} изоморфна полю \mathbb{C} .

Представление чисел матрицами не только подчеркивает тесную связь понятия матрицы с понятием числа, но и открывает способ конструирования новых чисел.

Для начала возьмем матрицы вида

$$D_2(x, y) = \begin{pmatrix} x & y \\ y & x \end{pmatrix},$$

которые выглядят ровно так же, как матрицы для определения целых чисел через натуральные, только $x, y \in \mathbb{R}$. Матричная арифметика на таких числах определяет так называемые **двойные гиперкомплексные числа**. Каждое двойное число можно представить в виде $x + jy$, где $j^2 = 1$, но при этом $j \neq \pm 1$. Точнее, положим

$$1 = D_2(1, 0), \quad j = D_2(0, 1)$$

Тогда $x + jy$ соответствует линейной комбинации матриц 1 и j с коэффициентами x и y , в полной аналогии с комплексным числом $x + iy$.

Алгебра двойных чисел содержит делители нуля и поэтому, в отличие от алгебры комплексных чисел, не является полем. Все пары делителей нуля можно определить формулой: $x(1 \pm j)$, $x \in \mathbb{R}$. Отметим, что это вписывается в формулу (3.4), где достаточно взять $x = y = 1$, $x' = y' = -1$.

Подробное исследование алгебры двойных чисел и ее применений в физике, а также сравнительный анализ с полем комплексных чисел можно найти в статье [85].

Взяв матрицу $D_2(x, y)$ как образец конструкции, построим матрицу

$$D_4(z, w) = \begin{pmatrix} z & w \\ w & z \end{pmatrix},$$

где $z = D_2(x_1, y_1)$, $w = D_2(x_2, y_2)$ — двойные числа. Операции с матрицами индуцируют арифметику на эти новые числа, которые называются **четверными гиперкомплексными числами**. Полагая

$$1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad j = \begin{pmatrix} j & 0 \\ 0 & j \end{pmatrix}, \quad i = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad k = \begin{pmatrix} 0 & j \\ j & 0 \end{pmatrix},$$

где $1 = D_2(1, 0)$, $j = D_2(0, 1)$, мы получаем базисные векторы $1, i, j, k$, обладающие таблицей умножения:

1	i	j	k
i	1	k	j
j	k	1	i
k	j	i	1

Заметим, что умножение здесь коммутативно, но при этом обладает делителями нуля. Ими могут быть все возможные суммы и разности базисных векторов, умноженные на вещественный скаляр (по формуле разности квадратов): $x(1 \pm i)$, $x(1 \pm j)$, $x(1 \pm k)$, $x(i \pm j)$, $x(i \pm k)$, $x(j \pm k)$, где $x \in \mathbb{R}$.

На пространстве четверных гиперкомплексных чисел вводится специальная метрика, называемая метрикой Бервальда–Моора. Получаемое пространство является одной из перспективных моделей геометрии физического мира (подробнее см. [78]).

Двойные числа не следует путать с **дуальными числами**, определяемыми матрицей $\begin{pmatrix} x & y \\ 0 & x \end{pmatrix}$. Дуальные числа, как и двойные, образуют двумерную коммутативную алгебру, не являющуюся полем по той же причине — в ней есть делители нуля. Наличие делителей нуля позволяет интерпретировать дуальные числа как модель для гипердействительных чисел, т. е. расширение вещественных чисел с помощью бесконечно малых. Действительно, всякое дуальное число можно записать в виде $x + \varepsilon y$, где $\varepsilon^2 = 0$. В матричном виде $\varepsilon = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$. Таким образом, плоскость пар чисел (x, y) можно интерпретировать как \mathbb{R} , где каждому вещественному числу x сопоставляется еще целое \mathbb{R} -подобное семейство бесконечно малых чисел, определяемых довеском εy . В теории гипердействительных чисел такое семейство принято называть **моной** числа x .

Возьмем теперь представление комплексных чисел в виде матрицы над \mathbb{R} и подставим вместо действительных чисел комплексные:

$$Q(z, w) = \begin{pmatrix} z & -w \\ \bar{w} & \bar{z} \end{pmatrix}$$

где \bar{z}, \bar{w} — комплексно сопряженные числа. В итоге мы получим гамильтоновы четырехмерные числа, известные также как **кватернионы**. Множество всех кватернионов обозначается \mathbb{H} .

Более того, продолжая процедуру удвоения (с использованием сопряжения на второй строке матрицы), мы будем получать все новые и новые числа в пространстве размерности 2^n . Такая процедура носит название **процедуры Кэли–Диксона**. При этом свойства новых алгебр с каждым шагом становятся все менее привлекательными. Так, имея поле \mathbb{R} с тождественным сопряжением ($\bar{x} = x$), мы получаем поле \mathbb{C} с нетождественным сопряжением, затем мы приходим к телу \mathbb{H} (потеряли коммутативность) гамильтоновых чисел, затем — к некоммутативной и неассоциативной алгебре \mathbb{O} (октанионы), все еще без делителей нуля. Дальнейшие алгебры уже будут иметь делители нуля.

Перед формулировкой следующей теоремы напомним, что тело отличается от поля тем, что в нем умножение может быть некоммутативным, при этом в теле, как и в поле, отсутствуют делители нуля.

Теорема 3.6 (Фробениуса). *Пусть тело \mathbb{L} содержит \mathbb{R} как подтело, для любых $x \in \mathbb{L}, y \in \mathbb{R}$ выполняется равенство $xy = yx$ (т. е. элементы \mathbb{L} коммутируют с действительными числами) и, кроме того, \mathbb{L} является конечномерной алгеброй над \mathbb{R} .*

Тогда \mathbb{L} изоморфно либо \mathbb{R} , либо \mathbb{C} , либо \mathbb{H} .

Проще говоря, требования обратимости ненулевых элементов и конечная размерность ограничивают нас только двумя конечномерными расширениями \mathbb{R} — полем \mathbb{C} и телом \mathbb{H} .

Доказательство теоремы Фробениуса можно найти, например, в [62].

Итак, как видим, матрицы и матричное представление чисел открывают нам в мире алгебр еще одну ветку числовых систем, пусть и не всегда привлекательных своими свойствами, которые, тем не менее, укладываются в наше представление о том, что такое алгебра. Тем более удивителен тот факт, что эти неординарные числа прекрасно работают в линейной алгебре и геометрии, открывая нам различные симметрии в мире преобразований пространств.

Кроме того, отметим, что матрицами можно представлять и группы. Например, группу перестановок S_n можно изоморфно отобразить в множество матриц размерности $n \times n$ вида

$$\begin{pmatrix} 0 & 1 & \dots & 0 \\ 0 & & \dots & 1 & 0 \\ 0 & \dots & 1 & \dots & 0 \end{pmatrix}$$

где в i -ой строке единица стоит на s_i -м месте в соответствии с перестановкой

$$\begin{pmatrix} 1 & 2 & \dots & n \\ s_1 & s_2 & \dots & s_n \end{pmatrix}$$

При этом обычное умножение матриц соответствует суперпозиции перестановок.

Как видим, матричное представление различных числовых объектов насыщено разнообразными примерами (и позже мы еще раз вернемся к нему в связи с расширениями поля \mathbb{Q}). На этом примере нам хотелось показать еще один **архетип математики**, а именно: **вариативность представления математических объектов**. Действительно, математика настолько неделимая наука, что очень многие вещи в ней как по волшебству начинают проявляться в самых, казалось бы, далеких друг от друга разделах. Матрицы, которые мы привыкли считать основой линейной алгебры, вполне себе могут породить всевозможные виды хорошо (и не очень) известных нам чисел, а также отвечать за представление групп. Понятие алгебры — вполне себе числовое, однако оно находит применение и в теории множеств, и в геометрии, и в информатике, и еще много где.

Поэтому в современной математике очень важно уметь видеть один и тот же (с точностью до изоморфизма) объект в своих многочисленных ипостасях, чтобы результаты, полученные для него в одной ветви математики, с успехом применять в других ветвях математики и других науках, которые, на первый взгляд, никак не связаны друг с другом. В этом архетипе, пожалуй, кроется та непостижимая и колоссальная мощь, которую дает математика человечеству.

3.3 Гауссовы целые числа

Великая теорема Ферма, известная своей неприступностью в течение 3 столетий, породила много нетривиальных методик исследований в математике и привела к созданию целых теорий. В свое время Д. Гильберта спрашивали, почему он не хочет расколоть орешек теоремы Ферма, на что он ответил: нет смысла убивать курицу, несущую золотые яйца [7]. Имея ввиду тот пласт новых веяний в математике XIX — начала XX веков, который породила теорема Ферма и вообще вся наука о диофантовых уравнениях.

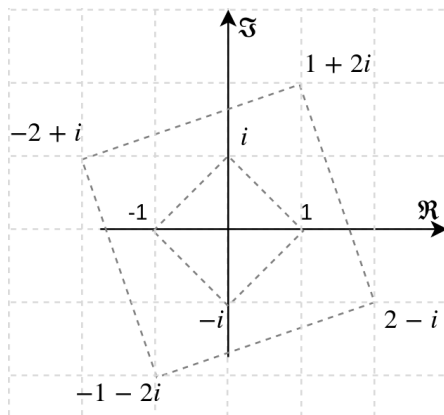


*Карл
Фридрих
Гаусс*

Этот и следующий разделы целиком основаны на цикле видеолекций д.ф.-м.н. А. Савватеева «Теорема Ферма» [sDgeroYKMEg], а также рекомендуем книгу [50].

Для начала мы познакомимся с целыми числами Гаусса. Рассмотрим на \mathbb{C} целочисленную решетку $\{x + iy \mid x, y \in \mathbb{Z}\}$ с обычными операциями над комплексными числами. Легко проверить, что это кольцо с единицей без делителей нуля, в котором обратимыми элементами являются только члены мультипликативной группы корней 4 степени из 1, т. е. $\pm 1, \pm i$.

Кольцо гауссовых целых чисел обозначается $\mathbb{Z}[i]$. Это подчеркивает тот факт, что данное кольцо получено из \mathbb{Z} присоединением числа i и замыканием с помощью линейных комбинаций. Это — прямой аналог группового кольца, если вместо группы корней взять группу \mathbb{Z}_2 и базисные векторы полученной двумерной алгебры обозначить как 1 и i . Но полная аналогия (и, соответственно, изоморфизм) здесь неуместна, т. к. произведение в групповом кольце $\mathbb{Z}[\mathbb{Z}_2]$ и в кольце $\mathbb{Z}[i]$ отличается. Действительно, в первом случае имеем:



$$(x, y) \cdot (x', y') = (xx' + yy', xy' + x'y) \quad (\text{напомним представление в виде } x + gy),$$

а во втором случае имеем:

$$(x + iy)(x' + iy') = (xx' - yy') + i(xy' + x'y).$$

В гауссовых числах можно построить свою теорию делимости.

3.3.1 Делимость и простые числа

Число $a + bi$ **делится на** $c + di$, если существует число $a' + b'i$ такое, что $a + bi = (c + di)(a' + b'i)$.¹⁴ Обозначение аналогично обычному в натуральных числах: $(c + di)|(a + bi)$. Например, число 2 делится на $(1 + i)$, т. к. $2 = (1 + i)(1 - i)$.

Нормой гауссова числа $a + bi$ называется величина

$$N(a + bi) = (a + bi)(a - bi) = a^2 + b^2$$

Несколько свойств нормы:

Norm1 $N(a + bi) = 0$ тогда и только тогда, когда $a = b = 0$;

Norm2 Нормы комплексно сопряженных чисел совпадают;

Norm3 Если норма нечётна, то она имеет вид $4k + 1$, никакая норма не может быть равна $4n + 3$;

Norm4 $N(zw) = N(z)N(w)$, где z, w — гауссовы числа.

Последнее свойство означает, что делителями единицы (обратимыми элементами) могут быть только числа с нормой 1, т. е. ± 1 и $\pm i$. Других обратимых нет.

Все гауссовы числа делятся на делители единицы.

Два гауссовых числа называют **ассоциированными**, если одно получается из другого умножением на делитель единицы. Ассоциированность является отношением эквивалентности, причем каждый класс эквивалентности включает ровно 4 числа, расположенных в углах квадрата с центром в 0. Например, $1 + 2i$, $-2 + i$, $-1 - 2i$ и $2 - i$ ассоциированы.

Свойства делимости гауссовых чисел очень похожи на таковые свойства в арифметике натуральных чисел, но есть и отличия.

Приведем несколько свойств:

Div1 Если гауссово число $a + bi$ делится на обычное целое число $c + i0$, то $c|a$ и $c|b$ в целых числах;

Div2 Если $z|w$ и $w|z$, то z и w ассоциированы;

Div3 Ассоциированность сохраняет делимость: если z и w ассоциированы, u и v ассоциированы, то $(z|u) \rightarrow (w|v)$;

Упражнение
3.26.
Проверьте
свойства.

Упражнение
3.27.

¹⁴В этом разделе мы по умолчанию считаем, что все использованные вещественные числа — целые, если не указано обратное.

Div4 z ($N(z) > 1$) имеет как минимум 8 делителей: своих ассоциированных и ассоциированных с 1;

Div5 Делители z являются делителями $N(z)$;

Div6 Норма $z = a + bi$ четна тогда и только тогда, когда $(1+i)|z$, в частности, если a и b имеют разную четность, то z не делится на $1+i$;

В кольце $\mathbb{Z}[i]$ можно любое число u разделить на любое число $v \neq 0$ с остатком, так что получится

$$u = qv + r, \quad N(r) < N(v). \quad (3.6)$$

При этом выбор чисел q и r можно строго ограничить, выбирая q как ближайшее гауссово число к комплексному u/v , а r как разность между u и qv . В случае, когда выбор q неоднозначен (может быть максимум 4 числа), можно договориться выбирать то, которое на координатной сетке находится левее и/или ниже.

На основе деления с остатком нетрудно получить выполнимость алгоритма Евклида (2.4) для гауссовых чисел и представление НОД в виде линейной комбинации исходных чисел. Докажем этот факт иным способом.

Лемма 3.2. *Для любых гауссовых чисел $u, v \neq 0$ существует гауссово число r такое, что:*

- 1) $r|u$ и $r|v$ (общий делитель);
- 2) если $(q|u) \wedge (q|v)$, то $q|r$ (наибольший общий делитель);
- 3) существуют гауссовы x, y такие, что $r = xu + yv$.

Кроме того,

- 4) число r , удовлетворяющее 1)–3), единственное с точностью до ассоциированности.

Доказательство. Рассмотрим множество $R(u, v) = \{xu + yv \mid x, y \in \mathbb{Z}[i]\} \setminus \{0\}$. В множестве норм $\{N(z) \mid z \in R\}$ существует наименьшее положительное число (т. к. нуля там быть не может). Пусть $r \in R$ такое число, у которого норма минимальная (оно может быть не единственное, выберем одно). Остается показать, что r — искомое.

Во-первых, r имеет вид $xu + yv$ по построению. Во-вторых, если $(q|u) \wedge (q|v)$, то очевидно, что $q|r$ также по построению r .

Докажем пункт 1).

Из (3.6) имеем: $u = rt + s$, где $N(s) < N(r)$. Подставляя представление r , имеем: $u = xut + yvt + s$, откуда $s = (1 - xt)u + (-yt)v$. Если $s \neq 0$, то $s \in R$ как линейная комбинация u и v , но тогда $N(s) \geq N(r)$ в силу выбора r , а это не так в силу (3.6). Следовательно, $s = 0$, откуда $r|u$. Аналогично, $r|v$.

Докажем пункт 4). Пусть $r' = x'u + y'v$ также удовлетворяет свойствам 1)–3). Тогда $r|r'$ и $r'|r$. Из первого следует, что $r' = rt$ и $N(r') = N(r)N(t)$, из второго следует, что $r = r't'$ и $N(r) = N(r')N(t')$. Таким образом, нормы $N(t)$ и $N(t')$ взаимно обратны в натуральных числах, откуда следует $N(t) = N(t') = 1$, т. е. t — делитель 1 и, следовательно, r и r' ассоциированы. \square

Доказанная лемма позволяет определить понятие НОД для гауссовых чисел с точностью до ассоциированности. В качестве НОД мы будем выбирать какое-то одно из четырех (наиболее удобного вида).

Гауссово число называется **простым**, если оно не имеет никаких делителей, кроме тривиальных (ассоциированных с 1 и самим собой), и не является делителем 1, т. е. простое число имеет ровно 8 делителей. Два гауссовых числа называются **взаимно простыми** (обозначается $u \perp v$), если их НОД — обратное число, т. е. 1 и остальные делители 1.

Докажите следующие свойства простых гауссовых чисел:

Упражнение
3.28.

Prim1 Если $a + bi$ простое, то $a - bi$ также простое;

Prim2 Если z простое, то его ассоциированные также простые;

Prim3 Если z простое и $z|uv$, то $(z|u) \vee (z|v)$;

Prim4 Норма простого, неассоциированного с $1 + i$, всегда нечетна, т. е. имеет вид $4k + 1$;

Prim5 Натуральное простое не всегда есть гауссово простое: $5 = (2 + i)(2 - i)$;

Prim6 Критерий Гаусса: $a + bi$ простое тогда и только тогда, когда 1) либо одно из чисел a, b нулевое, а второе — простое целое число вида $\pm(4k + 3)$, 2) либо a, b ненулевые и норма $N(a + bi) = a^2 + b^2$ — простое натуральное число;

Prim7 Следствие: простое натуральное вида $4k + 1$ не может быть простым гауссовым, простые натуральные вида $4k + 3$ являются простыми гауссовыми;

Prim8 Простое натуральное $4k + 1$ можно представить как сумму квадратов $a^2 + b^2$ (рождественская теорема Ферма);

Prim9 Если $N(a) \perp N(b)$ в натуральных числах, то $a \perp b$ в гауссовых числах.

Примеры простых гауссовых чисел: $\pm 3, \pm 7, \pm 3i; 1 \pm i, 1 \pm 2i, 1 \pm 4i$.

Для примера покажем свойство Prim3. Пусть простое $z|uv$. Предположим, что $\neg(z|u)$, тогда $z \perp u$, откуда по лемме 3.2 получаем, что $1 = xz + yu$. Умножаем на v : $v = xzv + yuv$. Справа оба слагаемых делятся на z , следовательно $z|v$. Аналогично, если $\neg(z|v)$, то $z|u$.

И свойство Prim9. Пусть $r = \text{НОД}(a, b)$ в гауссовых числах. Тогда $a = rt$, $b = rt'$ и $N(a) = N(r)N(t)$, $N(b) = N(r)N(t')$. Откуда $N(r)|N(a)$ и $N(r)|N(b)$. Тогда из условия $N(a) \perp N(b)$ следует, что $N(r) = 1$, т. е. r — ассоциированное с 1 гауссово число. Откуда $a \perp b$.

Для гауссовых чисел существует аналог основной теоремы арифметики 2.6:

Теорема 3.7 (Основная теорема арифметики гауссовых чисел).

Каждое ненулевое неассоциированное с 1 гауссово число раскладывается на гауссовы простые множители, причем это разложение единственно с точностью до ассоциированных с этими множителями простых чисел и порядка множителей, т. е. разложение имеет вид

$$\alpha_1 \dots \alpha_n = \beta_1 \dots \beta_n,$$

где пары (α_i, β_i) являются ассоциированными простыми числами (при этом a_i, a_j и b_i, b_j при $i \neq j$ также могут быть ассоциированными).

Доказательство теоремы прямо следует из свойства Prim3.

Пример: $5 = (2 + i)(2 - i) = (1 + 2i)(1 - 2i)$ (множители переводятся друг в друга умножением на i и на $-i$).

Лемма 3.3. *Если $(u \perp v) \wedge (uv = c^n)$, то существуют $a \perp b$ такие, что $u = a^n$ и $v = b^n$ и $c = ab$.*

Доказательство непосредственно следует из теоремы 3.7.

Заметим, что и в обычной арифметике целых чисел верна такая же теорема. Более того, как основная теорема арифметики 3.7, так и лемма 3.3 верны в любом **евклидовом кольце** (т. е. в таком кольце, где возможно деление с остатком в виде (3.6) при некоторой натурально-значной норме и, как следствие, алгоритм Евклида (2.4)). Этим свойством евклидовых колец мы еще воспользуемся в дальнейшем.

3.3.2 Некоторые приложения гауссовых чисел

Рассмотрим диофантово (т. е. в целых числах) уравнение

$$x^2 + 1 = y^3.$$

В гауссовых числах оно эквивалентно уравнению

$$(x + i)(x - i) = y^3.$$

Покажем, что $x + i \perp x - i$. Действительно, если это не так, т. е. $z|x + i$ и $z|x - i$, то $z|(x + i) - (x - i) = 2i$, откуда $z = 1 + i$ или ему ассоциированное. Кроме того, $z|y^3$, причем, поскольку $1 + i$ — простое, оно должно входить в разложение y^3 трижды, т. е. $z^3|y^3$, но тогда в разложение $x + i$ или $x - i$ входит $z^2 = 2i$, чего быть не может, т. к. $x \pm i$ не делится на 2 (см. свойство Div1). Следовательно, $x + i \perp x - i$.



Диофант
Александрий-
ский

Тогда в силу леммы 3.3 существует число $a + bi$ такое, что $x + i = (a + bi)^3$. Возводя в куб и сравнивая коэффициенты при i , находим, что $1 = b(a^2 - b^2)$. Это — уравнение в целых числах, поэтому $b = \pm 1$, откуда $a^2 = 0$ или 2. Но $a^2 = 2$ неразрешимо в целых числах, поэтому $a = 0$, откуда $x = 0$. Таким образом, единственно возможное решение в целых числах у исходного уравнения $x^2 + 1 = y^3$ — это $x = 0, y = 1$.

Рассмотрим **Теорему Ферма** при $n = 2$: $a^2 + b^2 = c^2$ (в натуральных числах). Ясно, что можно сразу считать, что все числа a, b, c попарно взаимно простые натуральные числа (иначе можно было бы сократить уравнение на общий множитель). Отсюда также следует, что a и b имеют разную четность. Действительно, если a и b четные, то таково же и c , а значит, они не взаимно простые. Если a и b нечетные, то $a^2 + b^2$ имеет остаток 2 при делении на 4, но c^2 может иметь остаток либо 0 (четное), либо 1 (нечетное). Таким образом, допускается только случай, когда a и b имеют различную четность. Тогда по свойству Div6 число $a + bi$ не делится на $1 + i$.

Заметим, что $(a + bi)(a - bi) = a^2 + b^2 = c^2$. Предположим, что НОД чисел $a + bi$ и $a - bi$ равен r и отличен от делителя 1. Тогда $r|2a$ и $r|2bi$. Но $a \perp b$ в натуральных числах, тогда $N(a) \perp N(b)$, откуда по свойству Prim9 $a \perp b$ в гауссовых числах. Это значит, что r есть НОД 2 и $2i$, т. е. $r = 1 + i$ или его ассоциированным. Но такое число не может быть делителем $a + bi$ и $a - bi$ по доказанному выше. Следовательно, $(a + bi) \perp (a - bi)$.

Тогда по лемме 3.3 существуют такие z, w , что $a + bi = z^2$, $a - bi = w^2$ и $c = zw$. Пусть $z = n + mi$, тогда $a + bi = n^2 - m^2 + 2nmi$, откуда $a - bi = n^2 - m^2 - 2nmi$, откуда $w = n - mi$ и $c = n^2 + m^2$.

Таким образом, мы получаем формулу **пифагоровых троек**:

$$a = n^2 - m^2, \quad b = 2nm, \quad c = n^2 + m^2,$$

где натуральные $n, m > 0$.

Рассмотрим теперь уравнение $x^4 + y^4 = z^4$, неразрешимость которого доказал еще сам Ферма методом, который мы покажем ниже. Он в чем-то перекликается с методом доказательства теоремы Гудстейна.

Бесконечный
спуск

Докажем более сильное утверждение: $x^4 + y^4 = z^2$ неразрешимо в целых положительных числах. Как и прежде, считаем сразу же, что $x \perp y$. Посмотрим на это уравнение как на уравнение второй степени: $(x^2)^2 + (y^2)^2 = z^2$.

Если оно разрешимо, то существуют ненулевые взаимно простые n, m такие, что

$$x^2 = n^2 - m^2, \quad y^2 = 2nm, \quad z = n^2 + m^2,$$

откуда вновь получаем уравнение второй степени $x^2 + m^2 = n^2$, а значит, его решение имеет вид:

$$x = a^2 + b^2, \quad m = 2ab, \quad n = a^2 + b^2,$$

где ненулевые $a \perp b$. Тогда для y имеет место равенство: $y^2 = 4nab$ и, поскольку число 2 простое (в обычных целых числах), $y = 2y'$.

Тогда $(y')^2 = nab$. Так как n, a, b попарно взаимно просты (это следует из того, что $a \perp b$ и $n = a^2 + b^2$), в силу леммы 3.3 (для обычных целых чисел) существуют такие s, t, k , что $n = s^2$, $a = t^2$, $b = k^2$. Подставляем это в равенство $n = a^2 + b^2$, получаем:

$$t^4 + k^4 = s^2,$$

где $t \perp k$ и $z > s > 0$ (это следует из того, что $s = \sqrt{n}$, $n^2 < z$).

Таким образом, имея одно решение (x, y, z) исходного уравнения, мы построили еще одно (t, k, s) , где $s < z$. Продолжая применять эти построения далее, мы получим бесконечную последовательность решений (t_j, k_j, s_j) такую, что $z > s > s_1 > s_2 > \dots$. Но это невозможно, т.к. в натуральном ряде, как и в любом ординале, не существует бесконечная строго убывающая последовательность.

Полученное противоречие доказывает неразрешимость уравнения $x^4 + y^4 = z^2$ в целых положительных числах, а значит, и неразрешимость уравнения $x^4 + y^4 = z^4$. Заметим, что отсюда сразу же следует справедливость теоремы Ферма для всех степеней n , кратных 4.

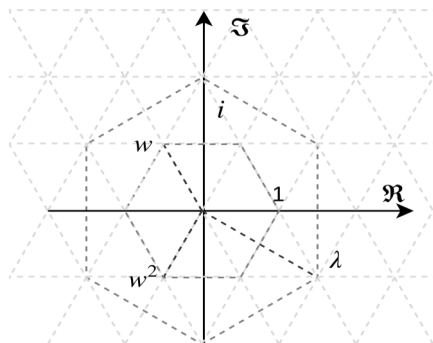
Предъявленный здесь метод доказательства называется **методом бесконечного спуска**, который мы отнесем к архетипическим методам. Он напоминает индукцию, только не доказующую, а опровергающую, поскольку приводит к противоречию. То же самое мы делали, когда строили строго убывающую последовательность ординалов при доказательстве теоремы Гудстейна. Только там мы уперлись не в противоречие, а в ноль за конечное число шагов.

3.4 Целые числа Эйзенштейна

Числа Гаусса, как мы видели, содержат корни 4 степени из 1. Рассмотрим теперь треугольную решетку чисел, содержащую корни 6 степени из 1, т.е. числа вида $\pm 1, (\pm 1 \pm i\sqrt{3})/2$, по формуле Эйлера это числа вида $e^{i\pi k/3}$, где $k = 0, 1, \dots, 5$.

Как и для чисел Гаусса в качестве порождающего элемента достаточно взять одно число $w = (-1 + i\sqrt{3})/2 = e^{2i\pi/3}$ ¹⁵, остальные получаются как линейная комбинация w и целых чисел. Таким образом, мы рассматриваем множество комплексных чисел

$$\mathbb{Z}[w] \Rightarrow \{a + bw \mid a, b \in \mathbb{Z}\}$$



Эти числа ложатся на треугольную решетку с углом наклона 60° и образуют коммутативное кольцо с 1. Они называются **числами Эйзенштейна** (в некоторых источниках — числами Эйлера).

Для чисел Эйзенштейна также вводится понятие нормы:

$$N(a + bw) = a^2 + b^2 - ab,$$

которая равна нулю тогда и только тогда, когда $a = b = 0$. На самом деле, это все та же норма комплексных чисел:

$$N(a + bw) = N(a - b/2 + ib\sqrt{3}/2) = (a - b/2)^2 + (b\sqrt{3}/2)^2 = a^2 + b^2 - ab.$$

Таким образом, у ненулевых чисел минимальная норма равна 1, и этой нормой обладают все корни 6 степени из 1. Они же являются обратимыми в кольце $\mathbb{Z}[w]$, и только они.

Норма чисел Эйзенштейна наследует свойства нормы комплексных чисел, в частности, она мультипликативна: $N((a + bw)(a' + b'w)) = N(a + bw)N(a' + b'w)$. Кроме того, $N(a + bw) = (a + bw)\overline{(a + bw)}$. При этом $\bar{w} = w^2 = -1 - w$, так что $N(a + bw) = (a + bw)(a - b - bw)$.

Для чисел Эйзенштейна выполняется деление с остатком в виде (3.6) и, соответственно, алгоритм Евклида (2.4). При этом, как и для гауссовых чисел, все отношения следует воспринимать с точностью до коэффициентов-обратимых.

Деление числа u на число v нацело означает, что $u = v\varepsilon$, где ε — одно из обратимых чисел. Как и гауссовы, числа Эйзенштейна делятся на классы ассоциированных. Для каждого числа (кроме нуля) их ровно 6 — они получаются умножением этого числа на обратимые и образуют правильный 6-угольник с центром в нуле. Соответственно, простым числом Эйзенштейна называется такое число, которое имеет ровно 12 делителей — свои ассоциированные и обратимые.

Упражнение 3.29. Проверьте, что $\mathbb{Z}[w]$ образует кольцо

¹⁵Привычный символ ω у нас занят, так что — w .

Пример: число 3 не является простым, т. к. $3 = (1 - w)(2 + w)$. Общепринятое обозначение: $\lambda = 1 - w$. Заметим, что $\bar{\lambda} = 2 + w$, так что $3 = \lambda\bar{\lambda}$. Число λ является простым числом Эйзенштейна.

В числах Эйзенштейна, как в любом евклидовом кольце, выполняется основная теорема арифметики 3.7 о разложении на простые с точностью до обратимых, а также ее следствие — лемма 3.3.

Теперь покажем одну техническую лемму для демонстрации работы с числами Эйзенштейна и затем перейдем к рассмотрению теоремы Ферма для случая $n = 3$.

Лемма 3.4. Для любого $z \in \mathbb{Z}[w]$ либо $\lambda|z$, либо $\lambda^4|z^3 + 1$, либо $\lambda^4|z^3 - 1$.

Доказательство. Для начала покажем следующую трихотомию:

$$(\lambda|z) \vee (\lambda|z + 1) \vee (\lambda|z - 1) \quad (3.7)$$

Предположим, что первое неверно, тогда $z = \lambda q + r$, причем $N(r) < N(\lambda) = 3$ (используем деление с остатком). Нетрудно проверить, что в этом случае r будет иметь норму 1, т. е. будет обратимым. Но тогда либо $r + 1$, либо $r - 1$ делится на λ . Действительно, $1 - 1 = 0$, $w - 1 = -\lambda$, $w^2 - 1 = -\bar{\lambda}$, т. е. степени w при сдвиге влево на 1 дают числа, кратные λ . Аналогично, остальные три обратимых при сдвиге вправо на 1 дают числа, кратные λ . Отсюда следует (3.7).

Теперь докажем лемму. Пусть z не делится на λ , тогда $z = \lambda q \pm 1$ при некотором q , откуда

$$z^3 \mp 1 = \lambda^3 q^3 \pm 3\lambda^2 q^2 + 3\lambda q \pm 1 \mp 1 = \pm 3\lambda^2 q^2 + \lambda^3 q^3 + 3\lambda q,$$

где мы согласовали знаки $+$ и $-$ с представлением $z = \lambda q \pm 1$. Заметим, что $3\lambda^2 q^2$ делится на λ^4 , т. к. $3 = \lambda\bar{\lambda}$. Осталось рассмотреть $\lambda^3 q^3 + 3\lambda q$, для чего заметим, что $\lambda^3 = -3\lambda w$ и $w^3 = 1$, откуда

$$\lambda^3 q^3 + 3\lambda q = 3\lambda q w^3 - 3\lambda w q^3 = 3\lambda w q(w^2 - q^2).$$

Если q делится на λ , то полученное выражение, очевидно, делится на λ^4 , в противном случае $q = \lambda r \pm 1$ в силу (3.7), но тогда $q^2 - 1$ делится на λ . Отсюда получаем, что

$$w^2 - q^2 = (w^2 - 1) - (q^2 - 1) = -\bar{\lambda} - (q^2 - 1),$$

что также делится на λ , а значит, исходное выражение $3\lambda w q(w^2 - q^2)$ делится на λ^4 . Лемма доказана. \square

Рассмотрим теперь теорему Ферма при $n = 3$, т.е. уравнение в целых числах вида

$$x^3 + y^3 = z^3, \quad (3.8)$$

где мы предполагаем, что x, y, z отличны от 0 и попарно взаимно просты (иначе все три можно сократить на общий делитель).

Упражнение 3.31. Заметим также, что x, y, z , будучи взаимно простыми в \mathbb{Z} , остаются взаимно простыми и в $\mathbb{Z}[w]$, что позволяет нам выйти в числа Эйзенштейна и рассмотреть более сильное утверждение: уравнение (3.8) неразрешимо в ненулевых числах Эйзенштейна. При этом предполагая сразу, что x, y, z попарно взаимно просты как числа Эйзенштейна.

Рассмотрим два случая.

Случай (I): ни одно из чисел x, y, z не делится на λ .

Случай (II): одно из чисел x, y, z делится на λ (два быть не может, т.к. они окажутся не взаимно простыми).

В случае (I) мы применяем лемму 3.4 и от равенства (3.8) переходим к равенству

$$\lambda^4 x_1 \pm 1 + \lambda^4 y_1 \pm 1 = \lambda^4 z_1 \pm 1,$$

где x_1, y_1, z_1 — какие-то числа Эйзенштейна, а знаки \pm не согласованы (т.е. имеют место 8 различных комбинаций знаков). Но тогда выражение $\pm 1 \pm 1 \pm 1$ должно быть кратно λ^4 . Однако, оно, очевидно, не может быть нулевым, а ненулевое выражение такого вида имеет норму не больше 9, в то время как норма λ^4 равна 81. Это значит, что данное выражение не может делиться на λ^4 . Таким образом, случай (I) невозможен.

В случае (II) можно без ограничения общности считать, что именно z делится на λ , в противном случае мы просто переобозначим переменные и перенесем их из одной части равенства (3.8) в противоположную. После чего мы еще больше усилим доказываемое утверждение, а именно: уравнение

$$x^3 + y^3 = z^3 \varepsilon, \quad (3.9)$$

где ε — произвольное обратимое, а z делится на λ , неразрешимо в ненулевых числах Эйзенштейна. Ясно, что отсюда автоматически следует теорема Ферма при $n = 3$ для случая (II). Докажем это новое утверждение.

Уточним делимость z на λ : пусть z делится на λ^k , но не делится на λ^{k+1} (такое k всегда существует).

Случай (IIa): $k > 1$. В этом случае покажем, что можно найти новую тройку x_1, y_1, z_1 такую, что она удовлетворяет уравнению (3.9), но при этом z_1 делится на λ^{k-1} , но не делится на λ^k . То есть перейдем от решения со степенью k к решению со степенью $k - 1$.

Перепишем уравнение (3.9) в виде:

$$(x + y)(x + yw)(x + yw^2) = z^3 \varepsilon = \lambda^{3k} q^3 \varepsilon, \quad (3.10)$$

где q не делится на λ . Путем несложных арифметических упражнений,¹⁶ учитывая взаимную простоту x, y, z , несложно показать, что степень λ распределена крайне неравномерно в сомножителях $(x+y)(x+yw)(x+yw^2)$, а именно: в одном из них находится степень λ^{3k-2} , а в двух других — только первая степень. Не теряя общности, можно считать, что $\lambda^{3k-2} | x+y$. Таким образом,

Вот где на самом деле нужны новые числа!

$$x+y = \lambda^{3k-2}a, \quad x+yw = \lambda b, \quad x+yw^2 = \lambda c,$$

где a, b, c взаимно просты и не содержат степеней λ .

После подстановки в (3.10) и сокращения на λ^{3k} получаем, что $abc = q^3\varepsilon$. Отсюда по лемме 3.3 (которая верна и в случае чисел Эйзенштейна) получаем, что $a = t^3\varepsilon_1$, $b = u^3\varepsilon_2$, $c = v^3\varepsilon_3$, где t, u, v взаимно просты и не делятся на λ .¹⁷

Кроме того, заметим, что $(x+y) + (x+yw)w + (x+yw^2)w^2 = 0$, откуда имеем

$$\lambda^{3k-2}a + \lambda bw + \lambda cw^2 = 0$$

или

$$u^3\varepsilon_2w + v^3\varepsilon_3w^2 = -(\lambda^{k-1}t)^3\varepsilon_1$$

или

$$u^3 + v^3\varepsilon_4 = (\lambda^{k-1}t)^3\varepsilon_5, \quad (3.11)$$

где $\varepsilon_4 = w\varepsilon_3/\varepsilon_2$ и $\varepsilon_5 = -\varepsilon\varepsilon_1/(w\varepsilon_2)$.

Теперь заметим, что правая часть делится как минимум на λ^3 , а в левой части (3.11) стоят кубы, для которых по лемме 3.4 имеет место представление $u^3 = \lambda^4g \pm 1$ и $v^3 = \lambda^4h \pm 1$, так что на λ^3 должно делиться выражение $\pm 1 \pm \varepsilon_4$, норма которого не превышает 4, в то время как норма λ^3 равна 27, следовательно, единственно возможный случай: $\pm 1 \pm \varepsilon_4 = 0$. Это значит, что $\varepsilon_4 = \pm 1$ и, значит, $v^3\varepsilon_4 = (v\varepsilon_4)^3$.

Подставляя полученное в (3.11) и производя замену $x_1 = u$, $y_1 = v\varepsilon_4$, $z_1 = \lambda^{k-1}t$, получаем равенство

$$x_1^3 + y_1^3 = z_1^3\varepsilon_4,$$

которое в точности соответствует равенству (3.9), только с пониженной степенью λ в правой части.

Таким образом, случай (IIa) методом бесконечного спуска сводится к ситуации, когда z делится только на первую степень λ .

¹⁶Нужно показать, что λ^2 не может одновременно делить никакие два множителя.

¹⁷Чтобы в точности попасть в условия леммы 3.3, мы сначала перенесли ε влево, соединив его с одним из множителей a, b, c , а затем вернули обратно вправо в представлении этого множителя. Чтобы одной записью закрыть все три случая, мы ввели три обратимых $\varepsilon_1, \varepsilon_2, \varepsilon_3$, подразумевая, что только один из них отличен от 1.

Случай (IIb): $k = 1$. Но этот случай невозможен по следующим причинам. Поскольку z не делится на λ^2 , запишем его в виде λq , где q не делится на λ , кроме того, $x^3 = \lambda^4 t \pm 1$ и $y^3 = \lambda^4 r \pm 1$ в силу леммы 3.4. Тогда

$$\lambda^4 t + \lambda^4 r \pm 1 \pm 1 = \lambda^3 q^3,$$

где $\pm 1 \pm 1$ не может быть равно 2 или -2 , поскольку 2 не делится на λ , но $\pm 1 \pm 1$ не может быть равно и нулю, поскольку в этом случае выражение слева будет делиться на λ^4 , а справа — нет. Других вариантов для $\pm 1 \pm 1$ нет, поэтому случай $k = 1$ также невозможен.

Итак, в случае (II) мы приходим к противоречию, т.е. уравнение (3.9) неразрешимо в ненулевых взаимно простых x, y, z . Следовательно, теорема Ферма верна при $n = 3$.

3.5 Линейные пространства и операторы

Ранее мы определили линейное пространство как модуль над полем. Мы также ввели понятие базиса и размерности пространства. Рассмотрим теперь два традиционных примера линейных пространств: \mathbb{R}^n и \mathbb{C}^n , используя общее обозначение L_n для того и другого пространства и общее обозначение F для их базового поля, когда нам будет неважна природа этого поля (\mathbb{R} или \mathbb{C}).

На пространстве L_n мы можем рассмотреть группу биекций $S(L_n)$ с операцией композиции. Среди таких биекций мы можем выделить такие $f \in S(L_n)$, которые удовлетворяют **условиям линейности**:

Lin1 $f(x + y) = f(x) + f(y)$ (аддитивность);

Lin2 $f(ax) = af(x)$ (однородность 1-го порядка),

где $x, y \in L_n$, a — число из базового поля L_n . Такие функции называются **обратимыми линейными операторами** над пространством L_n . Обратимость здесь обусловлена тем, что это биекции, т.е. данные операторы имеют обратные (в смысле обратной функции). В общем случае линейный оператор может быть необратим.

Обратимые линейные операторы над L_n образуют группу с операцией композиции, которая обозначается $GL(L_n)$ и называется **полной линейной группой** пространства L_n .

Отметим также, что линейные операторы (не обязательно обратимые) можно складывать и умножать на число из того поля, над которым они заданы. Это значит, что мы можем рассматривать векторное пространство линейных операторов над полем L_n . Более того, на этом пространстве определено умножение (как композиция операторов в группе $S(L_n)$), поэтому

Осталось рассмотреть многочлены над операторами, и будет «картина маслом».

можно говорить об алгебре линейных операторов, которая изоморфна алгебре квадратных матриц над полем L_n .

Введение базиса в пространстве L_n (размерность которого равна n) позволяет взаимно однозначно сопоставить каждому обратимому линейному оператору некоторую обратимую матрицу размера $n \times n$ с элементами из поля F . Действительно, если e_1, \dots, e_n — базисные векторы, и $a^{(1)}, \dots, a^{(n)}$ — их образы относительно линейного оператора f (т.е. $a^{(j)} = f(e_j)$), то, пользуясь единственностью разложения по базисным векторам $x = x_1 e_1 + \dots + x_n e_n$, $a^{(j)} = a_1^{(j)} e_1 + \dots + a_n^{(j)} e_n$ и $f(x) = y_1 e_1 + \dots + y_n e_n$, получаем, что $f(x) = x_1 a^{(1)} + \dots + x_n a^{(n)}$ и, кроме того,

$$y = \begin{pmatrix} a_1^{(1)} & \dots & a_1^{(n)} \\ \vdots & \ddots & \vdots \\ a_n^{(1)} & \dots & a_n^{(n)} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = Ax$$

где столбцы матрицы A составлены из координат векторов $a^{(1)}, \dots, a^{(n)}$ в выбранном базисе, x — столбец координат вектора x в этом же базисе, и y — столбец координат вектора $y = f(x)$ в этом базисе.

Для любой квадратной матрицы A задается ее определитель $\det A$. Одно из его определений таково:

$$\det A = \sum_{\sigma \in S_n} (-1)^{\text{sgn}(\sigma)} a_{1\sigma(1)} \dots a_{n\sigma(n)},$$

т.е. берем какую-нибудь перестановку σ из группы S_n и перемножаем элементы матрицы, строки которых пронумерованы подряд, а столбцы выбраны в соответствии с этой перестановкой, после чего суммируем все полученные произведения с коэффициентами, равными четности соответствующих перестановок. Таким образом, со знаком «плюс» складываются те произведения элементов A , которые получены с помощью четных перестановок (т.е. перестановок из группы A_n), а с «минусом» — остальные.

Некоторые свойства определителя:

Det1 $\det E = 1$;

Det2 $\det A^T = \det A$;

Det3 $\det(AB) = \det A \det B$;

Det4 $\det(A^{-1}) = (\det A)^{-1}$.

Это далеко не все важные свойства определителя, например, его можно интерпретировать как ориентированный объем параллелепипеда, построенного на трех векторах в \mathbb{R}^3 , составляющих матрицу A , определитель которой и

совпадает с ориентированным объемом. Кроме того, определитель помогает решать системы линейных уравнений вида $Ax = b$ методом Крамера.

Матрица, определитель которой равен 0, называется вырожденной. Матрица обратима тогда и только тогда, когда обратим ее определитель (т.е. когда она невырождена).

Обратимому линейному оператору соответствует невырожденная матрица. Обратно, по любой невырожденной матрице A легко восстанавливается соответствующий ей обратимый линейный оператор. Кроме того, обратному линейному оператору соответствует обратная матрица (в смысле умножения матриц).

Итак, любому обратимому линейному оператору ставится во взаимно однозначное соответствие обратимая квадратная матрица, если зафиксирован некоторый базис в пространстве L_n . При смене базиса меняются и матрицы всех линейных операторов некоторым единообразным способом, но при этом определители их матриц не меняются! Это легко понять из следующей диаграммы:

$$\begin{array}{ccc} x & \xrightarrow{A} & y \\ P \uparrow & & \downarrow P^{-1} \\ \tilde{x} & \xrightarrow{P^{-1}AP} & \tilde{y} \end{array}$$

Здесь матрица перехода P получается так: ее столбцы — это координаты новых базисных векторов в старом базисе. Вектор $y = Ax$ — это результат действия оператора A на вектор x , при этом как векторы x, y , так и матрица A выражены в координатах старого базиса. Далее, $\tilde{x} = P^{-1}x$ — это координаты вектора x в новом базисе (поскольку $x = P\tilde{x}$ — вычисление старых координат через новые и матрицу перехода), а $\tilde{y} = P^{-1}y$ — это координаты вектора $y = Ax$ в новом базисе. Итого получаем

$$\tilde{y} = P^{-1}y = P^{-1}Ax = (P^{-1}AP)\tilde{x},$$

откуда мы видим, что матрицей оператора A в новом базисе будет матрица $P^{-1}AP$.

Но тогда по свойствам определителя имеем:

$$\det(P^{-1}AP) = \det(P^{-1}) \det A \det P = \det A.$$

Естественно, мы предполагаем при этом, что P обратима. На самом деле, это всегда так, если матрица P является матрицей перехода от одного базиса к другому (в противном случае векторы второго базиса оказались бы линейно зависимыми).

Матрице перехода соответствует обратимый линейный оператор, который переводит старый базис в новый.

Множество всех обратимых матриц размера $n \times n$ над полем F обозначается $\text{GL}(n, F)$ или $\text{GL}(n)$ (если не требуется уточнить поле) и образует группу, изоморфную группе $\text{GL}(L_n)$. Поэтому все три обозначения могут равнозначно использоваться для данной группы.

Как видим, мы вновь открыли для себя группу перестановок, как в ее бесконечном виде ($\text{S}(L_n)$), так и в конечном ($\det A$).

При изучении линейных пространств особая роль отводится их топологизации, т. е. добавлению такой структуры, которая отвечала бы за геометрию в этих пространствах. Точнее, мы задаем на векторах произведение, результатом которого является число из базового поля. Это не делает линейное пространство алгеброй, но позволяет определить понятия нормы, расстояния, непрерывности. Иначе говоря, мы снова стараемся максимально уподобиться комплексным числам, как наиболее удобной рабочей площадке математического анализа.

Скалярным произведением векторов $x, y \in L_n$ называется величина $x \cdot y$, удовлетворяющая аксиомам:

Scal1 $x \cdot y = \overline{y \cdot x}$ (эрмитова симметричность);

Scal2 $(x + y) \cdot z = x \cdot z + y \cdot z$ (дистрибутивность);

Scal3 $(ax) \cdot y = a(x \cdot y)$ (ассоциативность);

Scal4 $x \cdot x \geq 0$ (положительная определенность), причем $x \cdot x = 0$ тогда и только тогда, когда $x = 0$,

где $x, y, z \in L_n$, $a \in F$. Свойства Scal2 и Scal3 называются вместе свойством линейности по первому аргументу функции $x \cdot y$. Вместе со свойством Scal1 линейность дает билинейность, т. е. линейность по обоим аргументам.

Конечно-мерное вещественное векторное пространство с заданным на нем скалярным произведением называется **евклидовым пространством**.

Ненулевые векторы **ортогональны** ($x \perp y$), если их скалярное произведение равно нулю. **Длиной** вектора x называется величина $|x| \triangleq \sqrt{x \cdot x}$. **Расстоянием** (евклидовым) между векторами x и y называется длина их разности, т. е. $|x - y|$. Несложно проверить, что (евклидово) расстояние удовлетворяет аксиомам метрики (см. раздел 5.1.5).

В линейной алгебре важную роль играет ортонормированный базис (ОНБ). Это такой базис e_1, \dots, e_n , что для любой пары его элементов верно равенство: $e_i \cdot e_j = \delta_{i,j}$, где $\delta_{i,j} = [i = j]$ ¹⁸ — символ Кронекера. Таким образом, все векторы ОНБ попарно ортогональны и имеют длину 1.

¹⁸Нотация Айверсона $[\varphi]$ принимает значение 1, если φ истинно, и 0 — в противном случае.

Если векторы x и y разложить по ОНБ в виде $x = x_1\mathbf{e}_1 + \dots + x_n\mathbf{e}_n$, $y = y_1\mathbf{e}_1 + \dots + y_n\mathbf{e}_n$, то из свойств скалярного произведения мы получим:

$$x \cdot y = x_1\bar{y}_1 + \dots + x_n\bar{y}_n = x^T\bar{y},$$

Упражнение 3.32. Докажите с помощью Scal1–Scal3.

последнее означает произведение вектора-строки (в силу транспонирования) с координатами x на вектор-столбец с комплексно сопряженными координатами вектора y в смысле произведения матриц. Это уже привычное со школы определение скалярного произведения.

если не считать комплексного сопряжения и n штук координат

Как видим, скалярное произведение в такой алгебраической форме, на первый взгляд, зависит от конкретного ОНБ. Но чудо состоит в том, что на самом деле скалярное произведение численно не зависит от выбора ОНБ!

Дело в том, что переход от первого ОНБ ко второму ОНБ осуществляется с помощью линейного преобразования, матрица которого обладает свойством унитарности (в вещественном случае — ортогональности).

Матрица A размерности $n \times n$ называется **унитарной** (ортогональной), если

$$A\bar{A}^T = E \quad (AA^T = E),$$

где E — единичная матрица соответствующей размерности. То есть для унитарной матрицы обратная ищется транспонированием и сопряжением.

Пусть P — матрица перехода от ОНБ1 к ОНБ2. Известно, что P — унитарная матрица (это следствие того, что ее столбцы ортогональны и имеют длину 1, поскольку их можно рассматривать как векторы ОНБ2 с координатами в ОНБ1).

Пусть \tilde{x}, \tilde{y} — координаты векторов x, y в ОНБ2. Тогда, как мы видели выше, $\tilde{x} = P^{-1}x$, $\tilde{y} = P^{-1}y$. Тогда:

$$\tilde{x} \cdot \tilde{y} = (P^{-1}x)^T \overline{P^{-1}y} = x^T ((P^{-1})^T \bar{P}^{-1}) \bar{y} = x^T (\bar{P}P^T) \bar{y} = x^T E \bar{y} = x \cdot y,$$

т. е. скалярное произведение сохраняется при смене одного ОНБ на другой или, в интерпретации линейных преобразований, скалярное произведение инвариантно относительно унитарного (ортогонального) преобразования, т. е. такого преобразования, которому соответствует унитарная (ортогональная) матрица. Отсюда, в частности, следует одно замечательное геометрическое свойство линейных операторов: *расстояние между векторами сохраняется при унитарных (ортогональных) линейных преобразованиях.*

Архетип инварианта

Вообще, если какая-то функция биективна и сохраняет расстояния (при этом она может действовать из одного пространства в другое), то она называется **изометрией**. Для линейных операторов верно обратное утверждение: если линейный оператор сохраняет расстояние, то он является унитарным (ортогональным).

Архетип изоморфизма!

Действительно, сохранение расстояния означает сохранение длины, поскольку $|x|^2 = |x - 0|^2$, в то же время¹⁹

$$\begin{aligned} x \cdot y &= \\ &= \frac{1}{4}((x+y) \cdot (x+y) - (x-y) \cdot (x-y) + i(x+iy) \cdot (x+iy) - i(x-iy) \cdot (x-iy)) = \\ &= \frac{1}{4}(|x - (-y)| - |x - y| + i|x - (-iy)| - i|x - iy|). \end{aligned}$$

Если в это формуле всюду заменить x на $f(x)$ и y на $f(y)$ и воспользоваться линейностью f , то мы получим, что $f(x) \cdot f(y) = x \cdot y$, т. е. линейная изометрия сохраняет скалярное произведение.

Отсюда следует, что f переводит ОНБ в ОНБ, а отсюда уже следует свойство унитарности (ортогональности) оператора f .

Выше мы потребовали, чтобы изометрия f была линейным оператором. Очередное чудо состоит в том, что для унитарности f достаточно, чтобы оно оставляло 0 на месте и было изометрией, но при условии, что рассматриваемое пространство — вещественное. Этот факт следует из теоремы Мазура—Улама, и мы не откажем себе в удовольствии привести здесь ее доказательство.

Теорема 3.8 (Мазура—Улама). *Если биекция $f : X \rightarrow Y$ между нормированными пространствами является изометрией, то она является аффинным преобразованием, т. е.*

$$f(tx_1 + (1-t)x_2) = tf(x_1) + (1-t)f(x_2), \quad t \in (0, 1).$$

Доказательство. Для начала докажем, что

$$f\left(\frac{x_1 + x_2}{2}\right) = \frac{f(x_1) + f(x_2)}{2}. \quad (3.12)$$

Мы воспользуемся методом, предложенным Bogdan Nica в статье [109]. Для произвольных двух точек x_1 и x_2 и изометрии f определим следующую вещественную функцию:

$$D(f) = \left| f\left(\frac{x_1 + x_2}{2}\right) - \frac{f(x_1) + f(x_2)}{2} \right|.$$

Нетрудно найти следующую оценку (с помощью неравенства треугольника):

$$D(f) \leq \frac{1}{2} \left| f\left(\frac{x_1 + x_2}{2}\right) - f(x_1) \right| + \frac{1}{2} \left| f\left(\frac{x_1 + x_2}{2}\right) - f(x_2) \right| = \frac{|x_1 - x_2|}{2}, \quad (3.13)$$

¹⁹Формула приведена для комплексного пространства, для вещественного она существенно проще.

где мы воспользовались свойством изометрии $|f(a) - f(b)| = |a - b|$.

Пусть $h(x) = f(x_1) + f(x_2) - x$. Очевидно, что h — изометрия. Кроме того, f^{-1} также является изометрией. Поэтому функция

$$g(x, f) = f^{-1}(h(f(x)))$$

также является изометрией. Кроме того, $g(x_1, f) = x_2$ и $g(x_2, f) = x_1$. Тогда

$$D(g) = \left| f^{-1} \left(f(x_1) + f(x_2) - f \left(\frac{x_1 + x_2}{2} \right) \right) - \frac{x_1 + x_2}{2} \right|,$$

поскольку f^{-1} — изометрия, применим ее к разности внутри скобок:

$$D(g) = \left| f(x_1) + f(x_2) - f \left(\frac{x_1 + x_2}{2} \right) - f \left(\frac{x_1 + x_2}{2} \right) \right| = 2D(f).$$

Пусть теперь $g_1 = f$, $g_{n+1}(x) = g(x, g_n)$. Тогда получим $D(g_n) = 2^{n-1}D(f)$. И если $D(f) > 0$, то при некотором n мы получим $D(g_n) > |x_1 - x_2|/2$. Но оценка (3.13) справедлива для всех изометрий и не зависит от них (она равномерная), а зависит только от точек x_1 и x_2 . Противоречие.

Следовательно, $D(f) = 0$, что и доказывает (3.12).

Осталось доказать аффинность: $f(tx_1 + (1 - y)x_2) = tf(x_1) + (1 - t)f(x_2)$.

Упражнение 3.33. Предлагаем сделать это читателю самостоятельно. Но в качестве подсказки заметим, что любое вещественное t можно приблизить двоичными дробями:

$$\frac{k}{2^n} \leq t \leq \frac{k+1}{2^n},$$

Соответственно, сначала нужно показать свойство аффинности для коэффициентов вида $t = k/2^n$, а затем воспользоваться приближениями, зависящими только от n . \square

Из свойства аффинности непосредственно вытекает, что изометрия, оставляющая ноль на месте ($f(0) = 0$) является линейным оператором. Действительно, полагая $x_2 = 0$, получаем, что $f(tx_1) = tf(x_1)$, где $t \in (0; 1)$. Если $t > 1$, то положим $y = tx_1$, тогда $f(y/t) = f(y)/t$, что соответствует $f(tx_1) = tf(x_1)$. Далее, $f(x_1 + x_2) = 2f((x_1 + x_2)/2) = f(x_1) + f(x_2)$. Наконец, в случае $t = -1$ имеем $f(-x) = f(x) + f(-x) - f(x) = f(x - x) - f(x) = -f(x)$.

Таким образом, как аддитивность, так и однородность f имеют место быть, т.е. f — линейный оператор. Наконец, из того, что f — линейная изометрия, по доказанному, получаем, что f есть ортогональный оператор. Таким образом, доказана

Теорема 3.9. *Если f — изометрия, сохраняющая ноль, действует над вещественным пространством, то f — ортогональный оператор.*

Заметим, что в комплексных пространствах эта теорема неверна. Например, отображение $z \mapsto \bar{z}$ сохраняет ноль и расстояния, но не является линейным, поскольку $iz \mapsto -i\bar{z}$, т. е. нарушается однородность.

Тем не менее, как ортогональные, так и, в превосходящей степени, унитарные операторы играют важную роль в линейной алгебре, в частности, при решении линейных однородных дифференциальных уравнений.

Множество всех унитарных (ортогональных) матриц образует подгруппу в группе $GL(n)$. Эта подгруппа обозначается $U(n)$ в комплексном случае и $O(n)$ — в действительном. Столбцы и строки унитарной матрицы образуют ОНБ.

Легко видеть, что $|\det A| = 1$, если A — унитарная (ортогональная) матрица, кроме того, такие матрицы соответствуют операторам поворотов и инверсии относительно одной из осей пространства L_n .

Частный случай унитарной матрицы — **специальная унитарная матрица** (соответственно, **специальная ортогональная матрица**), определяемая условием $\det A = 1$ (без модуля). Специальные унитарные матрицы имеют применение, например, в квантовой механике и физике элементарных частиц.²⁰ Условие $\det A = 1$ отсекает инверсии в $U(n)$ ($O(n)$), и, таким образом, оставляет только повороты, которые также образуют группу. В комплексном случае эта группа обозначается $SU(n)$, в действительном — $SO(n)$, и называется группой вращений.

Группы $U(n)$, $O(n)$, $SU(n)$, $SO(n)$ являются классическими представителями групп Ли [53, 54].

Комментарий 8. О пользе многомерных пространств.

Рассматривая линейные пространства, мы, как правило, изучаем \mathbb{R}^n или \mathbb{C}^n , т. е. немного обобщенную школьную геометрию в ее координатном аналитическом представлении. Оказывается, что эти многомерные пространства, хоть и не связаны напрямую с физическим миром, тем не менее, весьма востребованы в различных аналитических и оптимизационных задачах.

Действительно, многие жизненные и технологические процессы описываются огромным количеством физических параметров (например, медицинские показатели или характеристики автомобиля). Это значит, что мы живем в многомерном мире, где лишь 4 измерения отвечают за пространство и время.

Умение быстро и качественно анализировать многомерные большие массивы данных (BigData) — ключевая задача современных вычислительных систем, программистов и аналитиков данных. Во многом она опирается на линейную алгебру.

Рассмотрим простейшую задачу **машинного обучения** (Machine Learning).

²⁰Специальная унитарная группа $SU(3)$ лежит в основе одного из фундаментальных взаимодействий — квантовой хромодинамики.

Пусть имеется массив векторов (в общем случае n -мерных), которые описывают некоторые события (например, данные о погоде, дорожной ситуации, геометрии дорожного полотна, дорожных знаках) и соответствующих им откликов (t различных классов откликов, так что мы имеем таблично заданную частичную (!) функцию $\mathbb{R}^n \rightarrow \{1, \dots, t\}$). Это — известные данные и полученные отклики. Пусть далее мы встречаемся с новым набором данных (например, датчики беспилотного автомобиля их считывают с интервалом 0.1 секунды, по крайней мере, все, которые связаны с дорогой, разметкой, знаками).

Вопрос — какой должен быть отклик (от 1 до t) на эти новые данные?

Это так называемая классическая **задача машинного обучения с учителем**. Система имеет некий тренировочный набор (например, запись процесса вождения автомобиля профессиональным водителем), на котором «учится» прогнозировать нужные отклики («рулить» самостоятельно).

t действий в таких задачах принято обозначать различными цветами, так что получаем, что каждая точка из заданного набора векторов пространства \mathbb{R}^n имеет цвет, один из t заданных цветов.

Это не задача о раскраске плоскости!

Предположим, что $n = 2 = t$, то есть событие отображается точкой на плоскости, а обучающий отклик на событие обозначается одним из двух цветов — красным или синим. На рис. 3.2 событие (точка) имеет два параметра, лежащие в отрезке $[0; 1]$: неровность дороги (bumpiness) и уклон дороги (grade). Ответ системы: ехать быстро (fast, синий) или ехать медленно (slow, красный). Видно, что около нуля разрешается ехать быстро, а около точки $(1; 1)$ — медленно. Здесь обучающая картинка такова, что можно на глаз примерно определить линию разграничения, выше которой почти все точки будут красные, ниже — синие. Имея такую линию, построенную по эмпирическим данным, мы можем предсказывать (рекомендовать) отклик системы на произвольное событие: если точка оказалась выше линии — нужно ехать медленно, если ниже — можно быстро.

Разграничивающая два класса точек линия (а в n -мерном случае это $n - 1$ -мерная поверхность) называется **поверхностью решения** (decision surface) и является **моделью** системы, в которой происходит обучение. Здесь мы видим переход от конечных эмпирических данных к континуальной (абстрактной) модели.

Существует несколько методов построения поверхности решения.

LM1 Наивный байесовский классификатор (Naïve Bayes)

LM2 Метод k ближайших соседей.

LM3 Метод опорных векторов (SVM).

LM4 Деревья решений.

LM5 Метод случайных деревьев.

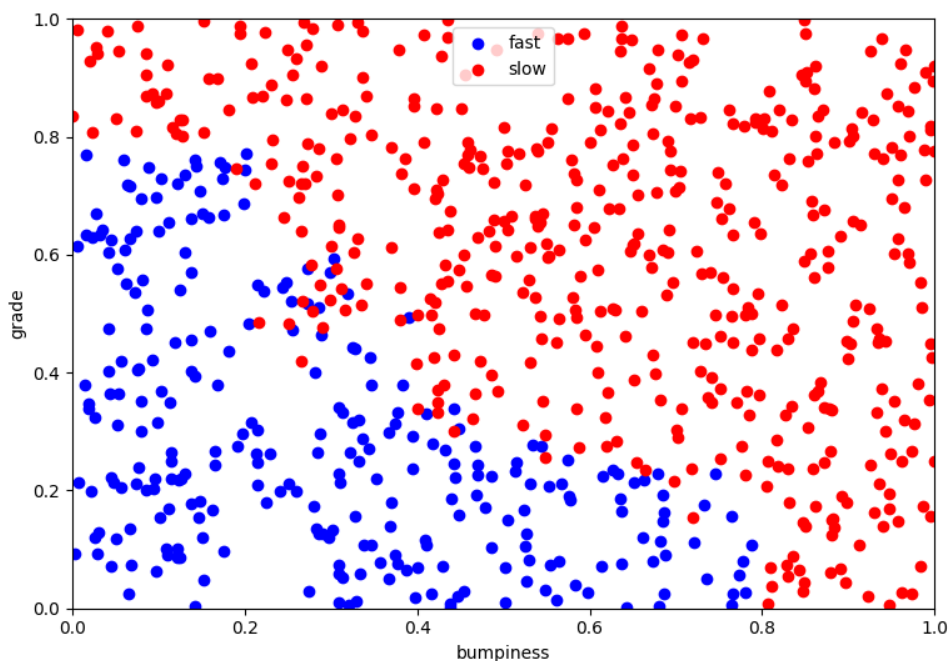


Рис. 3.2: Обучающий набор.

В этих методах заранее предполагается аналитический класс поверхности решения. Как правило рассматриваются кусочно-линейные поверхности (т.е. составленные из кусков гиперплоскостей, в двумерном случае — ломаная линия) или полиномы. Вычислительная простота формулы поверхности — одно из главных требований (представьте, что автомобиль не успеет сообразить, что нужно затормозить на красный свет!), а задача отыскания поверхности решения — это примерно то же самое, что задача о регрессии в статистике. Только здесь мы ищем не аппроксимирующую, а разделяющую функцию.

Случай t красок сводится к t задачам с двумя красками: для каждой краски мы рассматриваем дилемму «точка заданного цвета или какого-то другого?» Это называется сравнением «один против всех» (one-vs-all) или «один против остальных» (one-vs-rest).

Ближе всего к линейной алгебре стоит метод SVM. В этом методе мы ищем гиперплоскость, «наилучшим образом» разделяющую два класса точек. Пусть обучающий набор состоит из точек:

$$X = \{(x_1, c_1), \dots, (x_p, c_p)\}, \quad x_i \in \mathbb{R}^n, c_i \in \{-1, 1\},$$

где c_i отвечает за цвет (класс) точки. Предполагается также, что этот набор нормализован так, чтобы в нем не было значительных выбросов (слишком большие

векторы x_i могут испортить классификацию, это хорошо известно в статистике).

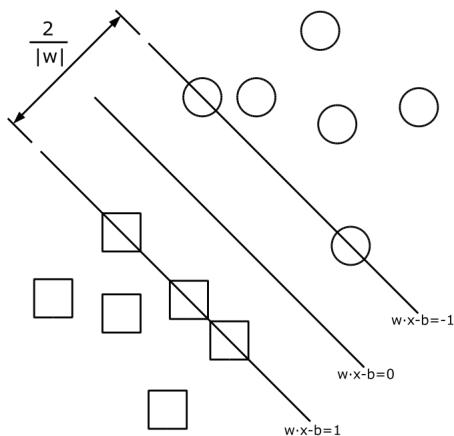
Далее производится поиск поверхности решения в виде плоскости, заданной уравнением

$$w \cdot x - b = 0,$$

где w — вектор, ортогональный к данной плоскости. Параметр $b/|w|$ равен расстоянию от плоскости до начала координат.

Считаем, что плоскость «наилучшим образом» разделяет два класса точек, если она равноудалена от ближайших к ней точек выборки X . Такие точки и называются **опорными векторами**. Смысл в том, что все остальные точки лежат еще дальше от поверхности решения, чем опорные векторы, а сама поверхность решения ищется примерно посередине между ними.

Если такая оптимальная плоскость существует, то выборка X называется **линейно разделяемой**. Для линейно разделяемой выборки задача поиска поверхности решения сводится к минимизации вектора w при условиях



$$w \cdot x_i - b \geq 1 \text{ при } c_i = 1, \quad w \cdot x_i - b \leq -1 \text{ при } c_i = -1.$$

Это — задача квадратичной оптимизации, которая решается методом Лагранжа.²¹

В этом случае мы гарантируем, что между двумя классами точек (определяемыми числами $c_i = \pm 1$) существует разделительная полоса (пространство между двумя параллельными гиперплоскостями), не содержащая (кроме как на границе) точек выборки X , а минимальность w обеспечивает максимально возможную ширину такой полосы, которая равна $2/|w|$. При этом поверхность решения представляет собой плоскость, находящуюся ровно посередине данной полосы.

В случае линейной неразделимости (как на рис. 3.2) метод остается тот же, но вводятся поправки, допускающие минимально возможные ошибки. Точнее, ограничительные неравенства теперь записываются следующим образом:

$$w \cdot x_i - b \geq 1 - \varepsilon_i \text{ при } c_i = 1, \quad w \cdot x_i - b \leq -1 + \varepsilon_i, \quad \varepsilon_i \geq 0.$$

²¹Точнее, методом Лагранжа решается двойственная к ней задача!

При этом в целевом функционале эти ошибки также учитываются: нужно минимизировать величину

$$|w|^2 + c(\varepsilon_1 + \dots + \varepsilon_n).$$

После нахождения поверхности решения некоторые точки одного или обоих классов обучающей выборки могут оказаться не с той стороны гиперплоскости, где находится основная масса точек того же класса, в чем и заключается неточность метода. Ясно, что на практике, как правило, мы имеем дело именно с таким случаем.

Далее производится проверка эффективности найденного решения, а именно, рассчитывается процент верной классификации точек обучающей выборки и тестовой выборки.²² Если он достаточно близок к 1, то такую модель можно использовать для классификации новых данных.

Когда модель найдена (т. е. найдены w и b) и проверена на тестовых данных, ею можно пользоваться для классификации новых данных. Для произвольной новой точки x вычисляется значение $w \cdot x - b$. Если это значение больше нуля, то точка относится к классу $c = 1$, а если меньше — к классу $c = -1$.

Если классов t штук, то строится t моделей (по правилу «один против всех») и для новой точки вычисляются значения $w_1 \cdot x - b_1, \dots, w_p \cdot x - b_p$, причем значение $c_i = 1$ означает попадание в i -ый класс. Далее находим такое i , при котором значение $w_i \cdot x - b_i$ будет максимальным и положительным, и тогда точка x относится к i -му классу. Если же все значения $w_i \cdot x - b_i$ оказались отрицательными, то точку относят к тому классу, к чей поверхности решения она ближе.

Описанный выше метод построения поверхности решения относится к линейным SVM. Однако существует метод (т.н. *kernel trick*), в котором скалярное произведение $w \cdot x$ заменяется произвольной (достаточно хорошей, гладкой) функцией $k(w, x)$, где w — вектор параметров, по которым производится оптимизация модели. Функции $k(w, x)$ называются **ядрами** модели SVM.²³

Постановка самой задачи остается прежней — минимизация функционала $k(w, x)$ при условиях, что для точек одного класса должно быть $k(w, x) > 1$, а для остальных точек $k(w, x) < -1$. В зависимости от сложности ядра разрешающая поверхность может выглядеть сколь угодно сложным образом и разделять два класса из обучающей выборки на все 100%.

Чаще всего используются полиномиальные и экспоненциальные ядра. Заметим, что ядерный метод SVM может повышать размерность исходного пространства наблюдений. Например, путем введения дополнительных зависимых

²²Обычно известные данные сразу же разделяются на обучающие и тестовые в соотношении примерно 1 к 3 или 1 к 4 случайным образом. Большая часть служит для обучения машины, а меньшая — для тестирования. Если машина ошибается в 2–3 процентах случаев как на обучающей, так и на тестовой выборке, то настройку можно считать удачной.

²³Обычно предполагается, что ядро симметрично по своим аргументам-векторам и неотрицательно определено, т. е. $k(x, x) \geq 0$.

размерностей. Так, если на плоскости картинка представляется нам не вполне удачной для линейного SVM, мы можем попробовать ввести третье измерение по формуле $x_3 = x_1x_2$, т.е. перемножить две исходных координаты события. В итоге вместо событий на плоскости (x_1, x_2) мы будем иметь события в пространстве (x_1, x_2, x_3) . Правильно выбирая зависимые измерения, мы можем свести исходную сложную задачу классификации на плоскости к линейной SVM в пространстве.

Более подробно о задачах машинного обучения вместе с многочисленными примерами кода на Python можно ознакомиться в [47, 48], а конкретно о методе SVM — в книге [44].

3.6 Экскурс в геометрию

В этом разделе мы рассмотрим некоторые задачи геометрии, не выходя за рамки уже определенных выше пространств. Речь идет о первых степенях \mathbb{R} (от 1-ой до 4-ой), включая дополнительные структуры вроде комплексных чисел и кватернионов, а также различные преобразования (функции) над ними.

Наша цель состоит прежде всего в анализе основных геометрических идей и методов в рамках т.н. Эрлангенской программы Феликса Клейна (1872).

Суть этой программы состоит в том, чтобы рассматривать геометрию как целостную науку о преобразованиях пространств и возникающих при этом инвариантах.

При этом под пространствами обычно понимаются векторные или топологические пространства, а также их алгебраические надстройки, под преобразованиями — группы функций со специальными свойствами (вроде группы линейных операторов), а под инвариантами — свойства подмножеств этих пространств (скажем, длины или углы геометрических фигур на плоскости).

В первом приближении можно выделить следующий список инвариантов, расположенный по возрастанию степени общности:

- Inv1 Сохранение длин (отношение нулевого порядка) — при движениях (изометриях);
- Inv2 Сохранение отношений длин (отношение первого порядка) — при подобиях (гомотетиях);
- Inv3 Сохранение двойного отношения (отношение второго порядка) — при проективных преобразованиях;
- Inv4 Сохранение операций/отношений — при изоморфизмах;



*Феликс
Христиан
Клейн*

Inv5 Сохранение связности — при гомеоморфизмах;

Inv6 Сохранение мощности множества — при биекциях.

Понятие инварианта тесно связано с архетипом редукции. Дело в том, что инвариант — это по определению *неизменная величина*. Мы видим какую-то сложную конструкцию, воздействуем на нее некоторыми преобразованиями, об инвариантах которых нам что-то известно, получаем в результате некоторую упрощенную (редуцированную) конструкцию и делаем выводы об ее свойствах. Затем, «отматывая назад» проделанные преобразования, восстанавливаем исходную конструкцию, понимая на основе инвариантов ее свойства.

Для пояснения мы можем воспользоваться тем примером, который уже был в разделе о комплексных числах. Аэродинамическая задача о воздушных потоках в плоскости сечения крыла самолета сводится преобразованием Жуковского к задаче о воздушных потоках, обтекающих круглый цилиндр (окружность в сечении). Таким образом, производя вычисления вихрей и нагрузок с окружностью (что намного проще с вычислительной и аналитической точки зрения) и понимая, что зависимости инвариантны относительно преобразования Жуковского, мы рассчитываем траектории в более простом случае, а затем переносим их на реальное крыло с помощью преобразования $w = z + 1/z$.

Геометрический пример (см. рис. 3.3). Дана окружность и точка A вне ее (в одной плоскости α), задача: провести касательные AB и AC к окружности, используя только линейку. Иначе говоря, необходимо путем простых построений получить на окружности точки касания этих прямых, а затем обосновать, что это действительно точки касания.

Само построение довольно простое. Мы берем линейку и проводим три произвольные прямые k, l, m так, чтобы они пересекали окружность. Получаем 6 точек пересечения, через которые строим два креста внутри образовавшихся криволинейных четырехугольников. Центральные точки этих крестов обозначены E и F . Через точки E и F проводим прямую до пересечения с окружностью и получаем точки B и C . Фокус в том, что это и есть искомые точки касания!

Вопрос: как это доказать?

Оказывается, что можно нашу рабочую плоскость α спроецировать на некоторую другую плоскость β пространства (проецирование осуществляется путем проведения прямых линий из некоторой выбранной в пространстве точки O — центра проекции) так, что точка A , из которой нужно провести касательные, уйдет на бесконечность, и задача сведется к тому, чтобы провести параллельные касательные к окружности, а это сделать намного проще (требуется построить диаметр — и точки касания готовы)! И поскольку проективные преобразования сохраняют точки касания, а прямые при этом

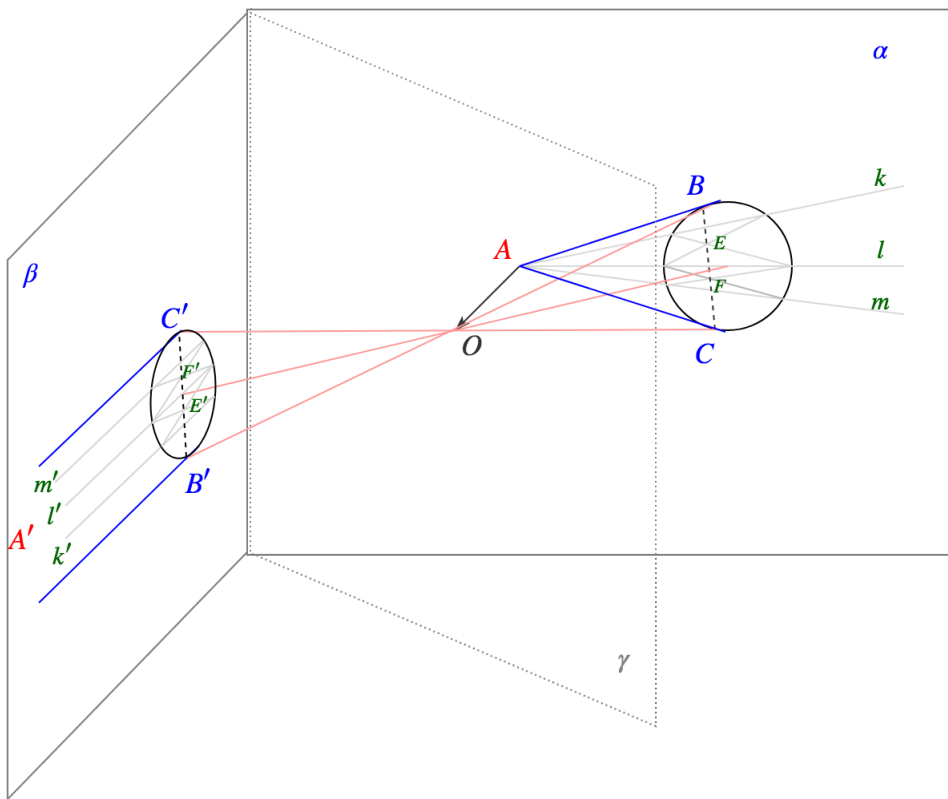


Рис. 3.3: Построение касательных одной линейкой.

переходят в прямые, то можно в два шага получить нужное построение в исходной конфигурации.

На рис. 3.3 показана вторая плоскость β , на которой также нарисована окружность, а в качестве центра проекции O выбрана середина отрезка, соединяющего центры этих окружностей. Точка O лежит в биссектральной плоскости γ , делящей угол между α и β пополам. Варьируя этот угол можно добиться того, что отрезок AO будет параллелен плоскости β . После чего начинаем проецировать изображение с плоскости α на плоскость β с помощью прямых, проходящих через точку O . При этом у нас нештрихованные точки переходят в штрихованные, причем прямые k' , l' , m' становятся параллельными, и та же самая конструкция построения крестов с центрами в точках E' и F' становится симметричной, так что в ней обосновать, что $B'C'$ является диаметром, уже не составит труда.

Этот пример и многое в нашем дальнейшем изложении в разделе о геометрии заимствованы из открытых видеолекций д.ф.-м.н. А. Савватеева «[Неевклидова геометрия](#)» [0oKpduJy1bM] и «[Геометрия и группы](#)» [h9TB9PTKvXU].

Кроме того, настоятельно рекомендуем ознакомиться с книгой В. В. Прасолова и В. М. Тихомирова «Геометрия» [88].

3.6.1 Движения в действительном пространстве

Рассмотрим для начала то, что называется *движением* или *изометрией*, т. е. такие преобразования прямой, плоскости и пространства, которые сохраняют расстояния.²⁴ Сюда же присоединим движения окружности и сферы.

Выше была приведена фундаментальная теорема Мазура–Улама и следствие из нее о том, что любая изометрия действительного пространства, сохраняющая точку 0 на месте, является линейным ортогональным оператором, т. е. принадлежит группе $O(n)$, где n — размерность пространства. Кроме того, было замечено, что у таких операторов модуль определителя равен 1, а значит, в действительном пространстве имеем либо $\det = 1$ (группа $SO(n)$), либо $\det = -1$.

Таким образом, напрашивается естественная (с теоретико-групповой точки зрения) классификация движений:

T1 Движения, являющиеся специальными линейными ортогональными операторами ($\det = 1$),

T2 Движения, являющиеся линейными ортогональными операторами с $\det = -1$,

T3 Движения, не сохраняющие точку 0 на месте.

Последний случай можно свести к первым двум тривиально. Пусть движение T таково, что $T(0) = b$. Тогда рассмотрим новое движение $T_1(x) = T(x) - b$. Оно будет изометрией, сохраняющей 0 на месте. Стало быть, T есть композиция движений первого или второго типа плюс константа (параллельный перенос). Если бы мы вели речь не об \mathbb{R}^n , а об аффинном пространстве, в котором все сонаправленные векторы равной длины отождествляются, то третий случай был бы вовсе исключен.

Движения прямой \mathbb{R}^1

Это легко продемонстрировать в самом простом случае — при $n = 1$, т. е. на прямой. Всякое движение там либо тождественно ($y = x$, где x — исходная координата, а y — результат применения движения к x), либо является симметрией относительно нуля ($y = -x$), либо сдвиг на константу одного из двух предыдущих ($y = \pm x + b$).

²⁴Пространства более высокой размерности мы оставляем читателю для самостоятельного ознакомления.

Если же мы определим симметрию относительно *произвольной* точки ($y = -(x - 2c)$), то все движения можно описать через симметрии: тождественное движение есть две одинаковые симметрии подряд, сдвиг — две разные симметрии, сдвиг с симметрией — симметрия со смещенным центром. Иначе говоря, все движения при $n = 1$ можно получить как композицию одной или двух симметрий на прямой (с различными, вообще говоря, центрами). Предлагаем читателю убедиться в этом самостоятельно.

Упражнение
3.34.

Движения плоскости \mathbb{R}^2

При $n = 2$ картина становится интереснее. Матрица специального линейного ортогонального оператора имеет вид

$$R_\phi^2 = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix}$$

Такой оператор осуществляет поворот плоскости вокруг нуля на угол ϕ в положительном направлении. Кроме того, если вспомнить представление комплексного числа в виде матрицы, то оказывается, что все такие повороты задаются умножением на комплексное число $z = \cos \phi + i \sin \phi$, т. е. на комплексное число с единичной окружности. Так что вектор $x + iy$ переходит при данном повороте в вектор $(x + iy)z = (x \cos \phi - y \sin \phi) + i(x \sin \phi + y \cos \phi)$, например, при $\phi = \pi/2$ вектор $1 + 0 \cdot i$ переходит в $0 + i \cdot 1$.

Кроме того, матрица линейного ортогонального оператора с отрицательным определителем на плоскости имеет вид:

Это
называется
инволюцией.

$$\begin{pmatrix} \cos \phi & \sin \phi \\ \sin \phi & -\cos \phi \end{pmatrix} = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Иначе говоря, остальные операторы группы $O(2)$, которые не входят в $SO(2)$, можно получить умножением специальных операторов на матрицу

$$\hat{E} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Это соответствует случаю $y = -x$ на прямой.

Матрица \hat{E} примечательна тем, что ее групповая степень равна 2, т. е. $\hat{E}\hat{E} = E$, а множество $\{E, \hat{E}\}$ является нормальной подгруппой в $O(2)$, фактор по которой изоморфен $SO(2)$. Кроме того, данная матрица, как оператор, осуществляет отражение (симметрию) комплексной плоскости относительно действительной оси: i переходит в $-i$ и т.д. Так что все движения, сохраняющие 0 и не являющиеся поворотами, получаются как композиция симметрии относительно действительной оси и поворота.

Таким образом, все движения плоскости, сохраняющие 0, являются либо поворотами ($SO(2)$), в число которых входит и поворот на нулевой угол, т. е.

id, либо композицией симметрии \hat{E} с поворотом. Очевидно (из теоретико-групповых соображений!), верно и обратное: все движения плоскости, сохраняющие 0, являются либо симметриями, либо композицией симметрии \hat{E} с какой-то другой симметрией.

Действительно, симметрия относительно оси l_ψ , проходящей через 0 и имеющей угол наклона ψ к положительной вещественной оси, имеет матрицу оператора инволюции

$$S_{l_\psi} = \begin{pmatrix} \cos 2\psi & \sin 2\psi \\ \sin 2\psi & -\cos 2\psi \end{pmatrix} = R_{2\psi}^2 \hat{E},$$

т.е. произвольная осевая симметрия, оставляющая 0 на месте, с углом наклона ψ представляется как композиция симметрии \hat{E} и поворота на угол 2ψ .

Обратно:

$$R_\phi^2 = \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix} = \begin{pmatrix} \cos(2\phi/2) & \sin(2\phi/2) \\ \sin(2\phi/2) & -\cos(2\phi/2) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = S_{l_{\phi/2}} \hat{E},$$

т.е. поворот на угол ϕ есть композиция отражения относительно действительной оси и отражения относительно оси с углом наклона $\phi/2$.

Все эти рассуждения полностью описывают движения окружности S^1 с центром в нуле. В некотором смысле они очень похожи на движения прямой. Действительно, сдвиг прямой соответствует повороту окружности (разница лишь в диапазоне параметра: на прямой это $(-\infty; +\infty)$, на окружности — $[0; 2\pi)$), а отражение прямой относительно произвольной точки соответствует симметрии окружности относительно произвольной прямой, проходящей через 0 (диапазоны параметра, соответственно, $(-\infty; +\infty)$ и $[0; \pi)$).

Наконец, рассмотрим случай, когда движение T не сохраняет точку 0 на месте. Тогда движение $T(x) - T(0)$, которое отличается от исходного параллельным переносом на вектор $T(0)$, сохраняет 0 и, следовательно, является одним из рассмотренных выше. Но параллельный перенос также можно выразить как композицию двух симметрий, только уже относительно смещенных прямых на плоскости (сравните со случаем симметрии со сдвигом на прямой).

Действительно, рассмотрим на плоскости отрезок $[0; T(0)]$. Затем выполним симметрию относительно его срединного перпендикуляра, затем симметрию относительно прямой, проходящей через $T(0)$ и параллельной срединному перпендикуляру. В итоге мы получим смещение на вектор $T(0)$ с помощью двух отражений. Следовательно, исходное движение T есть композиция двух параллельных симметрий.

Таким образом, аналогично одномерному случаю, всякое движение плоскости является либо симметрией (относительно какой-то оси, не обязательно

проходящей через 0), либо композицией двух симметрий, либо композицией трех симметрий (они же образуют симметрию с параллельным переносом, они же называются скользящей симметрией). Это есть содержимое известной **теоремы Шаля**.

Теорема Шаля выводится также из следующей леммы, доказательство которой мы предлагаем осуществить читателю самостоятельно, не основываясь на матрицах и алгебре пространства \mathbb{R}^2 — одними геометрическими методами.

Лемма 3.5 (о трех гвоздях). *Если движение оставляет на месте*

- 1) *три точки плоскости, не лежащие на одной прямой, то это тождественное преобразование (id),*
- 2) *иначе, если две точки, то это симметрия относительно оси, проходящей через эти точки,*
- 3) *ровно одну точку, то это поворот вокруг данной точки.*

Во втором случае используется программистская конструкция `elseif`, поскольку симметрия оставляет на месте не ровно 2 точки, а целую прямую точек, но никак не треугольник.

Итак, мы видим, что все движения плоскости мы можем классифицировать двумя способами. Первый (алгебраический) представляет любое движение как композицию поворота из группы $SO(2)$, отражения \hat{E} и параллельного переноса вдоль действительной оси. Топологию движений, таким образом, можно представить как цилиндр $S^1 \times \mathbb{R}$ с двумя поверхностями — внутренней и внешней (переход от одной к другой определяется отражением).

Второй (геометрический) представляет все движения либо как композиции одно-, двух- и трехкратных симметрий, либо как один из видов движения: параллельный перенос, поворот или скользящая симметрия. Заметим, что скользящая симметрия определяется тоже как композиция параллельного переноса и симметрии относительно оси этого переноса (в этом случае она коммутирует с переносом), так что она в себе содержит и простую симметрию (при переносе на нулевой вектор).

Дадим следующую таблицу композиций движений плоскости.

	\Rightarrow	\circlearrowleft	\nwarrow
\Rightarrow	\Rightarrow	\circlearrowleft	\nwarrow
\circlearrowleft	\circlearrowleft	\Rightarrow или \circlearrowleft	\nwarrow
\nwarrow	\nwarrow	\nwarrow	\Rightarrow или \circlearrowleft

Таблицу следует читать следующим образом. Символ \Rightarrow обозначает класс движений плоскости, являющихся параллельными переносами на произвольный вектор (в том числе нулевой), символ \circlearrowleft — класс движений, являющихся

поворотами относительно произвольного центра на произвольный угол, символ \leftarrow — класс движений, осуществляющих скользящую симметрию (сдвиг на произвольный вектор и отражение относительно оси данного вектора).

Композиция одного типа движения (левый столбец) с тем же или другим типом (верхняя строка) дает один из этих же типов движений (соответствующая ячейка).

Покажем, например, что два произвольных поворота — это либо параллельный перенос, либо поворот. Введем следующие обозначения. Пусть $z(\phi) = \cos \phi + i \sin \phi$, для точек A и B разность $A - B$ обозначает вектор \vec{BA} . Запомним также, что $z(\phi)^{-1} = z(-\phi)$. Далее, пусть первый поворот имеет центр O_1 и угол ϕ , второй поворот — центр O_2 и угол ψ . Предположим, что при последовательном применении данных поворотов точка X остается на месте. Тогда, учитывая что поворот вокруг нуля описывается умножением на комплексное число длины 1, получаем:

$$((X - O_1)z(\phi) + O_1 - O_2)z(\psi) + O_2 = (X - O_1) + O_1,$$

откуда легко найти, что

$$X = \frac{z(-\psi) - 1}{z(\phi) - z(-\psi)}(O_1 - O_2) + O_1.$$

Это выражение имеет смысл тогда и только тогда, когда $z(\phi) \neq z(-\psi)$, т.е. когда углы поворотов не компенсируют друг друга. При равенстве $z(\phi) = z(-\psi)$ компенсация углов приводит к тому, что их композиция является параллельным переносом. Во всех остальных случаях точка X находится однозначно и является центром результирующего поворота.

Чтобы найти угол нового поворота, отследим, куда при этом переходит точка O_1 относительно точки X :

$$\begin{aligned} \frac{(O_1 - O_2)z(\psi) + O_2 - X}{O_1 - X} &= \frac{(O_1 - O_2)z(\psi) - ((X - O_1)z(\phi) + O_1 - O_2)z(\psi)}{O_1 - X} = \\ &= z(\phi)z(\psi) = z(\phi + \psi). \end{aligned}$$

Таким образом, композиция двух поворотов приводит к новому повороту на суммарный угол. Аналогично можно получить остальные клетки таблицы композиций.

Этим полностью исчерпывается описание группы движений плоскости.

3.6.2 Движения в \mathbb{R}^3 , кватернионы

Огромная значимость комплексных чисел и их многочисленные применения заставили математиков искать их обобщение на трех- и более мерные пространства \mathbb{R}^n . Решение обнаружилось в 4-мерном пространстве (двумерном комплексном), это так называемые **числа Гамильтона** или **кватернионы**.

Выше мы уже определяли эти числа с помощью матриц специального вида над полем комплексных чисел (см. раздел 3.2), так что мы обладаем достаточным инструментарием для работы с кватернионами.

Итак, матричное представление кватернионов:²⁵

$$Q(z, w) = \begin{pmatrix} z & -w \\ \bar{w} & \bar{z} \end{pmatrix}, \quad z, w \in \mathbb{C}.$$

Аналогично определению комплексных чисел вектор (z, w) записывается как алгебраическое выражение $z + wj$, где j — новая мнимая единица, добавляющая одно комплексное измерение. При этом сложение таких векторов производится покомпонентно, а умножение почти так же, как умножение комплексных чисел:

$$(z + wj)(z' + w'j) = (zz' - w\bar{w}') + (zw' + \bar{z}'w)j,$$

эта формула интересна тем, что она расширяет соответствующую формулу умножения комплексных чисел, добавляя комплексное сопряжение в двух местах.

Таким образом, мы можем рассматривать кватернионы как двумерную (некоммутативную) алгебру над полем \mathbb{C} .

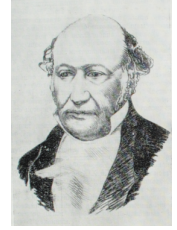
Полагая далее, что $z = a + bi$ и $w = c + di$, запишем $z + wj$ в виде 4-компонентного вектора

$$(a, b, c, d) = a + bi + cj + dk,$$

где $k = ij$ — еще одна новая мнимая единица, и $a, b, c, d \in \mathbb{R}$. Из определенного выше умножения (построенного на основе умножения матриц) нетрудно получить таблицу умножения мнимых единиц:

1	i	j	k
1	i	j	k
i	-1	k	-j
j	-k	-1	i
k	j	-i	-1

²⁵ Иногда определяют матрицей, где не только сопряжение, но и знак минус перед w стоит на нижней строке, но мы его наследовали от матричного представления комплексных чисел, оставив наверху. Тем более, что это никак не влияет на определение.



Уильям
Роуэн
Гамильтон

Упражнение
3.35.
Проверьте,
что это
соответ-
ствует
умножению
матриц
 $Q(z, w)$ и
 $Q(z', w')$

Заметим, что умножение здесь перестало быть коммутативным (кватернионы коммутируют только с действительными числами). Множество \mathbb{R}^4 с такими операциями сложения и умножения называется пространством кватернионов или **гамильтоновых чисел** и обозначается \mathbb{H} .

Наконец, пользуясь блочным представлением матриц и тем, что комплексные числа тоже представляются матрицами, мы можем от матрицы $Q(z, w)$ перейти к матрице 4×4 и представить кватернион $q = a + bi + cj + dk$ в следующем виде:

$$q = \begin{pmatrix} z & -w \\ \bar{w} & \bar{z} \end{pmatrix} = \begin{pmatrix} a & -b & -c & d \\ b & a & -d & -c \\ c & d & a & b \\ -d & c & -b & a \end{pmatrix} \quad (3.14)$$

Проверьте, что умножение таких матриц в точности дает тот же результат, что умножение матриц с комплексными числами вида $Q(z, w)$. Таким образом, кватернионы можно рассматривать как 4-мерную алгебру над полем \mathbb{R} . | Упражнение 3.36.

Итак, мы видим уже 4 ипостаси кватернионов: они представляются как

- двумерная (некоммутативная) алгебра над полем \mathbb{C} ;
- четырехмерная (некоммутативная) алгебра над полем \mathbb{R} ;
- подкольцо кольца матриц 2×2 над \mathbb{C} ;
- подкольцо кольца матриц 4×4 над \mathbb{R} .

Заметим, что определитель матрицы, представляющей комплексное число, равен норме этого комплексного числа:

$$\det \begin{pmatrix} a & -b \\ b & a \end{pmatrix} = a^2 + b^2 = |z|^2.$$

Аналогично определим и норму кватерниона:

$$|q|^2 \Rightarrow \det \begin{pmatrix} z & -w \\ \bar{w} & \bar{z} \end{pmatrix} = |z|^2 + |w|^2 = a^2 + b^2 + c^2 + d^2.$$

Это объясняет, почему при матричном определении кватернионов на второй строке стоят комплексно сопряженные числа. Кроме того, так определенная норма кватерниона является евклидовой нормой (квадратом длины) вектора (a, b, c, d) .

Далее, мы хотим, чтобы сопряжение кватернионов было согласовано с их представлением через комплексные числа, а также с вычислением нормы по

формуле $|q|^2 = q\bar{q}$. Для этого заметим, что транспонированная матрица комплексного числа является матрицей сопряженного к нему числа:

$$z = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}, \quad \bar{z} = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$$

Введем аналогичное определение для кватернионов, добавив к транспонированию комплексное сопряжение:

$$q = \begin{pmatrix} z & -w \\ \bar{w} & \bar{z} \end{pmatrix}, \quad \bar{q} = \overline{\begin{pmatrix} z & \bar{w} \\ -w & \bar{z} \end{pmatrix}} = \begin{pmatrix} \bar{z} & w \\ -\bar{w} & z \end{pmatrix}$$

Нетрудно видеть, что в результате мы получим число $\bar{z} - wj$, которое в действительных числах записывается как $a - bi - cj - dk$, то есть, если представить кватернион q как матрицу 4 порядка из действительных чисел, то матрица сопряженного числа \bar{q} будет просто транспонированной матрицей q (в полной аналогии с матрицами для комплексных чисел). При этом определитель такой матрицы будет равен $|q|^4$, т. к.

$$\begin{pmatrix} a & -b & -c & d \\ b & a & -d & -c \\ c & d & a & b \\ -d & c & -b & a \end{pmatrix} \begin{pmatrix} a & b & c & -d \\ -b & a & d & c \\ -c & -d & a & -b \\ d & -c & b & a \end{pmatrix} = \begin{pmatrix} |q|^2 & 0 & 0 & 0 \\ 0 & |q|^2 & 0 & 0 \\ 0 & 0 & |q|^2 & 0 \\ 0 & 0 & 0 & |q|^2 \end{pmatrix}$$

Определитель матрицы справа, с одной стороны, равен $|q|^8$, а с другой стороны, это квадрат определителя матрицы кватерниона (3.14), следовательно, определитель действительной матрицы кватерниона равен $|q|^4$.

Кватернион равен нулю тогда и только тогда, когда $|q| = 0$, т. е. при $a = b = c = d = 0$. Обратный по умножению кватернион выражается по формуле $q^{-1} = \bar{q}/|q|^2$. Таким образом, кватернионы образуют тело (поле без коммутативности умножения).

По аналогии с гауссовыми числами в кватернионах есть группа чисел, образующих ортонормированный базис в \mathbb{R}^4 , и им противоположных, а именно:

$$G = \{\pm 1, \pm i, \pm j, \pm k\}.$$

Это — группа по умножению порядка 8, она не является ни абелевой, ни циклической. В ней есть нормальная подгруппа, изоморфная \mathbb{Z}_2 , это группа $\{-1, 1\} \triangleleft G$. Фактором по ней является уже известная нам группа Клейна:

$$G/\{-1, 1\} \cong V_4,$$

которая изоморфна группе поворотов пространства кватернионов вокруг базисных осей $1, i, j, k$ на угол 180° .

Наконец, перейдем к тому сюжету, ради которого, собственно, кватернионы и были придуманы Гамильтоном. Рассмотрим семейство функций $R_q : \mathbb{H} \rightarrow \mathbb{H}$, проиндексированных всеми ненулевыми кватернионами и определяемых по формуле:

А мы-то думали, он хотел алгебру в \mathbb{R}^4 построить...

$$R_q(h) = qhq^{-1}. \quad (3.15)$$

Эти функции обладают следующими свойствами.

Теорема 3.10.

- 1) R_q не зависит от длины и знака q : $R_q = R_{tq}$ для любого $t \in \mathbb{R}^*$;
- 2) R_q сохраняет расстояние (т. е. является изометрией);
- 3) R_q сохраняет мнимое подпространство (если h не содержит действительной части, то и $R_q(h)$ таково же).

Первое свойство позволяет нам вместо всех R_q рассмотреть только такие, у которых $|q| = 1$. При этом стоит особо отметить, что q и $-q$ определяют одно и то же движение. Заметим, что в этом случае соответствующая кватерниону комплексная матрица $Q(z, w)$ принадлежит $SU(2)$, т. е. это специальная унитарная матрица из группы вращений.

Упражнение 3.37. Второе свойство в силу теоремы 3.9 означает, что R_q является ортогональным линейным оператором над \mathbb{R}^4 .
Докажите теорему. Третье свойство позволяет нам рассмотреть сферу S^2 в трехмерном мнимом подпространстве пространства \mathbb{R}^4 , базисом которого являются числа i, j, k :

$$S^2 = \{xi + yj + zk \mid x^2 + y^2 + z^2 = 1\}.$$

Это — обычная сфера единичного радиуса в трехмерном пространстве (поверхность мяча). В силу приведенной выше теоремы функции R_q переводят данную сферу в себя, сохраняя расстояния.

Можно показать, что множество $\{R_q \mid |q| = 1\}$ с операцией композиции ($R_q \circ R_{q'} = R_{qq'}$) образует группу, и эта группа изоморфна группе $SO(3)$ (специальная ортогональная группа, она же — группа вращений трехмерного пространства), действующей в мнимом пространстве (на осях i, j, k в \mathbb{H}).

Действительно, поворот двумерной мнимой сферы вокруг вектора $\bar{\xi} = ai + bj + ck$ на угол ϕ описывается кватернионом

$$q = \cos(\phi/2) + \sin(\phi/2)\bar{\xi}. \quad (3.16)$$

Соответствующая матрица поворота имеет вид

$$\begin{pmatrix} 1 - 2(b^2 + c^2) \sin^2(\phi/2) & 2ab \sin^2(\phi/2) - c \sin(\phi) & 2ac \sin^2(\phi/2) + b \sin(\phi) \\ 2ab \sin^2(\phi/2) + c \sin(\phi) & 1 - 2(a^2 + c^2) \sin^2(\phi/2) & 2bc \sin^2(\phi/2) - a \sin(\phi) \\ 2ac \sin^2(\phi/2) - b \sin(\phi) & 2bc \sin^2(\phi/2) + a \sin(\phi) & 1 - 2(a^2 + b^2) \sin^2(\phi/2) \end{pmatrix}$$

Если же вместо вращений рассматривать сами их индексы q , т. е. трехмерную сферу

$$S^3 = \{q \in \mathbb{H} \mid |q| = 1\},$$

с операцией умножения, то это будет группа с нормальной подгруппой \mathbb{Z}_2 , фактор по которой, т. е. группа S^3/\mathbb{Z}_2 , будет изоморфен группе $SO(3)$, поскольку факторизация по \mathbb{Z}_2 склеивает пары $q, -q$, что мы и видим при рассмотрении вращения R_q .

Таким образом, каждый кватернион q единичной длины в паре со своим противоположным $-q$ по формуле (3.15) задает некоторое вращение сферы, и обратно, всякое вращение сферы задается некоторым (единственным с точностью до знака минус) кватернионом единичной длины.

Этот факт позволяет нам построить известный гомоморфизм из группы $SU(2)$ в группу $SO(3)$. Действительно, в группе $SU(2)$ находятся комплексные матрицы 2×2 с определителем, равным 1, т. е. в точности матрицы вида $Q(z, w)$, задающие кватернионы длины 1. Таким образом, мы имеем естественный изоморфизм между $SU(2)$ и кватернионами, составляющими сферу радиуса 1 в пространстве \mathbb{H} .

Кроме того, отображение R_q , рассматриваемое уже как соответствие между кватернионом q единичной длины и оператором вращения сферы S^2 , является гомоморфизмом с единичных кватернионов на группу вращений $SO(3)$ трехмерного вещественного пространства, т. е. обычной сферы. При этом ядром такого гомоморфизма является пара $\{-1, +1\}$, поскольку R_q не различает знак q .

Композиция обоих гомоморфизмов и дает требуемый гомоморфизм из $SU(2)$ в $SO(3)$:

$$\begin{array}{ccc} SU(2) & \xrightarrow{\quad} & SO(3) \\ & \searrow Q(z,w) & \nearrow R_q \\ & \mathbb{H} & \end{array}$$

В кватернионах, по аналогии с числами Гаусса, можно построить систему «целых» чисел, именуемых числами Гурвица. К таковым числам относят все четверки $a + bi + cj + dk$, для которых $2a, 2b, 2c, 2d$ являются целыми числами одинаковой четности.

*Снова наши
герой
компьютер!*

Кватернионы имеют приложения в современных компьютерных системах, например, при моделировании вращений 3D-объектов, а также в GPS-позиционировании, поскольку компьютер требует максимально компактного представления операций, а представление четверкой чисел заметно экономичнее, чем матрицей 3×3 .

Для того, чтобы завершить описание всех движений в \mathbb{R}^3 , нам необходимо добавить, как и раньше, отражение относительно произвольной плоскости

(в частности, если эта плоскость проходит через 0, то это будет симметрия сферы) и параллельный перенос вдоль произвольного вектора.

Вращения сферы вместе с отражениями образуют группу $O(3)$ всех ортогональных преобразований пространства \mathbb{R}^3 . Как и в случае плоскости, любое такое движение можно рассматривать как композицию специального ортогонального преобразования (вращения) и некоторого выделенного отражения, и наоборот, как одно- или двухкратное отражение. Кроме того, существуют движения сферы, при которых не остается неподвижных точек. Это зеркальные вращения, при которых осуществляется вращение и отражение относительно плоскости, перпендикулярной оси вращения (при этом отражение коммутирует с вращением). Таким образом, все движения сферы можно свести либо к вращениям, отражениям и зеркальным вращениям, либо к одно-, двух- и трехкратным отражениям.

И мы вновь видим, что движения плоскости имеют тесную аналогию с движениями сферы (ранее мы видели аналогию между движениями прямой и окружности): осевое вращение сферы аналогично параллельному переносу плоскости и одновременно вращению плоскости вокруг точки, отражение сферы относительно плоскости аналогично симметрии плоскости относительно прямой, зеркальный поворот сферы аналогичен скользящей симметрии плоскости.

На основе движений сферы несложно получить движения пространства \mathbb{R}^3 , добавив к ним параллельный перенос. При этом у нас появятся такие типы движений как: **винт** (перенос с последующим вращением вокруг оси направления переноса), **скользящая симметрия** (перенос с последующей симметрией относительно плоскости, параллельной данному вектору) и **зеркальное вращение** (вращение с последующей симметрией относительно плоскости, перпендикулярной оси вращения).

Таким образом, все движения \mathbb{R}^3 делятся на следующие типы:

1. Параллельный перенос;
2. Винтовое движение (в частности, просто вращение, если перенос на 0);
3. Скользящая симметрия (в частности, просто отражение, если перенос на 0);
4. Зеркальное вращение.

Для этих четырех типов движений можно составить таблицу композиций, аналогичную той, которую мы делали для плоскости. Предлагаем читателю самостоятельно разобраться в данном вопросе.

Заметим, что параллельный перенос и вращение являются композицией двух отражений, скользящая симметрия и зеркальное вращение — трех, винтовое движение — четырех.

Соберем теперь в одну таблицу сравнения движения прямой, плоскости, пространства и двух сфер — одномерной (т. е. окружности) и двумерной (см. таблицу 3.2).

К данной таблице нужно сделать следующие пояснения:

1. Под *собственными* движениями принято понимать такие, которые не меняют ориентацию пространства, а под *несобственными* — те, которые меняют. Таким образом, если мы совершаем обход некоторой фигуры в положительном направлении, постоянно оставляя ее слева от направления движения, то образ нашего маршрута после симметрии или другого несобственного преобразования становится обходом в отрицательном направлении. Пространство при этом выворачивается наизнанку, и совершать подобное преобразование непрерывным образом, не выходя из самого пространства, невозможно. Так, симметрия прямой относительно точки предполагает поворот на 180° этой прямой в содержащей ее плоскости, а отражение 3-мерного пространства относительно плоскости — поворот на 180° этого пространства в более широком, 4-мерном пространстве. Примечательно, что именно несобственное движение (например, комплексное сопряжение) перестает быть линейным оператором над комплексным полем, оставаясь таковым над действительным (см. замечание к теореме 3.9).
2. Вращение окружности находится в столбце «Перенос», поскольку для внутренней метрики окружности ее вращение представляется именно как смещение или сдвиг. Грубо говоря, для жителей одномерной окружности ее вращение выглядит именно как перенос, и поэтому полностью аналогично переносу на прямой. Но с точки зрения плоскости, в которой находится эта окружность, это движение является именно вращением вокруг фиксированного центра, поэтому вращение окружности также находится в столбце «Поворот».
3. Вращение сферы отнесено в столбец «Перенос», поскольку локально, вблизи экватора этого вращения, оно выглядит в точности как параллельный перенос плоскости. Но оно же отнесено и в столбец «Поворот», поскольку вблизи полюсов вращения оно выглядит в точности как поворот плоскости вокруг центра.
4. По тем же причинам зеркальное вращение сферы соотносится как со скользящей симметрией пространства (вблизи экватора это выглядит именно так), так и с зеркальным вращением, поскольку именно таковым и является, с тем лишь ограничением, что отражение производится относительно плоскости, содержащей центр сферы.

Таблица 3.2: Сравнение движений.

	Собственные движения (не меняют ориентацию)		Несобственные движения (меняют ориентацию)		
	Перенос	Поворот	Смещение по- ворота	Симметрия	Смещенная симметрия
Прямая	сдвиг на число			относительно точки	
Окружность		вращение		осевая симметрия	
Плоскость	параллельный перенос	относительно точки		осевая симметрия	скользящая симметрия (перенос+симметрия)
Сфера	вращение вблизи экватора	вращение вблизи полюса		отражение относительно плоскости	зеркальное вращение (вращение+симметрия)
Пространство	параллельный перенос	осевое вращение	винт (перенос + вращение)	отражение относительно плоскости	скользящая симметрия (перенос+симметрия)

В завершении нашего разговора о движениях вещественных пространств скажем, что в пространстве \mathbb{R}^n любое движение может быть представлено как композиция не более чем $n + 1$ симметрий относительно гиперплоскостей размерности $n - 1$. Кроме того, они сводятся к ортогональным преобразованиям, переносам и композициям тех и других. В свою очередь, ортогональные преобразования могут быть представлены как композиции вращений и симметрий.

Комментарий 9. О разложении матриц линейных операторов.

Из вышесказанного можно заметить, что всякое движение, во-первых, является композицией смещения и линейного оператора, а во-вторых, матрицу оператора движения без смещения можно представить как произведение ортогональных матриц. Иначе говоря, матрицу линейного оператора, осуществляющего движение, можно разложить в произведение ортогональных матриц (вращений и симметрий). Это разложение хоть и напоминает разложение целого (гауссова, эйзенштейновского) числа по степеням простых, но все-таки не является единственным.

В более общем случае, т. е. когда оператор не является изометрией, его матрица может иметь более разнообразные разложения. Мы рассмотрим только два из них.

Пусть есть оператор $\hat{A} : X \rightarrow X$, действующий на линейном пространстве X над полем L . Если существует такое число $\lambda \in L$ и ненулевой вектор $v \in X$, что выполнено равенство $\hat{A}(v) = \lambda v$, то число λ и вектор v называются собственными (числом и вектором) для оператора \hat{A} . Очевидно, что оператор \hat{A} действует как растяжение вдоль вектора v . Очевидно также, что любой вектор αv также будет собственным, поэтому обычно выбирается собственный вектор единичной нормы (если на X есть норма).

Если речь идет о вещественном или комплексном конечномерном пространстве с выбранным базисом, то оператору \hat{A} соответствует некоторая матрица A , и собственное число и вектор оператора \hat{A} называются также собственными числом и вектором матрицы A .

Если матрица A размерности $n \times n$ имеет n линейно независимых собственных векторов, то ее можно представить в виде

$$A = VDV^{-1},$$

где V — ортогональная матрица, а D — диагональная матрица, в которой на главной диагонали стоят собственные числа матрицы A , а остальные элементы равны нулю. Такое разложение называется **спектральным** (спектр оператора — это множество всех его собственных чисел).

Наличие спектрального разложения у матрицы означает, что соответствующий ей оператор является растяжением (не путать с термином «гомотетия»,

который мы встретим чуть позже!). Действительно, преобразование VDV^{-1} , как мы знаем, является переходом от одного ОНБ к другому, в котором матрица оператора будет иметь вид D , т.е. станет диагональной. Но оператор с диагональной матрицей — это растяжение вдоль осей на коэффициенты, являющиеся его собственными числами! Например, тень мяча на полу — преобразование сферы в эллипс (если смотреть под углом, а не сверху), которое представляется линейным оператором, действующим из \mathbb{R}^3 в \mathbb{R}^2 и имеющим спектральное разложение (собственные числа при этом будут равны осям эллипса, если у исходного шара радиус равен 1).

Помимо спектрального разложения матрицы существует также сингулярное разложение.

Пусть A есть матрица размерности $m \times n$, т.е. это матрица линейного оператора, действующего из n -мерного пространства X в m -мерное пространство Y (оба над полем L). Если существуют два вектора $v \in X$, $u \in Y$ и число $\sigma \in L$ такие, что

$$Av = \sigma u, \quad A^*u = \sigma v,$$

то число σ называется сингулярным числом матрицы (оператора) A , а векторы v и u , соответственно, правым и левым сингулярными векторами матрицы (оператора) A . Здесь $A^* = \overline{A}^T$ (сопряженно-транспонированная матрица). Как обычно, векторы v и u выбираются так, чтобы они имели единичную норму.

Сингулярным разложением матрицы A называется разложение вида

$$A = U\Sigma V^*,$$

где U , V — это унитарные матрицы ($UU^* = E$ и $VV^* = E$) размерности, соответственно, m и n , состоящие из левых и правых сингулярных векторов, соответственно, а Σ — матрица $m \times n$, у которой на главной диагонали стоят сингулярные числа матрицы A , а остальные элементы равны нулю.

Таким образом, если матрица линейного оператора допускает сингулярное разложение, то данный оператор является композицией двух изометрий (поворотов и/или симметрий) (в соответствующих пространствах) и растяжения вдоль общих осей пространств (из-за различной размерности пространств X и Y растяжение может действовать не по всем осям). Сингулярное разложение — это обобщение случая спектрального разложения на матрицы произвольной размерности (не обязательно квадратные).

Сингулярное разложение используется, например, в задачах понижения размерности данных. Так, мы можем исходную матрицу A приблизить матрицей A_k той же размерности, но с заранее заданным рангом k . Близость понимается в смысле нормы Фробениуса разности $A - A_k$ (по сути, это просто евклидова длина вектора, если матрицу вытянуть в вектор длины $m \cdot n$). Оказывается, что наилучшим приближением матрицы $A = U\Sigma V^*$ будет матрица $A_k = U\Sigma_k V^*$, где

диагональная матрица Σ_k получена из Σ занулением всех диагональных элементов, кроме k наибольших. После некоторых преобразований приближение можно записать в виде разложения

$$A_k = U_k \Sigma_k V_k^*,$$

где Σ_k — квадратная матрица $k \times k$, а матрицы U_k и V_k имеют размерности, соответственно, $m \times k$ и $n \times k$. Матрица становится при этом меньше исходной, но сохраняет значительную долю информации об исходной матрице A .

Благодаря этому свойству сингулярное разложение находит широкое практическое применение в сжатии данных, обработке сигналов, численных итерационных методах для работы с матрицами, методе главных компонент, латентно-семантическом анализе и прочих областях.

Сингулярное разложение известно также как метод SVD (Singular Value Decomposition). На этом методе основаны многие рекомендательные системы в интернете.

Пример из машинного обучения. Если мы имеем m фильмов и n юзеров, составляющих отзывы о фильмах (рейтинг от 1 до 5), то мы можем заполнить ими матрицу оценок A размерности $m \times n$. При этом часть ячеек матрицы могут оказаться пустыми (нет отзыва данного юзера о данном фильме), а сама матрица может оказаться огромной (тысячи строк и столбцов). Здесь мы делаем допущение, что оценка i -го фильма j -м юзером есть скалярное произведение некоторого профиля фильма на профиль пользователя (векторы одинаковой размерности). То есть, если собрать в матрицу B все профили юзеров, а в матрицу C — фильмов, то должно быть $A \approx BC$ (хотя бы приблизительно). Поиск матриц профилей есть задача оптимизации при целевом условии $\|A - BC\| \rightarrow \min$ (здесь мы сравниваем только те элементы матриц, для которых нам известна оценка в матрице A). Решив такую задачу, мы получаем матрицы профилей и оценочную матрицу $A' = BC$.

Далее находим сингулярное разложение $A' = U\Sigma V^*$. После чего мы можем найти приближение Σ_k с относительно небольшим числом k (в пределах сотни), т. е. решить задачу понижения размерности и одновременно не допустить переобучения системы рекомендаций. После чего матрицу $B' = U\Sigma_k$ считаем оценкой профилей интересов пользователей, а матрицу $C' = V$ — профилей фильмов, после чего мы можем прогнозировать оценку фильма пользователем как соответствующее скалярное произведение строки и столбца этих матриц и предлагать персональные рекомендации этому пользователю (если ранее он уже оценил хотя бы несколько фильмов).

Помимо спектрального и сингулярного разложения существует еще с десяток видов разложений, среди которых наиболее известным, пожалуй, является жорданова нормальная форма.

3.6.3 Подобия в действительных пространствах

Подобия — это такие преобразования пространств, при которых расстояние между точками изменяется в фиксированное число раз. Так, если $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$ есть подобие, то существует число k такое, что для любых четырех точек $a, b, c, d \in \mathbb{R}^n$ ($a \neq b, c \neq d$)

$$\frac{|T(a) - T(b)|}{|a - b|} = \frac{|T(c) - T(d)|}{|c - d|} = k, \quad (3.17)$$

где k называется коэффициентом подобия и принимает действительные значения из интервала $(0; +\infty)$.

Частным случаем подобия является **гомотетия** (растяжение) — преобразование, определяемое числом $k \neq 0$ и точкой пространства O , при котором каждая точка X переходит в точку X' такую, что $\vec{OX'} = k\vec{OX}$. При этом точка O называется центром гомотетии, а $|k|$ — коэффициентом.

Из (3.17) легко видеть, что

$$\left| \frac{T(a) - T(b)}{k} \right| = |a - b|,$$

то есть T/k является изометрией, а значит, поддается вышеустановленной классификации движений.

Соответственно, исходное T получается из изометрии T/k путем умножения на коэффициент k , т.е. путем композиции с центральной (привязанной к нулю) гомотетией.

Таким образом, подобие есть композиция гомотетии (растяжения) и изометрии (движения).

В частности, на комплексной плоскости каждое ненулевое комплексное число z доставляет подобие $T(w) = wz$, которое является композицией гомотетии с коэффициентом $|z|$ и вращения, определяемого аргументом числа z . Такие преобразования еще называют **поворотными гомотетиями**.

Заметим, что подобия, сохраняющие 0, являются линейными операторами (теперь уже не обязательно ортогональными), но не всякое линейное преобразование является подобием. Например, линейное преобразование на плоскости с матрицей $\begin{pmatrix} a & 0 \\ 0 & 1/b \end{pmatrix}$, растягивающее векторы вдоль оси Ox в a раз и сжимающее векторы вдоль оси Oy в b раз — линейное, но не являющееся подобием.

Вообще, подобие — это всего лишь масштабирование (одинаковое по всем направлениям), поэтому все свойства изометрий с точностью до коэффициента увеличения размеров фигур остаются такими же как у движений. Например, подобия сохраняют углы, порядок точек на прямой, переводят прямые

в прямые, а окружности в окружности, площади фигур при подобиях увеличиваются в k^2 раз, а объемы — в k^3 раз. Подобия являются биекциями на пространстве \mathbb{R}^n и образуют группу преобразований, нормальной подгруппой которой является группа движений.

3.6.4 Аффинные преобразования и однородные координаты

Здесь необходимо пояснить понятие аффинного пространства. Линейное пространство \mathbb{R}^n состоит из точек-векторов, т. е. из одного сорта объектов, на которых заданы операции сложения и умножения на число (модуль над кольцом), причем они жестко привязаны к конкретной координатной сетке (ибо задаются как упорядоченные наборы чисел). В аффинном пространстве (которое, кстати, является объектом изучения школьной геометрии в виде плоскости) различают собственно точки линейного пространства и векторы, которые задаются парой точек (начало и конец вектора). При этом точки и векторы можно складывать и вычитать. Так, сумма точки и вектора дает точку, разность точек дает вектор, сумма векторов дает вектор.²⁶ Таким образом, мы в каждую точку аффинного пространства «вклеиваем» по одному экземпляру \mathbb{R}^n , отвечающему за формирование всех исходящих из этой точки векторов. При этом на векторах равенство задано так, что начало и конец вектора не берутся в расчет, важен только сам «приклеенный» в начало вектор. Иначе говоря, $\vec{AB} = \vec{A'B'}$ тогда и только тогда, когда равны векторы $B - A$ и $B' - A'$, т. е. равенство векторов наследуется из их исходного пространства и не зависит от точек. В аффинном пространстве нет системы координат, нет выделенного начала отсчета. Любую точку можно свободно рассматривать как 0, а координаты после этого задаст тот пучок векторов \mathbb{R}^n , который «приклеен» к этой точке.

Лучше поздно, чем никогда!

Полученная конструкция, состоящая из двух пространств \mathbb{R}^n и операции сложения точки и вектора обозначается \mathbb{A}^n и называется n -мерным **аффинным пространством**.

Не путать с множеством алгебраических чисел!

Если теперь над аффинным пространством (точнее, над его точками) задать метрику как длину соответствующего вектора, соединяющего две данные точки (и определяемую скалярным произведением этих векторов), то полученная конструкция, состоящая из двух пространств \mathbb{R}^n , скалярного произведения, метрики и операции сложения точки и вектора, называется n -мерным **евклидовым пространством**.

Такие громоздкие конструкции могут показаться избыточными с точки зрения сложности реализации и расходования памяти, если мы решим их

²⁶Более точно, для любых точек A, B и векторов v, w имеют место аксиомы: $(A+v)+w = A+(v+w)$, $A+0 = A$, $\exists v: A+v = B$. В общем случае векторы могут принадлежать любому линейному пространству, лишь бы эти операции были корректно определены и удовлетворяли данным аксиомам.

закладывать в ЭВМ. Однако, они дают определенную свободу геометрам в том смысле, что дают представление о пространстве без его кодирования декартовыми координатами. Скрывая в фундаменте числа, на поверхность мы выставляем точки и векторы как самостоятельные сущности, и начинаем ими оперировать примерно так же, как это делал еще Евклид почти 2.5 тысячи лет назад, ничего не зная о Рене Декарте. По-видимому, мир вообще устроен так, что упрощение понимания объекта сопровождается усложнением его реализации «в железе», и наоборот.

Как уже упоминалось выше, в аффинных пространствах параллельный перенос склеивается с id в том смысле, что он никак не действует на векторы (переносит точки, мы никак не меняем векторы, которые их соединяют). Поэтому любая изометрия (движение) в аффинном пространстве на его векторах представляет собой линейное отображение. При этом движения полностью сохраняют геометрический объект — не только его форму, но и размеры. Про фигуры, совмещаемые движением, даже говорят, что они равны.

Подобия в аффинном пространстве также являются линейными операторами на векторах, но более общего свойства — они сохраняют форму, но не размер, производя одинаковое во всех направлениях масштабирование.

Следующим естественным обобщением преобразований аффинного пространства будут все обратимые линейные операторы, т.е. группа $\text{GL}(n, \mathbb{R})$. Их основное свойство — линейность — позволяет сохранять линейные объекты, т.е. образами прямых будут по-прежнему прямые, образами плоскостей — плоскости, и т.д. Однако окружность, в отличие от преобразований подобия, при аффинных преобразованиях не сохраняется, она становится эллипсом, а шар — эллипсоидом. Для некоторых линейных преобразований сохраняется объем (или площадь) фигуры. Например, гиперболическое преобразование плоскости, заданное матрицей $\begin{pmatrix} a & 0 \\ 0 & 1/a \end{pmatrix}$ смещает точки фигуры по параллельным гиперболам, искажая саму фигуру, но не меняя ее площадь.

Выше мы неоднократно видели, что параллельный перенос вызывает определенные неудобства при описании как движений, так и подобий, не укладываясь в матричное представление. Параллельный перенос «работает» на уровне точек, а не векторов, а для них сдвиг в пространстве существенен. И с прикладной (вычислительной) точки зрения привлечение аффинного пространства если не невозможно, то крайне затруднительно. Тем не менее существует простой способ свести к одним и тем же матрицам как движения с подобиями, так и переносы. Для этого нужно увеличить размерность матрицы линейного оператора, тем самым выйдя в надпространство относительно рабочего пространства.

Для примера рассмотрим преобразования плоскости. Вместо точки (x, y) будем рассматривать точку $(x, y, 1)$, которая фактически является записью пропорции $x : y : 1$ и поэтому считается эквивалентной любой точке вида

$(\alpha x, \alpha y, \alpha)$, где $\alpha \neq 0$. Тогда матрицы поворота на угол ϕ относительно начала координат и последующего параллельного переноса на вектор (x_0, y_0) по отдельности и в рамках одного оператора выглядят следующим образом:

$$\begin{pmatrix} \cos \phi & -\sin \phi & 0 \\ \sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & x_0 \\ 0 & 1 & y_0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} \cos \phi & -\sin \phi & x_0 \\ \sin \phi & \cos \phi & y_0 \\ 0 & 0 & 1 \end{pmatrix}$$

Последняя матрица получается умножением второй на первую (в порядке справа налево, как положено для композиций операторов). Нетрудно убедиться в том, что все рассмотренные нами преобразования поворота, переноса, симметрии, гомотетии, различного масштабирования по осям, а также их композиции, записываются такими матрицами и образуют подгруппу в полной линейной группе $GL(3)$. Более того, они образуют группу, изоморфную подгруппе группы $PGL(2)$ — группе проективных преобразований, которая, вообще говоря, чуть меньше, чем $GL(3)$.

Таким образом, расширение рабочего пространства на одно измерение приводит нас к получению более удобной аналитической формы изучения преобразований исходного пространства.

Примечательно, что именно такие *однородные координаты*, сохраняющие пропорции, используются для построения следующего, более общего объекта геометрии — проективного пространства, и связанных с ним проективных преобразований. В связи с этим далее мы рассмотрим начала проективной геометрии.

3.6.5 Проективная геометрия

По словам Ф. Клейна, «*проективная геометрия — это вся геометрия*». Иначе говоря, все, чем занимается геометрия (в понимании Эрлангенской программы) можно описать в терминах проективной геометрии.

Для того, чтобы расширить аффинное пространство до проективного, нам необходимо включить в него бесконечно удаленные точки. Дело тут вот в чем. Как мы уже видели выше (рис. 3.3) при проецировании²⁷ одной плоскости на другую (в нашем примере — плоскости α на плоскость β с центром проецирования O) некоторые точки проецируются параллельно результирующей плоскости и, тем самым, «уходят на бесконечность». При этом, если центр проецирования сдвигать, то направления этих бесконечностей меняют угол. Получается, что нам нужно добавить не одну виртуальную бесконечность, а много — столько, сколько углов направлений, т. е. целый континуум.

Конечно, в более простом случае, когда речь идет о проецировании одной прямой на другую (обе лежат в одной плоскости) никаких углов нет, и

²⁷ Важно не путать *проецирование* с *проектированием*, как и геометра с менеджером проекта.

можно отделаться добавлением всего лишь одной формальной точки ∞ , тем самым получив то, что в анализе называется расширенной прямой, а в топологии — окружностью S^1 . Но в общем случае требуется добавить целую «бесконечную гиперплоскость», которая обернет все пространство примерно так же, как сфера оборачивает шар, только с одним маленьким топологическим «но»: противоположные направления на этой сфере потребуются склеить, т. к. проецирование «не замечает» направления.

Таким образом, проективная плоскость помимо привычных финитных точек содержит бесконечно удаленную прямую, как бы обрамляющую эту плоскость по периметру (на бесконечности). Проективные преобразования — это суперпозиции проектирований из точки одной плоскости на другую. С топологической точки зрения проективная прямая — это окружность, а проективная плоскость — это неориентируемая двумерная поверхность, получаемая из круга приклеиванием по его границе листа Мёбиуса.

При проективных преобразованиях эквивалентными становятся не только окружность и эллипс (как при аффинных), но все кривые второго порядка — окружность, эллипс, гипербола, парабола (бесконечные точки позволяют достроить их разорванные графики до непрерывной кривой). Кроме того, в проективном пространстве можно ввести понятие расстояния.

Наконец, оказывается, что любое алгебраическое тело (т. е. поле за вычетом требования коммутативности умножения) порождает единственное соответствующее ему проективное пространство, а локально компактных непрерывных тел всего три — действительные числа, комплексные числа и кватернионы.

Снова та же тройка, как в теореме 3.6

Перейдем от беллетристики к конкретике.

Формально вещественное проективное пространство \mathbb{RP}^n над полем \mathbb{R} определяется следующим образом. В \mathbb{R}^{n+1} (т. е. линейном пространстве размерности $n+1$ над полем \mathbb{R}) рассматривается пучок всех прямых, проходящих через точку 0. Множество этих прямых и есть базовое множество проективного пространства \mathbb{RP}^n . На проективном пространстве можно определить метрику и движения, наследуя их из исходного пространства \mathbb{R}^{n+1} . Поскольку линейные преобразования \mathbb{R}^{n+1} сохраняют прямые, очевидно, что они корректно действуют на проективном пространстве. Кроме того, линейные преобразования \mathbb{R}^{n+1} описывают полностью все движения и подобию в \mathbb{R}^n при помощи перехода к однородным координатам. Но с помощью однородных координат легко интерпретировать и проективное пространство.

Рассмотрим для начала случай $n = 1$. Тогда проективным пространством будут все прямые на плоскости, содержащие 0. Выделим на плоскости прямую l , заданную уравнением $y = 1$ ($l \notin \mathbb{RP}^1$!) Каждая прямая, проходящая через 0 и не совпадающая с осью Ox , пересекает прямую l в единственной точке $(x, 1)$. Мы имеем взаимно однозначное соответствие между такими прямыми и точками \mathbb{R} , а ось Ox при этом соответствует бесконечно удаленной

точке.

Заметим, что однородные координаты $x : 1$, сохраняющие пропорцию, в точности пробегают прямую, проходящую через 0 и точку $(x, 1)$. Все такие точки (кроме точки O) мы считаем эквивалентными. То есть $(x, 1) \propto (tx, t)$ ($t \neq 0$).

Кроме того, любое невырожденное ($\det \neq 0$) линейное преобразование плоскости, заданное матрицей $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, переводит элементы проективного пространства в элементы проективного пространства, т. е. осуществляет преобразования проективного пространства, именуемые *проективными преобразованиями*.

Выше мы накладывали ограничения на матрицы этих преобразований, чтобы получить только движения и подобия. В случае же проективного пространства таких ограничений нет, кроме требования невырожденности матрицы. Ниже мы увидим, к чему это приводит.

Но возникает резонный вопрос: при чем тут проецирование? Почему пространство называется проективным?

Пусть точка $(X, 1)$ пробегает все элементы проективной прямой (иначе говоря, X пробегает \mathbb{R} и точку ∞ , а отношение $X : 1$ задает все элементы \mathbb{RP}^1). На рисунке 3.4 мы выделили красным прямую l , заданную уравнением $y = 1$, и ввели на ней относительную координату X , дублирующую первую координату пропорции $X : 1$.

Пусть теперь пропорция $X' : 1$ определяется нашей матрицей линейного преобразования плоскости:

$$\begin{pmatrix} X' \\ 1 \end{pmatrix} \propto \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} X \\ 1 \end{pmatrix} = \begin{pmatrix} aX + b \\ cX + d \end{pmatrix} \propto \begin{pmatrix} \frac{aX+b}{cX+d} \\ 1 \end{pmatrix}$$

Иначе говоря, пропорция $X' : 1$ задает те же прямые, что и пропорция $(aX + b) : (cX + d)$. То есть, по-просту говоря, X' выражается как дробно-линейная функция от X с коэффициентами из матрицы линейного оператора плоскости.

Далее сделаем следующее. Нам нужно построить прямую m (синего цвета) так, чтобы выполнялись условия:

1. X' является относительной координатой на m , сохраняющей расстояния (т. е. относительные координаты $X' = 0$ и $X' = 1$ на этой прямой отстоят друг от друга на 1 в абсолютных координатах плоскости),
2. относительной координате X на прямой l соответствует относительная координата $X' = (aX + b)/(cX + d)$ на прямой m путем проецирования l на m через центр O .

Упражнение 3.39. Проверьте, что сохранение пропорции — это отношение эквивалентности.

Знак \propto означает «пропорционально».

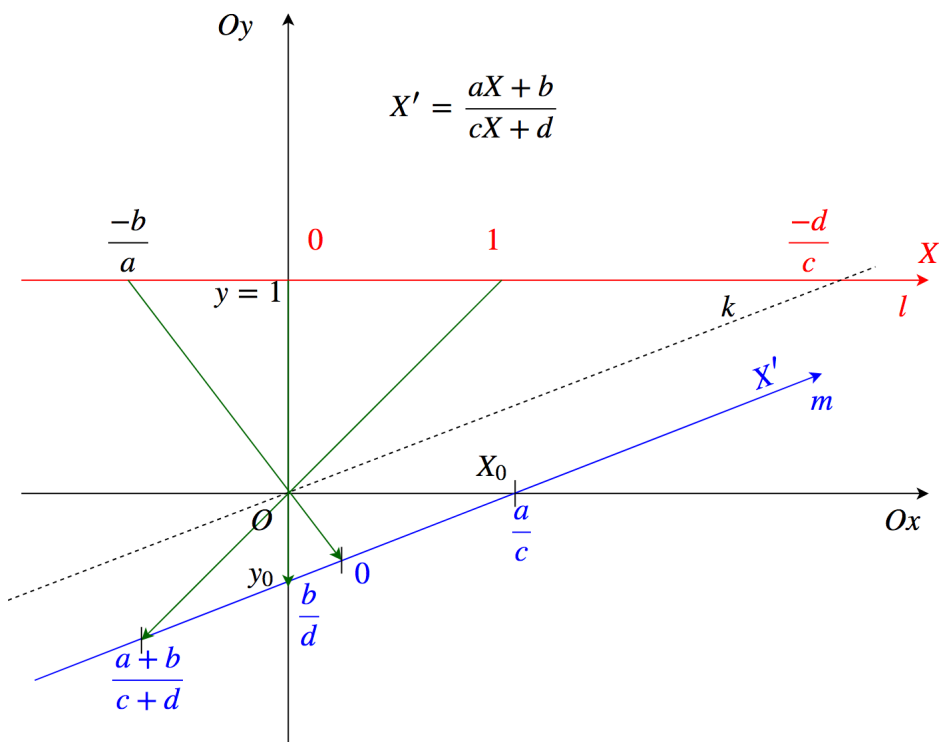


Рис. 3.4: Проективная прямая

Для начала нужно выяснить угол наклона прямой m . Для этого воспользуемся тем, что точка $X = -d/c$ должна уходить в бесконечность при проецировании (т. к. она доставляет 0 в знаменателе пропорции). Проведем вспомогательную прямую k через O и $(-c/d, 1)$. Далее, ее нужно сдвинуть параллельно так, чтобы точка $X = \infty$ перешла в точку пересечения m и оси Ox . На m это будет относительная координата $X' = a/c$ (предел при $X \rightarrow \infty$), а на оси Ox это будет неизвестная точка $(X_0, 0)$. На оси Ox она находится по той причине, что при проецировании через O мы приходим из бесконечности параллельно прямой l . Наконец, чтобы зафиксировать масштаб на прямой m , мы должны учесть, что точка $X = 0$ переходит в $X' = b/d$, которая в абсолютных координатах имеет представление $(0, y_0)$.

Само проецирование на рис.3.4 показано зелеными стрелками. В дополнение показано, что точка $-b/a$ в координатах прямой l переходит в точку 0 на прямой m , а точка 1 на прямой l переходит в точку $(a+b)/(c+d)$ на прямой m .

Длина отрезка $[(X_0, 0), (0, y_0)]$ равна $a/c + b/d$, что позволяет зафиксировать прямую m в одном из двух положений: либо выше точки O , либо ниже,

причем положительные направления m при этом будут противоположными. Проще всего выбрать то, которое сонаправлено прямой l . На рис.3.4 мы предполагаем, что $a, b, c > 0$ и $d < 0$, другие случаи предлагаем читателю разбирать самостоятельно на конкретных примерах.

Из подобия треугольников получаем:

$$\frac{X_0}{-d/c} = \frac{y_0}{-1} = \frac{a/c + b/d}{\sqrt{1 + d^2/c^2}},$$

откуда

$$\begin{cases} X_0 &= -\operatorname{sgn}(c)d \left(\frac{a}{c} + \frac{b}{d} \right) / \sqrt{c^2 + d^2} \\ y_0 &= -|c| \left(\frac{a}{c} + \frac{b}{d} \right) / \sqrt{c^2 + d^2} \end{cases}$$

Заметим, что всегда $c^2 + d^2 > 0$, что следует из условия невырожденности матрицы преобразования. Отсюда же следует, что круглая скобка всегда отлична от нуля. Если же $c = 0$ (что соответствует обычному линейному преобразованию), то переходя к пределу $c \rightarrow 0$ получаем упрощенные формулы

$$\begin{cases} X_0 &= \infty \\ y_0 &= a/d \end{cases}$$

т. е. прямая m будет параллельна l , причем будет отстоять от Ox на расстояние $|a/d|$, обеспечивая тем самым при проекции сжатие масштаба с коэффициентом $|a/d|$.

Если $d = 0$, то получаем²⁸

$$\begin{cases} X_0 &= -b/c \\ y_0 &= \infty \end{cases}$$

т. е. прямая m будет параллельна оси Oy и будет отстоять от нее на расстояние $|b/c|$.

Итак, мы видим, что в проективном пространстве \mathbb{RP}^1 линейные преобразования плоскости \mathbb{R}^2 обеспечивают движения, подобия и дробно-линейные преобразования, причем мы можем явным образом построить все необходимые проекции, показывающие, как точка X исходного пространства переходит в точку X' пространства-образа. Поскольку мы включаем в рассуждения и бесконечную точку, эти действия описывают преобразование всего пространства \mathbb{RP}^1 .

²⁸Строго говоря, можно выбрать $X_0 = b/c$, т. к., как мы уже отмечали, расположение m допускает два симметричных варианта в зависимости от требуемого направления положительной полуоси X' . Выбор $X_0 = -b/d$ обеспечивает рост X' при росте X , хотя при этом X' отрицательно и приближается к нулю.

На самом деле, несложно показать, что дробно-линейные преобразования прямой сохраняют двойное отношение. Пусть на прямой l в ее относительных координатах заданы 4 точки: X_1, X_2, X_3, X_4 , а на прямой m в ее относительных координатах лежат точки X'_1, X'_2, X'_3, X'_4 — образы X_1, X_2, X_3, X_4 относительно дробно-линейного преобразования с матрицей $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Среди них может быть и бесконечно удаленная точка. Тогда:

Упражнение 3.40.
Докажите!

$$[X_1; X_2; X_3; X_4] \Rightarrow \frac{(X_3 - X_1)(X_4 - X_2)}{(X_3 - X_2)(X_4 - X_1)} = \frac{(X'_3 - X'_1)(X'_4 - X'_2)}{(X'_3 - X'_2)(X'_4 - X'_1)}.$$

Определение. **Проективными преобразованиями** называются такие преобразования, которые сохраняют двойное отношение любых 4-х точек, лежащих на прямой.

Теорема 3.11. *Преобразование проективное тогда и только тогда, когда оно дробно-линейное.*

Из этой теоремы, в частности, следует аналог леммы о трех звездах: если проективное преобразование прямой имеет три неподвижные точки, то оно тождественно. Другая интерпретация этого факта состоит в следующем: проективное преобразование прямой однозначно задается тремя парами соответствующих точек. Обобщение этого факта на \mathbb{RP}^n : проективное преобразование однозначно задается двумя наборами из $n + 2$ точек общего положения и попарным соответствием точек одного набора точкам другого набора (*основная теорема о проективных преобразованиях*).²⁹

Напомним, что двойное отношение обладает свойством (2.13), т. е. все возможные перестановки 4-х переменных X_1, X_2, X_3, X_4 (которых ровно 24) разбиваются на 6 классов по 4 перестановки, в каждом из которых величина двойного отношения постоянна.

Рассмотрим случай $n = 2$. Здесь проективное пространство составляют все прямые в \mathbb{R}^3 , проходящие через 0. Каждая прямая описывается тройкой чисел (ξ, η, ζ) , из которых хотя бы одно отлично от нуля, причем тройки (ξ, η, ζ) и $(t\xi, t\eta, t\zeta)$, $t \in \mathbb{R}$, описывают одну и ту же прямую. Иначе говоря, координаты точек на такой прямой сохраняют пропорцию $\xi : \eta : \zeta$. Если прямая имеет ненулевой угол наклона относительно плоскости xOy , то она пересекается с плоскостью $z = 1$ в некоторой единственной точке $(x, y, 1)$, и пропорция $x : y : 1$ будет полностью описывать эту прямую. На рис. 3.5 это прямая l .

При этом мы получаем взаимно однозначное соответствие между прямыми, не лежащими в xOy и точками плоскости $z = 1$. Так что можно сказать,

²⁹Набор $n + 2$ точек называется набором *точек общего положения*, если соответствующие им векторы обладают тем свойством, что любые $n + 1$ из них образуют базис в \mathbb{R}^{n+1} .

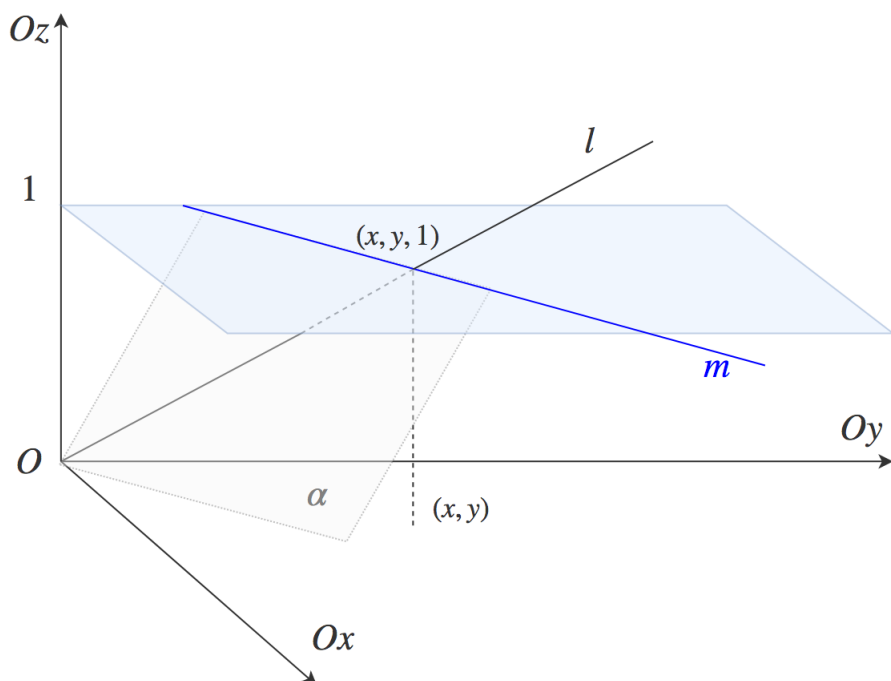


Рис. 3.5: Проективная плоскость

что проективное пространство \mathbb{RP}^2 — это плоскость \mathbb{R}^2 плюс множество прямых, лежащих в xOy и проходящих через 0. Эти прямые представляют собой бесконечно удаленные точки пространства \mathbb{RP}^2 и кодируются углом поворота, т. е. вещественным числом в интервале $[0, \pi)$. Кроме того, можно сказать, что \mathbb{RP}^2 — это прямая сумма \mathbb{R}^2 и \mathbb{RP}^1 (ведь \mathbb{RP}^1 определяется именно как пучок прямых на плоскости, проходящих через 0), и для всех последующих размерностей имеет место интерпретация $\mathbb{RP}^n = \mathbb{R}^n \sqcup \mathbb{RP}^{n-1}$.

Упражнение 3.41.
Почему не 2π ?

В пространстве \mathbb{RP}^2 появляются прямые (не как точки пространства, а как прямые этого пространства). Каждая прямая пространства \mathbb{RP}^2 — это плоскость в исходном \mathbb{R}^3 , содержащая 0. На рис. 3.5 она обозначена α . Всякая такая плоскость задается уравнением $ax + by + cz = 0$, где хотя бы одно из чисел a, b, c отлично от нуля. Интерпретировать ее в плоскости $z = 1$ можно естественным образом — пересечением плоскости α с плоскостью $z = 1$. Это пересечение — прямая m , обозначенная синим цветом на рис. 3.5.

Заметим, что плоскость xOy также является прямой в \mathbb{RP}^2 , только бесконечно удаленной. Так что в \mathbb{RP}^2 любые две прямые пересекаются (пересечение прямой и бесконечно удаленной прямой есть бесконечно удаленная точка). В

проективном мире нет параллельности! Кроме того, в \mathbb{RP}^2 , как и положено, через любые две точки можно провести прямую, и притом только одну. Действительно, точки — это прямые, проходящие через 0, через них проходит единственная плоскость, содержащая 0, она и определяет единственную прямую, проходящую через данные точки (даже если одна или обе эти точки бесконечно удаленные).

Заметим, что уравнение $ax + by + cz = 0$ можно рассматривать двояко: это уравнение плоскости (x, y, z) , перпендикулярной вектору (a, b, c) , но это же и уравнение плоскости (a, b, c) , перпендикулярной вектору (x, y, z) . Такой двойственный взгляд на уравнение (как на символьное выражение) приводит к определению двойственных объектов на проективной плоскости. Мы говорим, что точка $M \in \mathbb{RP}^2$ **двойственна** прямой $m \subset \mathbb{RP}^2$ (прямая m двойственна точке M), если прямая l в \mathbb{R}^3 , задающая точку M , перпендикулярна плоскости α в \mathbb{R}^3 , задающей прямую m (рис.3.6).

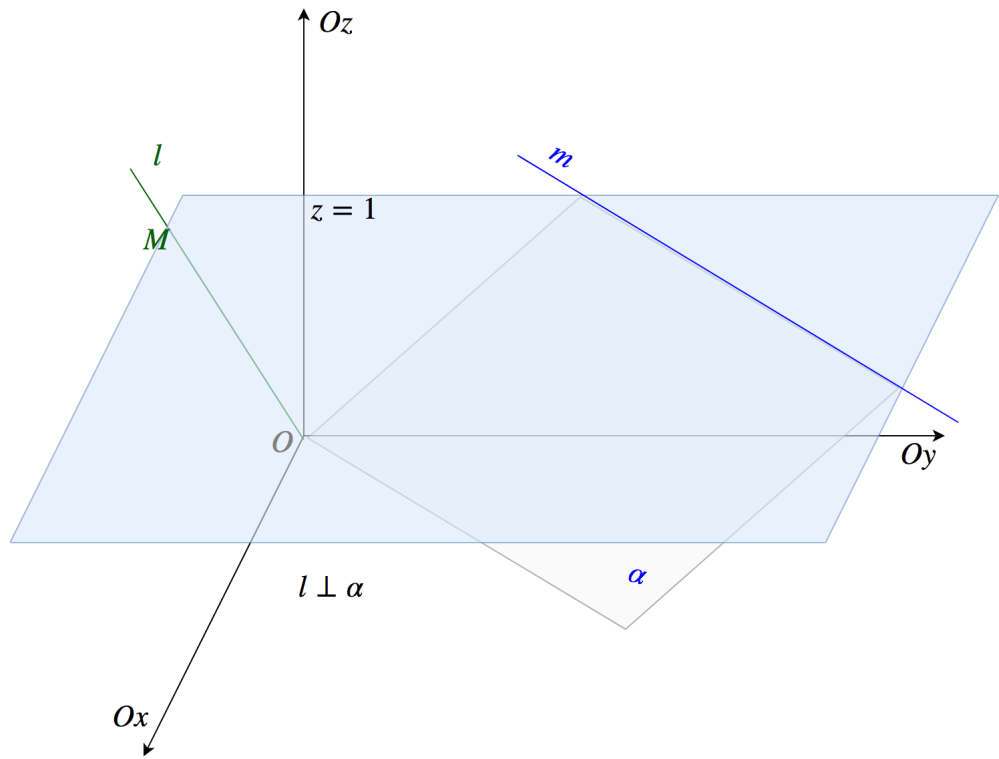


Рис. 3.6: Двойственные прямая m и точка M

Здесь мы выходим на еще один фундаментальный архетип математики: **двойственность** интерпретации смежных сущностей.

Ясно, что по любой прямой m можно построить единственную двойственную ей точку M , и наоборот, по любой точке M можно построить единствен-

ную двойственную ей прямую m (включая бесконечно удаленные объекты).

Добавим еще одно понятие — инцидентность. Точка M инцидентна прямой l (прямая l инцидентна точке M), если $M \in l$ (здесь принадлежность следует интерпретировать как вложение прямой, задающей точку M , в плоскость, задающую прямую l). Ясно, что если точка инцидентна двумя прямой, то она является точкой их пересечения, а если прямая инцидентна двум точкам, то она проходит через них.

Пусть точки M , M_1 и M_2 двойственны прямой m , m_1 и m_2 , соответственно. Тогда точка M инцидентна прямой m_1 и m_2 тогда и только тогда, когда прямая m инцидентна точкам M_1 и M_2 .

Упражнение 3.42.
Проверьте это утверждение.

Пользуясь этим простым утверждением, как базой индукции по сложности формулы, легко доказать следующую теорему [88].

Теорема 3.12 (о двойственном соответствии). *Если истинна формула φ , составленная из атомарных формул, описывающих отношение инцидентности точек и прямых, то после замены переменных на обозначения двойственных им объектов формула останется истинной.*

Например, если мы что-то доказали о расположении некоторых точек относительно сторон треугольника ABC и пересекающей его прямой l , то такое же утверждение справедливо о расположении двойственных им прямых относительно точек, двойственных сторонам треугольника и прямой l .

Аналогично случаю $n = 1$ для каждой невырожденной матрицы 3×3 можно построить плоскость, проекция через O на которую даст дробно-линейное преобразование, записываемое в однородных координатах следующим образом:

$$\begin{pmatrix} X' \\ Y' \\ 1 \end{pmatrix} \propto \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix} = \begin{pmatrix} a_{11}X + a_{12}Y + a_{13} \\ a_{21}X + a_{22}Y + a_{23} \\ a_{31}X + a_{32}Y + a_{33} \end{pmatrix} \propto \begin{pmatrix} \frac{a_{11}X + a_{12}Y + a_{13}}{a_{31}X + a_{32}Y + a_{33}} \\ \frac{a_{21}X + a_{22}Y + a_{23}}{a_{31}X + a_{32}Y + a_{33}} \\ 1 \end{pmatrix}$$

Нормирование первых двух координат на третью и дает дробно-линейное преобразование проективной плоскости \mathbb{RP}^2 . И в двумерном, и во всех остальных случаях теорема 3.11 остается справедливой.

Немного о группах

Следует особо отметить, что если все коэффициенты матрицы, порождающей данное преобразование, умножить на одно и то же число $\lambda \neq 0$, то преобразование не изменится, поскольку пропорция сохранится (мы просто сократим числитель и знаменатель дробно-линейной функции на λ). Это значит, что пропорциональные матрицы дают одно и то же проективное преобразование.

На самом деле, это единственный «склеивающий» фактор, отличающий разноеобразие всех невырожденных матриц от всех проективных преобразований.

Таким образом, полная линейная группа $GL(n, \mathbb{R})$ гомоморфно отображается на группу проективных преобразований $(n - 1)$ -мерного проективного пространства (т. е. группа невырожденных дробно-линейных преобразований), а ядро данного гомоморфизма состоит из матриц λE ($\lambda \neq 0$) и образует группу гомотетий пространства \mathbb{R}^n , изоморфную \mathbb{R}^* (группа ненулевых действительных чисел по умножению). **Группа проективных преобразований** \mathbb{RP}^{n-1} обозначается $PGL(n, \mathbb{R})$. Без потери общности можно считать, что в нее входят обратимые матрицы с условием $|\det| = 1$ (причем, если матрица с определителем -1 получается умножением на -1 из другой матрицы, то мы ее выбрасываем в силу факторизации).

В группе PGL принято выделять группу собственных преобразований, обозначаемую PSL . Она образует подгруппу группы PGL , определяемую тем условием, что определитель матриц в данной группе равен строго 1 (в то время как в PGL он может быть равен и -1). В том случае, если n нечетное,³⁰ имеет место равенство $PSL(n, \mathbb{R}) = PGL(n, \mathbb{R})$, а для четного n группа $PSL(n, \mathbb{R})$ будет собственной нормальной подгруппой $PGL(n, \mathbb{R})$, причем фактор $PGL(n, \mathbb{R})/PSL(n, \mathbb{R})$ будет изоморфен \mathbb{Z}_2 .

Напомним, что ранее мы сталкивались с группой $SL(n, \mathbb{R})$ всех матриц, определитель которых равен 1. При четном n среди таких матриц есть те, которые задают одно и то же проективное отображение, они отличаются умножением на матрицу $-E$. Поэтому $PSL(n, \mathbb{R}) = SL(n, \mathbb{R})/\pm E$ при четном n .

Приведем простой пример, иллюстрирующий отличие упомянутых групп. Возьмем матрицу $A = \begin{pmatrix} \lambda & 0 \\ 0 & -\lambda \end{pmatrix}$, $\lambda > 1$. Очевидно, что $\det A = -\lambda^2$. Ясно, что $A \in GL$, т. к. она обратима, и в то же время она не входит в PGL , PSL и SL . После нормирования на коэффициент мы получим матрицу $A/\lambda = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ с определителем $= -1$. Эта матрица доставляет проективное преобразование $y = -x$, но не принадлежит ни SL , ни PSL .

С другой стороны, взяв матрицу $B = \begin{pmatrix} -\lambda & 0 \\ 0 & -\lambda \end{pmatrix}$, $\lambda > 1$, мы получим, что $B \in GL$, но при этом она не входит в PGL , PSL и SL . После факторизации по условию $\det = 1$ мы найдем соответствующую ей (одну из двух) матрицу $B/\lambda = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$, которая принадлежит SL , но не принадлежит ни PGL , ни PSL (т. к. там есть уже матрица E , отличающаяся на коэффициент -1).

³⁰На самом деле, в более общем случае, нужно, чтобы корень n степени извлекался в том теле, над которым заданы матрицы. В частности, $PSL(n, \mathbb{C}) = PGL(n, \mathbb{C})$ независимо от четности n .

Факторизационные переходы, описанные выше, мы покажем на следующей диаграмме.

$$\begin{array}{ccc}
 \mathrm{GL} & \xrightarrow{\det=1} & \mathrm{SL} \\
 \downarrow / \lambda E & & \downarrow / \pm E \\
 \mathrm{PGL} & \xrightarrow{\det=1} & \mathrm{PSL}
 \end{array}$$

Шаг в комплексный проективный мир

До сих пор мы рассматривали только вещественное проективное пространство \mathbb{RP}^n . На самом деле, оно определяется для произвольного тела \mathbb{T} : точками пространства \mathbb{TP}^n являются прямые линейного пространства \mathbb{T}^{n+1} , проходящие через 0. Под прямой при этом понимается множество точек, получаемое из произвольного вектора умножением всех его координат на произвольное число из тела \mathbb{T} .

Например, в качестве \mathbb{T} можно рассмотреть поле рациональных чисел, и тогда \mathbb{QP}^n будет включать все прямые, проходящие через 0 и точки с целыми координатами. Мы уже неявно встречались с этим пространством, когда строили рациональные числа через целые (см. рис. 2.7 на стр.148), только в том случае мы их наносили на целочисленную решетку, но если их вложить в \mathbb{Q}^2 или даже в \mathbb{R}^2 , мы получим изоморфное множество прямых.

Интересны также случаи комплексного проективного пространства и пространства, формируемого на основе конечных полей (простейший пример которых — поле вычетов \mathbb{Z}_p по простому модулю). На комплексной проективной прямой \mathbb{CP}^1 изучаются дробно-линейные комплексные функции, ярким представителем которых является функция Жуковского.

Проективное пространство \mathbb{CP}^1 определяется как множество всех комплексных прямых в \mathbb{C}^2 , проходящих через 0, которые описываются пропорциями $z : w$,³¹ при этом пропорции $z : 1$ задают конечные точки \mathbb{CP}^1 и топологически составляют обычную комплексную плоскость, а пропорции $z : 0$ (с точностью до комплексного коэффициента это просто одна точка $(1, 0)$) задают бесконечно удаленную точку \mathbb{CP}^1 .

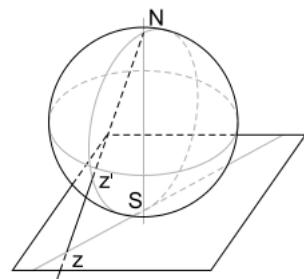


Рис. 3.7: Риманова поверхность $\mathbb{C} \cup \{\infty\}$.

³¹Пропорции определяются с точностью до коэффициента $\lambda \in \mathbb{C} \setminus \{0\}$. Поэтому комплексная прямая в \mathbb{C}^2 — это геометрически более сложный объект, чем вещественная прямая в \mathbb{R}^4 .

Таким образом, топологически пространство \mathbb{CP}^1 представляет собой комплексную плоскость \mathbb{C} , дополненную одной бесконечной точкой, которая стягивает эту плоскость за ее бесконечно удаленный край, превращая в сферу S^2 . Таким образом, \mathbb{CP}^1 можно интерпретировать как обычную сферу, на которой выделена одна точка ∞ . Мы уже сталкивались с этим объектом, когда вводили понятие расширенной комплексной плоскости $\overline{\mathbb{C}}$ и определяли понятие вычета функции в бесконечно удаленной точке (см. раздел 2.4.7). Другое название \mathbb{CP}^1 — *сфера Римана*.

На рис. 3.7 мы видим, как можно непрерывным (топологическим) способом сопоставить сферу S^2 и пространство \mathbb{CP}^1 . Для этого нужно из верхней точки сферы (обозначена как северный полюс N) провести прямую через точку z' сферы до пересечения с плоскостью \mathbb{C} в точке z . Это соответствие непрерывно, т. к. малым отклонениям на сфере соответствуют малые отклонения на плоскости. Сама же точка N переводится в ∞ , поскольку ее проецирование происходит параллельно плоскости \mathbb{C} .

Проективное пространство \mathbb{CP}^2 — более сложный объект. Оно определяется как множество всех комплексных прямых в пространстве \mathbb{C}^3 , содержащих 0. Иначе говоря, его объектами являются прямые, определяемые пропорциями $z : w : 1$ (они отвечают за конечные точки), и прямые $z : w : 0$ (они отвечают за бесконечно удаленные точки). Первые можно трактовать как точки пространства \mathbb{C}^2 или же \mathbb{H} (все кватернионы), вторые — как точки пространства \mathbb{CP}^1 , топологически эквивалентного S^2 .

Внутри \mathbb{CP}^2 мы можем построить проекцию \mathbb{H} с центром в 0 на бесконечно удаленную компоненту этого пространства, поставив в соответствие каждой точке (z, w) (определяющей прямую в базовом \mathbb{C}^3) точку $z : w$. Поскольку при пропорциональных точках (z, w) и $(\lambda z, \lambda w)$ получается одна и та же бесконечно удаленная точка, проще сразу предполагать, что мы проецируем сферу S^3 , заданную уравнением $|z|^2 + |w|^2 = 1$, на сферу Римана (эквивалентную обычной сфере S^2) по правилу $(z, w) \mapsto z/w$ (при $w = 0$ получаем точку ∞).

При таком отображении все точки вида $(ze^{i\phi}, we^{i\phi})$ переходят в одну точку сферы Римана z/w , при том что на исходной сфере S^3 пробегается целая окружность. И обратно, каждой точке s на сфере Римана при нашем проективном отображении в исходной сфере S^3 соответствует ровно одна окружность $(ste^{i\phi}, te^{i\phi})$, где вещественное $t = 1/\sqrt{1 + |s|^2}$.

Вспоминая теперь, что сфера Римана топологически эквивалентна S^2 , мы получаем отображение из S^3 в S^2 , при котором над каждой точкой сферы S^2 в прообразе висит окружность S^1 . Такое отображение называется **расслоением Хопфа** сферы S^3 над сферой S^2 :

$$S^1 \hookrightarrow S^3 \rightarrow S^2.$$



Бернард
Риман

Чтобы представить его наглядно, рекомендуем посмотреть, например, [этот ролик](#) [iXkQ7YL2N54].

На этом мы расстаемся с проективным миром, чтобы коротко взглянуть на мир неевклидовой геометрии.

3.6.6 Нестандартные геометрии

Прежде чем двигаться дальше, необходимо разобраться в понятийном аппарате и аксиоматике геометрии.

Ранее мы подробно разбирали аксиоматику теории множеств Цермело–Френкеля и чуть-чуть коснулись аксиоматик некоторых числовых структур — групп, колец, полей, не отделяя их, правда, от самой теории множеств.

Сразу оговоримся, что история вопроса здесь очень богата и заслуживает обстоятельного изучения всеми, кто интересуется геометрией. Собственно говоря, математика как строгая наука началась именно с геометрических аксиом Евклида. Впоследствии эти аксиомы были детально изучены (в основном из-за пятого постулата о параллельных) и уточнены. На сегодняшний день существует несколько эквивалентных подходов к формализации геометрии, но чаще всего сами геометры предпочитают работать в готовом топологическом или линейном пространстве, а разбор аксиоматик оставить логикам.

Тем не менее, стоит подчеркнуть, что история аксиом Евклида очень напоминает историю аксиом теории множеств. Обе прошли этап «наивной» аксиоматики (одна 2 тысячи лет, вторая — несколько десятилетий), обе имеют конструктивную часть, постулирующую операции с конечными объектами, обе не могут дать полноценной картины мира без привлечения неконструктивной аксиомы, связанной с потенциальной бесконечностью (аксиома о параллельных в любом ее виде и аксиома бесконечного кардинала). То же самое можно сказать и об арифметике.

Эти три кита математики (арифметика, геометрия, теория множеств) по своему выстраивают конструктивный мир с неконструктивной надстройкой и великолепно согласуются в формальной логике и теории моделей, примерно на уровне континуума (если смотреть на иерархию универсумов множеств). Тем не менее, теория множеств, пожалуй, является наиболее простой и удобной в использовании формальной теорией, позволяющей укладывать логические утверждения в множества, т. е. в некие существующие объекты. Арифметика является самой «числовой», позволяя многие вещи сводить к вычислительным процессам. Но и арифметика, и теория множеств не являются полными в том смысле, что на их языке можно сформулировать такие утверждения, которые невозможно ни доказать, ни опровергнуть в этих теориях. Геометрия находится в более выигрышном положении, поскольку еще предложенная Гильбертом аксиоматика геометрии полна и непротиворечива (постольку, поскольку непротиворечива арифметика), таковыми же являются и

другие современные аксиоматики геометрии. Но вместе с тем она сложнее в плане языка и формулировки аксиом и беднее в плане построения трансфинитных конструкций и чисел (арифметику и теорию множеств погрузить в геометрию нельзя).

В этом разделе мы будем оперировать «школьными» аксиомами Евклида, постоянно имея ввиду готовую конструкцию линейного пространства \mathbb{R}^n . «Настоящая» формальная геометрия будет представлена читателю позже, в разделе 4.1.5. «Наивные» аксиомы Евклида имеют следующий вид:

- Eukl1 От всякой точки до всякой точки можно провести прямую (построение отрезка);
- Eukl2 Ограниченную прямую (отрезок) можно непрерывно продолжать по прямой (построение прямой);
- Eukl3 Из всякого центра и всяким раствором может быть описан круг (построение окружности);
- Eukl4 Все прямые углы равны между собой (инвариантность углов);
- Eukl5 Если прямая, падающая на две прямые, образует внутренние односторонние углы, в сумме меньшие двух прямых, то, продолженные неограниченно, эти две прямые встретятся с той стороны, где углы в сумме меньше двух прямых.

Первые три постулата позволяют производить построения циркулем и линейкой. Четвертый постулат впоследствии был выведен из остальных и, строго говоря, аксиомой не является. Пятый постулат — самое известное и сомнительное требование, имеющее богатую историю исследований.

Нужно подчеркнуть, что геометрия, прежде всего, реализует принцип сохранения длин и углов при движениях. При этом, определение угла невозможно без понятия движения. Так, прямая задает развернутый угол, а прямой угол — это такой, который, будучи совмещен со своей копией движением, задает развернутый (проще говоря, прямой — это половина развернутого). Таким способом можно задавать дробные части развернутого угла.³² При этом размер угла не зависит от масштаба, выбранного на линейке. Кроме того, аналогично можно определять и расстояния. Приняв какой-то отрезок за единицу, мы можем многократно его откладывать, а также дробить на части, тем самым определяя расстояние как кратчайший путь из мерных отрезков. При этом, показателем того, что три точки лежат на одной прямой, является

³²На эту тему существует обширная область исследований о построении циркулем и линейкой, достаточно глубоко изученная в 19-м веке. Здесь мы позволим себе обойти ее вниманием.

то, что расстояние между крайними точками есть в точности сумма расстояний от них до третьей (говорят, что тогда третья точка лежит между этими двумя).

Таким образом, определяя каким-то способом движение, мы можем разделить расстояние и меры углов.

Здесь нам нужно отметить два основных аспекта евклидовой геометрии:

1. возможность неограниченно расширять геометрические конструкции;
2. пятый постулат (аксиома параллельных).

Исключая эти два аспекта (заменяя их противоположными) вместе или по отдельности, можно получать различные геометрии.

Мы начнем с того, что по следам Гаусса, Лобачевского и Бойаи посмотрим на геометрию, в которой выполняется отрицание 5-го постулата в том смысле, что на плоскости через любую точку вне данной прямой можно провести более одной прямой, параллельной данной (под параллельностью понимается то, что они нигде не пересекаются).³³



Николай
Иванович
Лобачевский

Модель Клейна

Это — наиболее простая модель геометрии Лобачевского, называемая также *моделью Бельтрами*.

Рассмотрим на \mathbb{R}^2 единичный круг S и все проективные преобразования, биективно переводящие его в себя. Известна следующая

Теорема 3.13. *Существует проективное преобразование плоскости, переводящее круг в себя, а центр круга в любую внутреннюю точку этого круга.*

Доказательство можно найти, например, в [87]. Оно достаточно наглядное и сводится к тому, что над кругом мы строим конус с осью, перпендикулярной плоскости круга и проходящей через его центр, а затем начинаем наклонять конус, сдвигая его вершину так, чтобы она оказалась над нужной нам внутренней точкой. Остальное — дело техники.

Я нашёл
этому
поистине
чудесное
доказатель-
ство,
но...

Проективные преобразования по определению сохраняют двойное отношение. Поэтому именно через него определяется расстояние в данной модели. Для любых двух *внутренних* точек A, B круга S определим расстояние по

³³ Отметим, что в проективной геометрии мы наблюдали обратную ситуацию — любые две прямые пересекаются. Поэтому проективная геометрия также не является в полном смысле евклидовой, хотя включает ее в себя как часть пространства (все конечные точки).

формуле:

$$\rho_K(A, B) = \frac{1}{2} |\ln[M, N, A, B]| = \frac{1}{2} \left| \ln \frac{(A - M)(B - N)}{(A - N)(B - M)} \right|,$$

где M и N — точки пересечения (евклидовой) прямой, содержащей A и B , с границей круга S (см. рис. 3.8). Причем, неважно, с какой стороны находится M , с какой — N . От их расположения значение выражения $\rho_K(A, B)$ не зависит. Нормировочный множитель $1/2$ обычно выбирается (как это часто бывает в математике) из эстетических соображений, приводящих в последующих расчетах к более простым коэффициентам. Иногда выбирается коэффициент $c/2$, где параметр c играет роль радиуса кривизны геометрии и задает целое семейство геометрий Лобачевского.

Выражение, стоящее под знаком логарифма, всегда положительно, а переменная A и B местами приводит к его переворачиванию, т. е. выходу -1 за логарифм, что ликвидируется модулем. Поэтому $\rho_K(A, B)$ симметрично и неотрицательно. Несложно показать, что $\rho_K(A, B) = 0$ тогда и только тогда, когда $A = B$. Кроме того, выполняется неравенство треугольника: $\rho_K(A, C) + \rho_K(C, B) \geq \rho_K(A, B)$, причем равенство достигается ровно тогда, когда C лежит на прямой AB между точками A и B . Отсюда, в частности следует, что в модели Клейна прямыми являются хорды круга S . Таким образом, прямые исходного пространства являются также и прямыми в модели Клейна. С тем лишь ограничением, что бесконечная прямая укладывается в интервал. Действительно, $\rho_K(A, M) = \infty$.

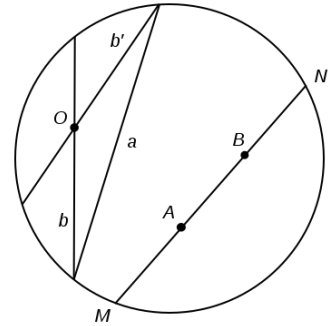


Рис. 3.8: Модель Клейна.

прямыми являются хорды круга S . Таким образом, прямые исходного пространства являются также и прямыми в модели Клейна. С тем лишь ограничением, что бесконечная прямая укладывается в интервал. Действительно, $\rho_K(A, M) = \infty$.

Углы же в данной модели совпадут с евклидовыми только в самом центре круга S .

Модель Клейна удовлетворяет всем аксиомам Евклида за исключением 5-го постулата. Действительно, для произвольной прямой a и точки O вне ее можно провести целый континуум прямых от b до b' (см. рис. 3.8), которые не пересекаются с a , т. е. параллельны ей.

Модель Пуанкаре

Данную модель геометрии Лобачевского можно построить на основе модели Клейна. Пусть задана сфера S^2 единичного радиуса, а в ее экваториальной плоскости лежит модель Клейна. Возьмем произвольную точку A в модели Клейна и спроецируем ее параллельно оси Oz на нижнюю полусферу S^2 в

точку A_1 , затем найдем точку A' пересечения отрезка NA_1 с экваториальной плоскостью xOy , как показано на рис. 3.9.

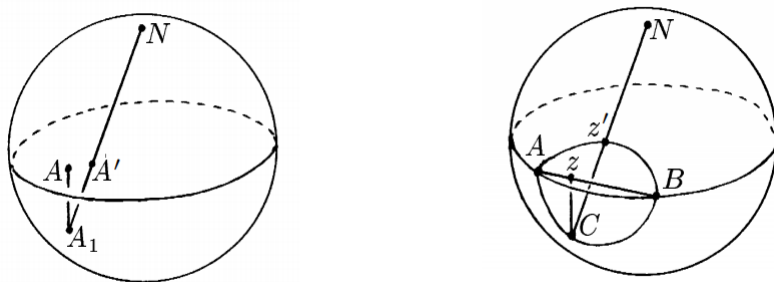


Рис. 3.9: Построение модели Пуанкаре в круге.

Соответствие $A \mapsto A'$ преобразует экваториальный круг в себя взаимно однозначно и непрерывно. При таком преобразовании хорда (прямая в модели Клейна) AB переходит в дугу на экваториальной плоскости ($z \mapsto z'$), которая представляет собой дугу окружности, причем в точках A и B она окажется перпендикулярна экватору. Тем самым, мы определенным образом выгнули хорду AB в сторону центра экваториального круга.

Таким образом, мы получаем некоторое преобразование (с помощью проекций) модели Клейна в некоторую новую модель, превращая хорды в дуги окружностей, перпендикулярные экватору.

После этого мы вводим в плоскости экватора комплексные координаты и для произвольных 4-х точек этой плоскости задаем двойное отношение (в арифметике комплексных чисел):³⁴

$$[z_1; z_2; z_3; z_4] \Rightarrow \frac{(z_3 - z_1)(z_4 - z_2)}{(z_3 - z_2)(z_4 - z_1)}.$$

Пусть тогда AB — хорда в модели Клейна, точки z и w лежат на ней, а z' и w' — их образы при рассматриваемом преобразовании экваториального круга. Можно показать следующее равенство:

$$[A; B; z; w] = |[A; B; z'; w']|^2.$$

Полагая теперь

$$\rho_P(z', w') \Rightarrow |\ln |[A; B; z'; w']||,$$

получаем, что

$$\rho_P(z', w') = \frac{1}{2} |\ln |[A; B; z; w]|| = \rho_K(z, w),$$

³⁴Ранее мы уже приводили тот факт, что такое отношение является вещественным тогда и только тогда, когда все точки лежат на одной прямой или окружности — см. теорему 2.19.

то есть расстояние в модели Клейна индуцирует расстояние в нашу новую модель, а дуги окружности, перпендикулярные экватору (в частности, диаметры экватора), представляют собой прямые в этой модели. Круг с заданным расстоянием ρ_P называется *моделью Пуанкаре в круге*.

Здесь мы видим, что прямые перестали быть евклидовыми прямыми (в отличие от модели Клейна), зато в данной модели углы сохраняются при движениях (угол между дугами — это угол между касательными в точке их пересечения). Поэтому такую модель геометрии Лобачевского называют также **конформной моделью** в круге.³⁵

Понимая, как связаны модели Клейна и Пуанкаре, легко увидеть прямые, параллельные данной — они перейдут в соответствующие дуги окружности (рис. 3.10).

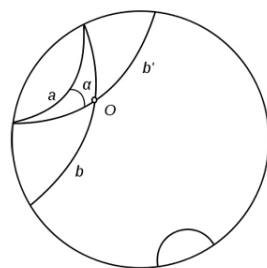


Рис. 3.10: Параллельность

Помимо рассмотренных нами двух моделей геометрии Лобачевского можно указать еще несколько известных моделей.

Прежде всего, это модель Пуанкаре в верхней полуплоскости \mathbb{C} , которая, в общем-то, получается из модели в круге с помощью специального отображения на комплексной плоскости (преобразование Кэли).

Другой моделью является гиперболическая модель в пространстве Минковского, построенная на верхней чаше гиперболида ($z > 0$), задаваемого равенством $x^2 + y^2 - z^2 = -1$. Прямые в такой модели получаются как сечения гиперболоида плоскостями, проходящими через 0, и представляют собой гиперболы (что и дает название данной модели и геометрии Лобачевского в целом).

Дадим несколько свойств геометрии Лобачевского:

- 1° любая инцидентная пара точка+прямая переводится в любую инцидентную пару точка+прямая движением (т. е. дробно-линейным преобразованием в евклидовых координатах);
- 2° треугольник полностью определяется углами (с точностью до движения);
- 3° если углы треугольника равны α, β, γ , то его площадь равна $\pi - (\alpha + \beta + \gamma)$ (если радиус кривизны c , то данное выражение умножается на c^2);
- 4° следствие предыдущего: сумма углов треугольника $< \pi$;
- 5° длина гиперболической окружности больше длины евклидовой окружности того же радиуса.

³⁵С понятием конформности мы уже сталкивались, рассматривая некоторые свойства комплексных функций.

Наличие моделей, в которых выполняются все аксиомы геометрии Евклида за исключением пятого постулата, означает, что отрицание пятого постулата логически совместимо с т.н. **абсолютной геометрией** (первые 4 постулата), а наличие стандартной модели (плоскость \mathbb{R}^2) означает, что пятый постулат также логически совместим с абсолютной геометрией. Таким образом, пятый постулат является независимым от абсолютной геометрии утверждением и может добавляться сам или его отрицание в зависимости от требований, предъявляемых к моделям пространства. Например, в теории относительности Эйнштейна гравитация считается следствием кривизны пространства, т. е. его геометрии, и поэтому более «правильной» геометрией физического пространства следует считать неевклидову геометрию. Более точно, геометрия пространства скоростей частиц в специальной теории относительности является геометрией Лобачевского [92].

В современных исследованиях (см., например, работы Сергея Сипарова [90]) производятся весьма успешные попытки с помощью геометрии специального вида наглядно и без привлечения сложных для восприятия сущностей смоделировать кватернионно-механические свойства материи, наблюдаемые в экспериментах (которые в классической литературе «объясняются» моделью Бора), а также полностью устранить такое эфемерное понятие в космологии как «темная материя», введенное для объяснения наблюдаемых свойств распределения материи на галактических масштабах, не укладывающихся в классическую теорию гравитации. Ведь если на минутку забыть о достижениях науки и посмотреть на мир «глазами» компьютера, то все, что мы имеем, — это серию раздражений нервной системы наблюдателя, вызываемую внешними раздражителями, среди которых можно выделить как других наблюдателей, так и явления неразумной среды (природы). Все, что мы хотим, — это упорядочить и отклассифицировать такие раздражения, пользуясь какой-то логической (алгоритмической) схемой с целью управлять этими раздражениями в угоду своему разуму. Таким образом, выбор адекватной и максимально простой для интуиции математической модели физических явлений (например, геометрии), вполне может быть инструментом такого упорядочения. Физики, конечно, будут пенять нам на физический смысл таких моделей, однако для главного героя нашей книги — компьютера — это понятие не является существенным.

3.6.7 Геометрия на сфере

Выше мы уже отмечали, что одним из аспектов евклидовой геометрии является возможность неограниченно расширять конструкции (например, чертить окружность сколь угодно большого радиуса), это свойство входит в определение абсолютной геометрии, однако у нас под ногами (в буквальном смысле) находится геометрия, в которой и это требование нарушается. Мы

говорим о **сферической геометрии**.

Рассмотрим уже излюбленную нами единичную сферу S^2 в пространстве \mathbb{R}^3 (ее уравнение: $x^2 + y^2 + z^2 = 1$) с центром O . Кратчайшим путем из точки A в точку B (в смысле метрики, индуцированной на эту сферу из объемлющего евклидова пространства) будет т.н. *дуга большого круга*, т.е. дуга окружности, полученной пересечением сферы плоскостью α , содержащей точки A , B и O (см. рис. 3.11).

Сфера S^2 с заданным таким способом расстоянием является моделью сферической геометрии. Отметим, что движения сферы, которые были нами ранее изучены, сохраняют это расстояние, т.е. являются движениями в сферической геометрии. И никаких новых движений при этом не возникает.

Прямыми в сферической геометрии служат окружности большого радиуса, т.е. сечения сферы плоскостью, проходящей через O . Ясно, что любые две прямые на сфере пересекаются в двух точках (представьте себе меридианы и экватор на глобусе). На рис. 3.11 две прямые пересекаются в точках P и Q . **Угол** между прямыми определяется как угол между касательными в точке пересечения прямых (на рис. 3.11 это угол θ в точке P). Соответственно, мы можем говорить о треугольниках и более сложных фигурах на сфере.

Окружностью называется множество точек, равноудаленных от заданной, коим является сечение сферы плоскостью (в частности, прямая является окружностью), причем ее радиус измеряется по сфере и, следовательно, всегда больше, чем радиус этой же окружности в \mathbb{R}^3 (на рис. 3.11 окружность с центром C и криволинейным радиусом r).

Если точки A, O, B не лежат на одной прямой, то плоскость α определяется однозначно, а дуга AB выбирается как меньшая из частей окружности. Длина дуги AB объявляется расстоянием между точками A и B на сфере и равна ϕ , где ϕ — мера угла AOB в радианах (в случае единичной сферы).

Если же A, O, B лежат на одной прямой, то точки A и B диаметрально противоположны, и расстояние между ними равно π .³⁶

Как и в проективной геометрии, определим точку и прямую как **инцидентные**, если данная точка лежит на данной прямой.³⁷ Больше того, как и в проективной геометрии здесь присутствуют **двойственные** прямая и пара точек (экватор и два соответствующих полюса), и выполняется та же самая

³⁶Поскольку радиус сферы равен 1, длины дуг совпадают с величиной их углов в радианах. Это упрощает все вычисления. Для того, чтобы масштабировать вычисления на сферу произвольного радиуса R , достаточно все линейные размеры умножать на R , а площади — на R^2 . R при этом называется радиусом кривизны сферической геометрии.

³⁷На самом деле, связь между проективной плоскостью и сферической геометрией намного глубже. Здесь мы точно так же можем рассматривать точки как пересечение прямых, проходящих через O , со сферой, а прямые — как пересечение плоскостей, проходящих через O , со сферой. С той только разницей, что каждая прямая задает две точки, а не одну. Поэтому проективная вещественная плоскость — это сфера с отождествленными диаметрально противоположными точками.

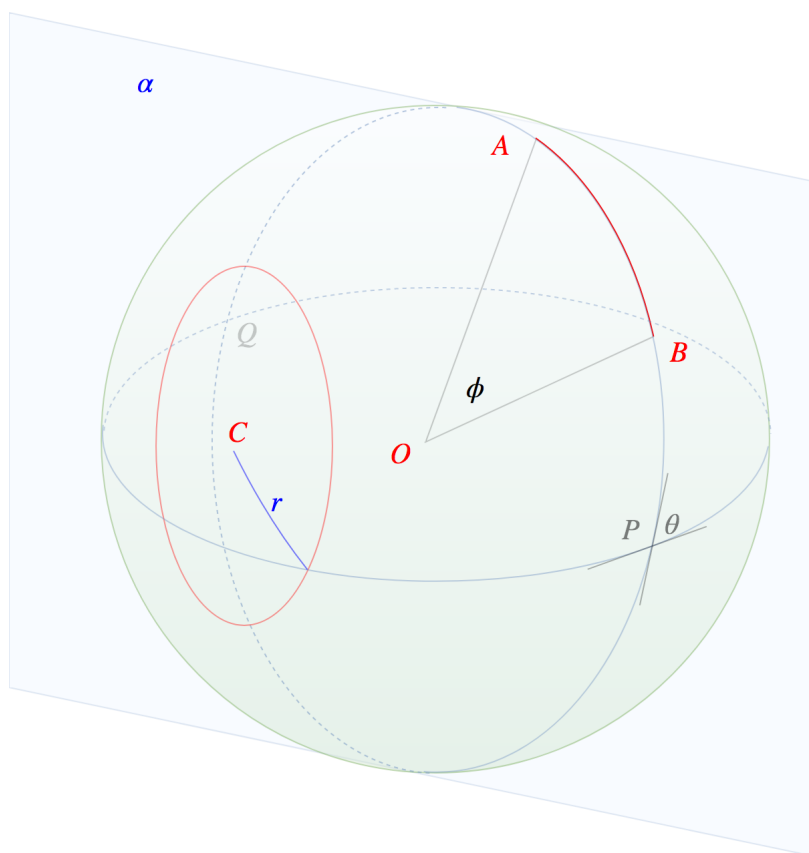


Рис. 3.11: Отрезок, прямая и окружность.

теорема 3.12 о двойственном соответствии, только двойственность принято называть **полярным соответствием**.

Кроме того, у каждой фигуры есть диаметрально противоположная (симметричная относительно центра O). Поэтому в задачах о двойственности всегда удобно рассматривать объекты парами. Например, у треугольника ABC (см. рис. 3.12) есть равный ему противоположный треугольник $A'B'C'$. Оба треугольника формируются тремя ограничивающими прямыми (на рисунке — синяя, желтая и зеленая), у каждой из которых есть пара двойственных точек-полюсов: у прямой $ABA'B'$ (желтой) это точки X и X' , у прямой $BCB'C'$ (синей) — Y и Y' , у $ACA'C'$ (зеленой) — Z и Z' . Таким образом, пара треугольников ABC и $A'B'C'$ двойственна другой паре треугольников XYZ и $X'Y'Z'$, причем стороны x, y, z треугольника ABC соответствуют углам при вершинах X, Y, Z , а углы при вершинах A, B, C — сторонам α, β, γ треугольника XYZ .

Это соответствие сторон и углов таково, что численно угол в радианах равен длине соответствующей стороны, вычитенной из π (на единичной сфере): $\angle X = \pi - x$, $\angle Y = \pi - y$, $\angle Z = \pi - z$. Вершины двойственного треугольника берутся из той же полусферы, в которой лежит исходный треугольник. Например, отрезок BC определяет синюю прямую с двумя полярными точками Y и Y' , но по одну сторону от синей прямой лежат точка A исходного треугольника и точка Y . В случае, когда треугольник составляет ровно $1/8$ сферы (все углы и все стороны равны $\pi/2$), он совпадает со своим двойственным (т. н. автополярный треугольник).

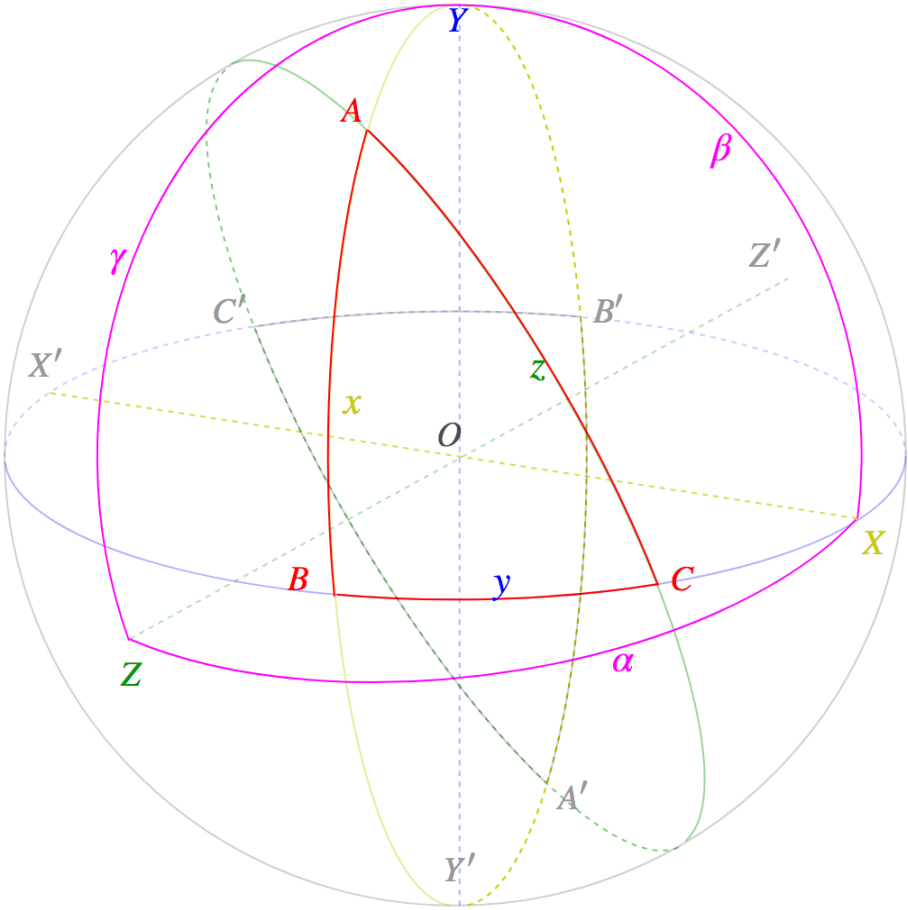


Рис. 3.12: Двойственность. «Штрихованные» точки — на задней полусфере.

Отсюда, в частности, следует, что каждый треугольник однозначно (с точностью до движения) задается только углами (т. к. они взаимно однозначно соответствуют сторонам двойственного треугольника). Перечислим некоторые свойства сферической геометрии:

- 1° любая инцидентная пара точка+прямая переводится в любую инцидентную пару точка+прямая движением сферы;
- 2° треугольник полностью определяется углами (с точностью до движения);
- 3° если углы треугольника равны α, β, γ , то его площадь равна $\alpha + \beta + \gamma - \pi$ (если радиус сферы R , то данное выражение умножается на R^2);
- 4° следствие 1 предыдущего: сумма углов треугольника $> \pi$;
- 5° следствие 2: периметр треугольника $< 2\pi$ ($2\pi R$);
- 6° длина сферической окружности меньше длины евклидовой окружности того же радиуса;
- 7° не существует окружности радиуса больше $\pi/2$ ($R\pi/2$), причем равенство достигается в случае прямой;
- 8° никакая область сферы не изометрична никакой области плоскости.

Последнее свойство выражает известную проблему карт земной поверхности. Проблема заключается в том, что достаточно большой участок поверхности планеты невозможно нарисовать на плоскости без искажения расстояний (для малых участков это либо незаметно, либо они сами по себе являются плоскими).

Геометрия на сфере, также как и геометрия евклидова пространства, допускает естественное обобщение на высшие размерности (сфера S^n в пространстве \mathbb{R}^{n+1}). Гиперплоскостью в S^n является сфера S^{n-1} с тем же центром. Соответственно, симметрии в S^n рассматриваются относительно таких гиперсфер.

Несмотря на то, что три геометрии — евклидова, сферическая (и родственная ей проективная плоскость) и Лобачевского (она же — гиперболическая) — отличаются друг от друга, и местами довольно сильно, они, тем не менее, имеют очень похожие фундаментальные законы, зависящие лишь от радиуса кривизны этих геометрий, который у евклидовой плоскости равен 0, у плоскости Лобачевского (и вообще любой гиперболической геометрии) отрицателен (пропорционален i^2), а у сферической (и вообще любой эллиптической римановой) геометрии — положителен.

Гиперболическая геометрия поставляет нам соответствующую гиперболическую тригонометрию с функциями

$$\operatorname{sh} t = \frac{e^t - e^{-t}}{2} = -i \sin(it), \quad \operatorname{ch} t = \frac{e^t + e^{-t}}{2} = \cos(it),$$

которые формально совпадают с таковыми же функциями обычной тригонометрии при соответствующей замене действительной переменной на чисто мнимую.

Теорему синусов во всех трех геометриях можно сформулировать одинаково:

$$\frac{l(a)}{\sin \alpha} = \frac{l(b)}{\sin \beta} = \frac{l(c)}{\sin \gamma},$$

где $l(r)$ — длина окружности радиуса r , стороны a, b, c противолежат углам α, β, γ , соответственно.

Теорема косинусов:

$$\begin{aligned} \cos \frac{a}{R} &= \cos \frac{b}{R} \cos \frac{c}{R} + \sin \frac{b}{R} \sin \frac{c}{R} \cos \alpha, \\ \operatorname{ch} \frac{a}{R'} &= \operatorname{ch} \frac{b}{R'} \operatorname{ch} \frac{c}{R'} - \operatorname{sh} \frac{b}{R'} \operatorname{sh} \frac{c}{R'} \cos \alpha, \end{aligned}$$

первая — для сферической геометрии, вторая — для геометрии Лобачевского. Нетрудно видеть, что одна в другую переходят заменой $R = iR'$. В евклидовой геометрии, как мы помним со школы, теорема косинусов имеет несколько другой вид: $a^2 = b^2 + c^2 - 2bc \cos \alpha$, но ее легко получить, умножив обе части равенства на R^2 и переходя к пределу при $R \rightarrow \infty$, пользуясь асимптотикой для тригонометрических функций. Так что и тут мы видим единство трех геометрий.

Дифференциал длины дуги в некоторых специально выбранных внутренних координатах:

$$ds^2 = du^2 + \cos^2 \left(\frac{u}{R} \right) dv^2; \quad ds^2 = du^2 + \operatorname{ch}^2 \left(\frac{u}{R'} \right) dv^2$$

для сферической и гиперболической геометрий, соответственно. И снова, обе формулы переходят друг в друга заменой $R = iR'$, а переходя к пределу при $R \rightarrow \infty$, получаем евклидов случай $ds^2 = du^2 + dv^2$.

Площадь треугольника выражается общей формулой:

$$S = R^2(\alpha + \beta + \gamma - \pi),$$

где в случае сферической геометрии $R > 0$, а в случае геометрии Лобачевского $R = iR'$, $R' > 0$ (что дает $S = (R')^2(\pi - \alpha - \beta - \gamma)$).

Для евклидовой геометрии здесь трудно усмотреть какую-либо известную формулу площади треугольника в пределе при $R \rightarrow \infty$. Тем не менее, зная дифференциал длины дуги, можно увидеть, что такой предельный переход возможен. Действительно, $\alpha + \beta + \gamma - \pi = \Delta\alpha + \Delta\beta + \Delta\gamma$, где каждая дельта — это разница между углом сферического треугольника ABC и соответствующим ему углом евклидова треугольника ABC (на рис. 3.13 сферический треугольник ABC выделен красным). Поэтому в силу теоремы синусов дельту углов можно оценить дельтой противолежащих им сторон, которая из предыдущего оценивается величиной порядка $1 - \cos^2(u/R)$ (u стремится к

a , b или c — одной из сторон евклидова треугольника), а это, в свою очередь, величина порядка $1/R^2$. Таким образом, $R^2(\Delta\alpha + \Delta\beta + \Delta\gamma)$ имеет конечный предел. Из геометрических построений ясно, что он должен быть равен площади евклидова треугольника ABC .³⁸

Вообще, при замене R на iR все метрические формулы геометрии Лобачевского (сохраняющие при этой замене геометрический смысл) переходят в соответствующие формулы сферической геометрии. При $R \rightarrow \infty$ те и другие дают в пределе формулы евклидовой геометрии (либо теряют смысл). Стремление к бесконечности величины R означает, что масштабный отрезок является бесконечно малым по сравнению с радиусом кривизны. То обстоятельство, что при этом формулы неевклидовой геометрии переходят в пределе в формулы евклидовой геометрии, означает, что для малых (по сравнению с радиусом кривизны) неевклидовых фигур соотношения между их элементами мало отличны от евклидовых.

Постепенный подъем от евклидовых пространств к аффинным, проективным (и далее — к неевклидовым геометриям) в очередной раз демонстрирует нам *архетип неограниченного расширения* математического знания. Каждый раз, когда мы видим перед собой более-менее законченную (разработанную, изученную, понятую) область алгебры, геометрии, анализа или иной, более сложной ветви, мы хотим отыскать некий естественный (в данный момент) путь ее обобщения и расширения. Отыскав такой путь и наметив способы его осмысления, мы тут же оглядываемся назад и с удивлением обнаруживаем новые свойства прежних, якобы досконально изученных областей. И здесь срабатывает второй архетип — *архетип трансцендентного восприятия*, о котором мы уже говорили ранее.

Примеры рассмотренных геометрий являют собой наглядный материал для сравнения бытовой и математической интуиции. Так, бытовая интуиция противится неевклидовой геометрии и не может признавать дуги окружностей прямыми. Даже несмотря на то, что сферическую геометрию мы с легкостью обнаруживаем на глобусе. Математическая интуиция, наоборот, нащупывает самую суть понятия (прямая — это кратчайшее расстояние, а расстояние должно удовлетворять таким-то аксиомам), после чего оперирует этой сутью, отвлекаясь от ее конкретных реализаций. Если математическая интуиция сработала верно, то мы получаем красивейшие теории, в которых

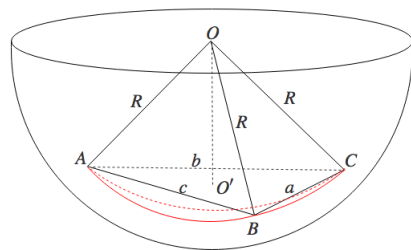


Рис. 3.13: Треугольник.

Сопряженный архетип?

³⁸ Для более точных выкладок следует рассмотреть пирамиду $ABCO$ и посчитать площадь евклидова треугольника ABC как сумму площадей проекций $AO'B$, $BO'C$, $AO'C$ граней этой пирамиды, каждая из которых равна $0.5R^2 \sin(\delta)$, где δ — угол при вершине O данной пирамиды. Здесь мы также видим три слагаемых с множителем R^2 .

основные понятия, доступные нам по бытовой интуиции, находятся в тех же самых (или подобных) отношениях друг с другом, хотя «физически» представляют собой нечто совершенно иное. Таким образом, математическая интуиция позволяет нащупывать новые закономерности там, где бытовая интуиция просто не в состоянии ничего разглядеть.

Возвращаясь к началу раздела о геометрии, вспомним об иерархии инвариантов: (Inv1) сохранение длин, (Inv2) сохранение отношений, (Inv3) сохранений двойных отношений, (Inv4) сохранение операций, (Inv5) сохранение связности, (Inv6) сохранение мощности множества.

С первыми тремя мы разобрались. Более того, рассматривая двойное отношение как меру длины (через логарифм), мы обнаружили, что на ее основе возникают более интересные геометрии (Лобачевского). Одновременно с этим мы поняли, что расстояние можно задавать не только как длину вектора в евклидовом пространстве, и даже вскользь упомянули аксиомы метрики. К общему понятию метрического пространства мы еще вернемся, когда будем вести речь об инварианте непрерывности, ключевой «фишке» топологических пространств. Сохранение мощности было в достаточной мере освещено в разделе о кардинальных числах, но то и дело этот инвариант появляется в любом разделе математики — от комбинаторики до топологии.

Рассмотрение 4-го типа инвариантов предполагает изучение неких движений, сохраняющих набор заданных операций. Но такими движениями являются, как мы уже отмечали, различные изоморфизмы, связывающие разнородные математические структуры. Этот инвариант также достаточно общий для того, чтобы повсеместно встречаться в математике, и его проявления мы еще не раз встретим.

Тем не менее, мы сделаем определенный логический шаг, переходя от конечных векторных пространств к счетно-мерным алгебрам многочленов (см. стр. 228), хотя и посмотрим на них больше с алгебраической, нежели чем с геометрической точки зрения.

3.7 Многочлены

Для погружения в данную тему помимо классических книг [55, 59, 60] рекомендуем читателю книгу [75], а также курс видеолекций д.ф.-м.н. А. Савватеева «Теория Галуа» [Rir4DM3Y9hE].

Многочленом от одной переменной x над коммутативным кольцом K с единицей³⁹ называется всякое выражение вида

$$P_n(x) = c_0 + c_1x + \dots + c_{n-1}x^{n-1} + c_nx^n,$$

³⁹Всюду в этом разделе мы считаем рассматриваемые кольца коммутативными и с единицей.

где символ x называется переменной, а числа $c_0, \dots, c_n \in K$. С такого рода записями мы уже сталкивались, когда говорили о канторовской нормальной форме для ординалов. При этом коэффициенты (ординалы) у нас не были элементами кольца. Тем не менее, часто такую форму записи называют многочленом безотносительно природы составляющих ее переменных и констант.

Каждой такой записи естественным образом взаимно однозначно сопоставляется последовательность $(c_0, \dots, c_n, 0, 0, \dots) \in K^\omega$ с хвостом нулей.

Поэтому формально кольцо многочленов определяется как множество функций из ω в K с конечным носителем. Напомним, что именно так определяется групповое кольцо, только областью определения функций там является некоторая группа. Так что можно считать, что кольцо многочленов над K задается как часть группового кольца $K[\mathbb{Z}]$, где все функции отличны от нуля лишь для неотрицательных аргументов. Проще говоря, $K[\mathbb{N}]$ — кольцо многочленов над K с операциями

$$(f + g)(n) = f(n) + g(n), \quad (fg)(n) = \sum_{k+m=n} f(k)g(m), \quad (\alpha f)(n) = \alpha f(n),$$

где $f, g \in K[\mathbb{N}]$, $\alpha \in K$. Эти операции согласуются с представлением многочлена в виде $P_n(x)$, если x считать «каким-то» числом. Чтобы подчеркнуть, что все многочлены мы хотим записывать в привычном виде от переменной x , кольцо многочленов принято обозначать $K[x]$.⁴⁰

Кроме того, напомним, что многочлены образуют счетно-мерную алгебру над кольцом K , а также могут рассматриваться как часть счетно-мерного векторного пространства (см. раздел 3.1.5).

Если мы теперь сделаем небольшой шаг назад и вспомним наши упражнения в геометрии, то можем отметить важный качественный скачок, когда мы переходили от линейных функций (отвечающих за поворотные гомотетии и сдвиги) к дробно-линейным (отвечающим за проективные преобразования). Аналогично определяется и понятие **дробно-рациональной функции** как отношение двух многочленов:

$$R_{n,m}(x) \doteq \frac{c_0 + c_1x + \dots + c_{n-1}x^{n-1} + c_nx^n}{d_0 + d_1x + \dots + d_{m-1}x^{m-1} + d_mx^m}.$$

При этом операции с ними определяются ровно так же, как операции над рациональными дробями. Предполагается при этом, что в знаменателе стоит не тождественный ноль (т. е. хотя бы один d_i отличен от нуля).

Нетрудно убедиться в том, что дробно-рациональные функции образуют

⁴⁰Обозначение не слишком удачное, но вполне устоявшееся.

поле (их сложение, вычитание, умножение и деление не выводят за пределы такого вида выражений). Это поле принято обозначать $K(x)$.⁴¹

Т. е. кольцо квадратное, а поле круглое? :-/

3.7.1 Конечные алгебраические расширения

Если кольцо K вкладывается в какое-то более широкое кольцо (или поле) F , как, например, $\mathbb{Z} \subset \mathbb{R}$ или $\mathbb{Q} \subset \mathbb{R}$ или $\mathbb{Q} \subset \mathbb{C}$, то, взяв какое-то множество $S \subseteq F$, можно определить **расширение кольца** K с помощью присоединения к нему множества S . Обозначим через $K[S]$ минимальное подкольцо F , содержащее множества K и S . При этом под минимальностью следует понимать теоретико-множественный минимум по отношению вложения базовых множеств этих колец. Аналогично, $K(S)$ есть минимальное подполе F , содержащее K и S (при этом предполагается, что K и F — поля).

И здесь снова мы встречаемся с архетипами *порождающего элемента* и *базового множества*!

Нетрудно показать, что $K[S]$ есть пересечение всех подколец F , содержащих K и S как подмножества (аналогично — для $K(S)$). При этом нужно иметь ввиду, что от внешнего кольца F в данном определении (в отличие от определения $K[x]$) мы не можем отказаться, поскольку оно связывает элементы множеств S и K операциями. Вполне может быть, что один и тот же элемент $s \in S$ в другом объемлющем кольце F' может обладать совсем другими алгебраическими свойствами, например, вместо уравнения $s^2 = -1$ он будет обладать свойством $s^4 = -1$, и в таком случае расширение $K[S]$ может быть неизоморфно тому, которое получается внутри F . Поэтому всегда важно понимать, в какой числовой вселенной мы находимся, и не терять ориентиры.

Легко видеть, что как $K[x]$, так и $K[S]$ являются алгебрами над K , поэтому мы можем говорить об их размерности, как векторных пространств над K . Если эта размерность конечная, то такая алгебра называется **конечным расширением** кольца K . Если E является расширением K , то E обозначают $E \supset K$ ($K \subset E$) или E/K . Мы будем пользоваться первым обозначением, чтобы не путать расширение с факторизацией. Размерность расширения (мощность базиса) $E \supset K$ обозначается $[E : K]$ и называется **степенью расширения**. Например, $[\mathbb{C} : \mathbb{R}] = 2$, $[\mathbb{Q}[\sqrt{2}, \sqrt{3}] : \mathbb{Q}] = 4$, $[\mathbb{Z}[i] : \mathbb{Z}] = 2$, $[\mathbb{Z}[w] : \mathbb{Z}] = 2$, $[\mathbb{R} : \mathbb{Q}] = 2^\omega$, $[\mathbb{R}[x] : \mathbb{R}] = \omega$.

Далее, если имеются конечные расширения $F \supset E \supset K$, образующие, как говорят, **башню расширений**, то их размерности связаны жестким арифметическим правилом.

⁴¹Это поле также является полем частных кольца $K[x]$, поскольку кольцо $K[x]$ есть область целостности.

Теорема 3.14 (о башне). Если имеются конечные расширения $F \supset E \supset K$, то

$$[F : E] \cdot [E : K] = [F : K].$$

Доказательство. Пусть $n = [E : K]$ и e_1, \dots, e_n — базис E над K , $[F : E] = m$ и f_1, \dots, f_m — базис F над E . Тогда векторы $f_1 e_1, \dots, f_1 e_n, \dots, f_m e_1, \dots, f_m e_n$ порождают все пространство $F \supset K$. С другой стороны, они линейно независимы над K , т. к. если предположить, что существует нетривиальная линейная комбинация

$$k_{11}f_1e_1 + \dots + k_{1n}f_1e_n + \dots + k_{m1}f_me_1 + \dots + k_{mn}f_me_n = 0$$

с коэффициентами $k_{ij} \in K$, то мы получим нетривиальную линейную комбинацию

$$(k_{11}e_1 + \dots + k_{1n}e_n)f_1 + \dots + (k_{m1}e_1 + \dots + k_{mn}e_n)f_m = 0,$$

которая зануляет базис $F \supset E$,⁴² а это противоречит определению базиса.

Таким образом, произведения $f_i e_j$ перечисляют некоторый базис $F \supset K$, значит, $[F : K] = mn$. \square

Определение расширения как минимального кольца (поля) — это *теоретико-множественное определение* и весьма общее. Но можно выбрать и другой подход, конструктивный: создавать элементы алгебры над K при помощи алгебраических операций, производимых над элементами множества S и кольца K . Поскольку эти операции сводятся к сложению, вычитанию и умножению (в случае поля — еще и делению), элементы алгебры можно генерировать с помощью полиномов над K с произвольным (конечным) количеством переменных:

$$P_{n,\dots,m}(x, \dots, y) = \sum_{i \leq n, \dots, j \leq m} x^i \dots y^j k_{i,\dots,j},$$

где коэффициенты $k_{i,\dots,j} \in K$ (в случае поля — дробно-рациональных функций от многих переменных). А именно, пусть задано множество $S \subseteq F$ в кольце F , содержащем K , тогда

$$E = \{P_{n,\dots,m}(s, \dots, t) \mid s, \dots, t \in S\}.$$

Легко видеть, что E — кольцо (сумма и произведение полиномов дает снова полином над K), содержащее в себе K . Следовательно, оно содержит и минимальное кольцо, т. е. $K[S] \subseteq E$. В то же время, поскольку $K \cup S \subseteq K[S]$

⁴²То, что эта комбинация также нетривиальна, следует из того, что e_i образуют базис $E \supset K$.

и $K[S]$ — кольцо, то в кольце $K[S]$ лежат и все числа вида $s^i \dots t^j k_{i,\dots,j}$, $k_{i,\dots,j} \in K$, $s, \dots, t \in S$, и их суммы. Но тогда $E \subseteq K[S]$. Таким образом, мы видим, что оба определения расширения K с помощью множества S дают одну и ту же алгебру над K . Этим удобно пользоваться в различных ситуациях. Аналогично устанавливается равенство понятий минимального поля и поля, полученного с помощью дробно-рациональных функций от многих переменных.

Пусть $K \subseteq F$ и $\alpha \in F$. Если при некотором многочлене $P_n \in K[x]$ имеет место равенство $P_n(\alpha) = 0$ (т. е. этот многочлен *занимает* α над K), то число α называется **алгебраическим** над K , в противном случае α называется **трансцендентным** над K . В частности, число $z \in \mathbb{C}$ называется алгебраическим, если оно алгебраическое над \mathbb{Q} , трансцендентным — если оно трансцендентное над \mathbb{Q} .

Для произвольного расширения $E \supset K$ говорят, что оно является **алгебраическим расширением** K , если все его элементы — алгебраические над K .

Наконец, расширение $K[S]$ (и $K(S)$) называется **конечно порожденным**, если S — конечное непустое множество. В том случае, когда S состоит из одного элемента α , конечно порожденное расширение $K[\alpha]$ ($K(\alpha)$) называется **простым**. Конечно порожденное расширение колец можно построить как башню простых расширений:

$$K[\alpha, \dots, \beta] = K[\alpha] \dots [\beta],$$

последовательно присоединяя к полученным кольцам элементы S . Действительно, представляя элементы данных расширений в виде полиномов от степеней α, \dots, β , легко показать, что эти множества совпадают. Чуть позже мы сможем установить, что аналогичное равенство справедливо и для полевых расширений $K(\alpha, \dots, \beta)$.

Комментарий 10.

Здесь мы можем перекинуть мостик к числам, определяемым с помощью матриц (см. раздел 3.2). А именно, рассмотрим матрицы вида

$$M(x, y) = \begin{pmatrix} x & 2y \\ y & x \end{pmatrix}$$

Это — матричное представление числа $x + y\sqrt{2}$, $x, y \in \mathbb{Q}$. Заметим, что оно отличается от представления комплексного числа тем, что вместо $-y$ написано $2y$, что как раз и определяется добавлением новой компоненты i такой, что $i^2 = 2$ (в случае \mathbb{C} : $i^2 = -1$).

Проверьте, что обычные матричные операции совпадают в данном случае с операциями над числами в поле $\mathbb{Q}(\sqrt{2})$. При этом, поскольку $x, y \in \mathbb{Q}$, в данной алгебре нет делителей нуля, т. е. не требуется какая-либо дополнительная

Введенные нами понятия обладают следующими свойствами.

Теорема 3.15.

- (1) Простое алгебраическое расширение поля является конечным;
- (2) Конечное расширение алгебраично;
- (3) Конечное расширение является конечно порожденным;
- (4) Алгебраическое конечно порожденное расширение поля конечно.

Доказательство. Рассмотрим простое алгебраическое расширение $K[\alpha]$ поля K . Так как оно алгебраическое, то существует зануляющий α многочлен из $K[x]$:

$$k_0 + k_1\alpha + k_2\alpha^2 + \cdots + k_n\alpha^n = 0, \quad k_n \neq 0,$$

откуда $\alpha^n = -(k_0 + k_1\alpha + k_2\alpha^2 + \cdots + k_{n-1}\alpha^{n-1})/k_n$ (K — поле!). Любой элемент $x \in K[\alpha]$ можно представить в виде

$$x = q_0 + q_1\alpha + q_2\alpha^2 + \cdots + q_m\alpha^m, \quad q_i \in K,$$

и если степень $m > n - 1$, то, используя представление для $k_n\alpha^n$, полученное выше, можно за конечное число шагов понизить эту степень до $n - 1$:

$$x = r_0 + r_1\alpha + r_2\alpha^2 + \cdots + r_{n-1}\alpha^{n-1}, \quad r_i \in K$$

Таким образом, x раскладывается по векторам $1, \alpha, \alpha^2, \dots, \alpha^{n-1}$. (1) доказано.

Для доказательства (2) достаточно заметить, что если $x \in E$ и $[E : K] = n$, то векторы $1, x, x^2, \dots, x^n$ линейно зависимы над K , а это и означает, что x зануляется некоторым многочленом над K .

Докажем (3). Пусть $[E : K] = n$ и s_1, \dots, s_n — базис. Необходимо показать, что базис порождает E , т. е. $E = K[S]$. Поскольку каждый элемент E является линейной комбинацией элементов базиса, ясно, что $E \subseteq K[S]$. С другой стороны, поскольку E — кольцо, содержащее $K \cup S$, оно содержит в себе минимальное кольцо $K[S]$, т. е. $E \supseteq K[S]$.

(4). Пусть расширение $E \supset K$ алгебраическое и конечно порожденное, т. е. $E = K[\alpha, \dots, \beta]$. Из доказанного ранее мы можем представить E в виде $K[\alpha] \dots [\beta]$ как башню простых расширений. Далее, всякий элемент, алгебраический над K , очевидно, алгебраичен и над каждым этажом этой башни. Поэтому в силу (1) расширение каждого следующего этажа является конечным расширением над предыдущим. Отсюда по теореме о башне, используя индукцию, нетрудно получить, что расширение $E \supset K$ конечное над K . \square

Отметим, что в пунктах (1) и (4) существенным требованием является то, что K — поле.

3.7.2 Свойства многочленов

Фундаментальнейшее свойство многочленов над кольцом выражается следующей теоремой.

Теорема 3.16 (Безу́ [Bezout]). Пусть K — кольцо, и $P \in K[x]$ — многочлен над этим кольцом. Тогда существует многочлен $Q \in K[x]$ такой, что

$$P(x) = (x - c)Q(x) + P(c)$$

для любого $c \in K$.

Мы предлагаем читателю самостоятельно разобраться в том, что в кольце многочленов возможно деление с остатком так, что остаток от деления будет иметь степень ниже, чем делитель. Аналогичные свойства ординалов, чисел Гаусса и Эйзенштейна мы уже рассматривали выше. После этого доказательство теоремы Безу становится тривиальным.

Из теоремы Безу, в частности, следует, что число $c \in K$ является корнем уравнения $P(x) = 0$ тогда и только тогда, когда P нацело делится на $(x - c)$. Отметим, что мы в данном случае не говорим о корнях, лежащих в каких-либо расширениях K , тем не менее, иногда формулировка теоремы Безу приводится в предположении, что x пробегает некоторое расширение кольца K , но в этом случае корень c может выпасть из K , а многочлен Q — из $K[x]$. Например, $x^2 + 1 \in \mathbb{R}[x]$, но нацело делится на $x - i$, причем получаем $Q = x + i \in \mathbb{C}[x]$.

В том случае, когда K не является полем, может оказаться так, что многочлен $P_n(x)$ имеет корней больше, чем n . Это связано с тем, что над кольцом разложение многочлена на множители бывает неоднозначным (не выполняется основная теорема алгебры).

Например, над кольцом \mathbb{Z}_8 выполняется равенство многочленов $x^2 - 1 = (x - 7)(x - 1) = (x - 3)(x - 5)$ при всех $x \in \mathbb{Z}_8$. Так что многочлен $x^2 - 1$ имеет четыре корня: 1, 3, 5, 7. При этом он остается многочленом второй степени. Такое обилие корней, как мы видим, объясняется неоднозначностью его разложения на простые множители. Еще пример: уравнение $x^2 + 1 = 0$ имеет бесконечно много решений в кватернионах (вся мнимая единичная сфера).

Многочлены над полем

Однако, в том случае, когда K — поле, ситуация становится более привлекательной. Алгебра многочленов в этом случае напоминает алгебру гауссовых чисел. Кольцо многочленов над полем является целостным (не имеет делителей нуля) и евклидовым (роль нормы в нем выполняет степень многочлена),

и в нем выполняется теорема о делении с остатком произвольного многочлена P на произвольный многочлен V в том виде, в каком мы ее встречали ранее (см. раздел 3.3.1):

$$P = QV + R, \quad \deg R < \deg V,$$

причем частное Q и остаток R определяются однозначно и принадлежат тому же кольцу, что и исходные многочлены P и V .⁴³ Кроме того, в таком кольце многочленов выполняется алгоритм Евклида, позволяющий определить НОД двух многочленов. А значит, не составит труда определить взаимно простые многочлены как такие, у которых НОД является делителем 1.

Аналогично гауссовым числам многочлены можно считать ассоциированными, если один из другого получается умножением на делитель единицы. При этом делителями единицы в $K[x]$ являются все многочлены-константы, отличные от нуля (т.е. многочлены нулевой степени). Простым или **неприводимым над K** называется многочлен из $K[x]$, который не делится нацело ни на какой неассоциированный с ним многочлен степени > 0 и не является делителем 1 (т.е. сам имеет степень > 0). Простых многочленов над любым полем бесконечно много. К таковым, например, относятся все многочлены степени 1 (в случае бесконечного поля их уже бесконечно много, а в случае конечного поля бесконечность простых многочленов доказывается ровно так же, как бесконечность простых чисел в арифметике).

Ввиду наличия деления с остатком основная теорема арифметики 2.6 и 3.7 имеет место и в кольце многочленов над полем.

Теорема 3.17 (о разложении многочленов). *Если K — поле, то каждый многочлен из $K[x]$ степени > 0 раскладывается в произведение конечного числа неприводимых многочленов, причем это разложение определяется однозначно с точностью до ненулевых множителей из K .*

Совершенно аналогично выполняется и лемма 3.3 о взаимно простых многочленах.

Неприводимость многочленов сильно зависит от базового поля K . Вспомним, что в обычной арифметике число 5 является простым, а в числах Гаусса раскладывается на множители вида $2 \pm i$. Точно также, многочлен $x^2 + 1$ неприводим над \mathbb{R} , но раскладывается на множители вида $x \pm i$ над полем \mathbb{C} . Из основной теоремы алгебры следует, что над \mathbb{C} приводим любой многочлен степени > 1 . Известно также, что над \mathbb{R} любой многочлен можно разложить в произведение многочленов 1 и 2 степени. Например, $x^4 + 4 = (x^2 - 2x + 2)(x^2 + 2x + 2)$, а дальнейшее разложение над \mathbb{R} невозможно.

Имеет место следующая теорема, полностью аналогичная теореме 3.4 для целых чисел.

⁴³Напомним, что степень многочлена-константы ($n = 0$) считается равной 0, если эта константа отлична от 0, и $-\infty$, если она равна нулю (по закону логарифма).

Теорема 3.18. Пусть $f \in K[x]$ — ненулевой многочлен. Тогда следующие утверждения равносильны:

- (1) f — неприводимый над K ;
- (2) идеал (f) — простой;
- (3) идеал (f) — максимальный.

Отсюда и из теоремы 3.3 следует, что факторизация по главному идеалу (f) , порожденному неприводимым многочленом, является полем. Например, факторизация $\mathbb{R}[x]$ по идеалу $(x^2 + 1)$ будет полем, причем $\mathbb{R}[x]/(x^2 + 1) \cong \mathbb{C}$, каждый класс эквивалентности в $\mathbb{R}[x]/(x^2 + 1)$ является комбинацией классов $a[x^2 + 1] + b[x]$, где $[x^2 + 1]$ соответствует единице, а $[x]$ — мнимой единице, $a, b \in \mathbb{R}$.⁴⁴

Короче говоря, мы видим, что многочлены не зря отнесены нами в главу о числах. Их арифметика очень похожа на обычную арифметику и позволяет чувствовать себя довольно свободно, но нужно постоянно держать в уме некоторую «гауссовость» этих чисел, связанную с ассоциированностью (которая, впрочем, сильно проще, чем в случае чисел Гаусса или Эйзенштейна).

Есть здесь, конечно и свои нюансы, с которыми мы еще столкнемся в дальнейшем.

Наконец, предполагая, что K является полем, установим тесную связь между $K[\alpha]$ и $K(\alpha)$.

Теорема 3.19. Если число α алгебраическое над полем K , то (1) $K[\alpha] = K(\alpha)$ и (2) размерность $[K[\alpha] : K]$ равна степени минимального над K многочлена, аннулирующего α .

Если число α трансцендентное над полем K , то (3) размерность $[K[\alpha] : K]$ счетная.

Доказательство. Чтобы показать (1), достаточно доказать, что любая рациональная дробь в точке α может быть представлена как многочлен в этой точке. Точнее, пусть $I(x)$ — минимальный многочлен такой, что $I(\alpha) = 0$ (минимальность подразумевается в смысле невозможности понизить степень I), для определенности полагая, что старший коэффициент в $I(x)$ равен 1 (такие многочлены называют *нормализованными*). Пусть также дана дробь

$$R(x) = \frac{Q(x)}{P(x)},$$

где $P, Q \in K[x]$ и $P(\alpha) \neq 0$. Пользуясь евклидовостью кольца $K[x]$, мы можем найти НОД многочленов $P(x)$ и $I(x)$, причем он будет выражаться в виде

⁴⁴На самом деле существует более общий результат: если K — кольцо, и K^* — кольцо невырожденных матриц вида $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$, то $A^* \cong A[x]/(x^2 + 1)$.

(алгоритм Евклида)

$$\gcd(x) = I(x)k_1(x) + P(x)k_2(x).$$

При этом полученный НОД есть многочлен степени не больше 0, т.к. иначе он бы делил $I(x)$, но тогда $I(x)$ не был бы минимальным. Следовательно $\gcd(x) \equiv \beta$ (константа из поля K). Кроме того, $\beta \neq 0$, поскольку НОД не может быть нулем (это также видно из алгоритма Евклида). Но тогда

$$0 \neq \beta = \gcd(\alpha) = I(\alpha)k_1(\alpha) + P(\alpha)k_2(\alpha) = P(\alpha)k_2(\alpha),$$

поскольку α — корень $I(x)$. Ясно, что $k_2(\alpha) \neq 0$, и тогда $P(\alpha) = \beta/k_2(\alpha)$.

Следовательно, $R(\alpha) = Q(\alpha)k_2(\alpha)/\beta$, т.е. рациональная дробь представлена как значение многочлена из $K[x]$.

Чтобы установить (2), необходимо показать, что $[K[\alpha] : K] = \deg I(x)$. Рассмотрим числа $1, \alpha, \dots, \alpha^n$, где $n = \deg I(x)$. Ясно, что они линейно зависимы, поскольку $I(\alpha)$ является линейной комбинацией, зануляющей их. Следовательно, $[K[\alpha] : K] \leq \deg I(x)$.

С другой стороны, если размерность меньше n , то существует многочлен меньшей степени, зануляющий α , в противоречии с минимальностью $I(x)$.

Таким образом, $[K[\alpha] : K] \geq \deg I(x)$.

Для доказательства (3) достаточно заметить, что все векторы $1, \alpha, \alpha^2, \dots, \alpha^n, \dots$ линейно независимы, иначе мы бы нашли многочлен, зануляющий α . То, что базис $K[\alpha]$ не более чем счетен, следует из того, что любой элемент этого пространства может быть разложен по векторам $1, \alpha, \alpha^2, \dots$ \square

Известно, что число π трансцендентное, поэтому в силу пункта (3) расширение $\mathbb{Q}[\pi]$ имеет счетную размерность над \mathbb{Q}

В дополнение к этой теореме скажем, что можно показать также следующие утверждения:

- 1) $K(\alpha) = \{c_0 + c_1\alpha + \dots + c_{k-1}\alpha^{k-1} \mid c_i \in K\}$, где $k = [K(\alpha) : K]$;
- 2) $K(\alpha) = K[\alpha] \cong K[x]/(I)$, где α — алгебраическое над K число, $I(x)$ — минимальный многочлен, зануляющий его;
- 3) $K(\alpha) \cong K(x)$ и $K[\alpha] \cong K(x)$, где α — трансцендентное над K число.

Покажем первое свойство. Необходимо показать, что любой многочлен степени выше k можно редуцировать до k -ой степени в точке α . Пусть $I(x)$ — минимальный многочлен, зануляющий α , и пусть элемент поля $K(\alpha)$ задан формулой $Q(\alpha) = q_0 + q_1\alpha + \dots + q_n\alpha^n$. Пользуясь делением с остатком, получаем, что

$$Q(x) = I(x)T(x) + R(x) = R(x),$$

где $\deg R < \deg I = k$. Таким образом, все элементы поля $K(\alpha)$ (а мы показали выше, что и дробные рациональности сводятся к многочленам) записываются как многочлены степени не выше k в точке α .

Из доказанной теоремы следует также, что

$$K[\alpha] \dots [\beta] = K(\alpha) \dots (\beta). \quad (3.18)$$

Наконец, заметим, с одной стороны, что $K(\alpha) \dots (\beta)$ — поле, содержащее в себе поле $K(\alpha, \dots, \beta)$. А с другой стороны, по установленному ранее тождеству, $K(\alpha) \dots (\beta) = K[\alpha, \dots, \beta]$ — кольцо, содержащееся в поле $K(\alpha, \dots, \beta)$. Следовательно, для конечно порожденных полей мы получаем равенство, аналогичное таковому для конечно порожденных колец:

$$K(\alpha, \dots, \beta) = K(\alpha) \dots (\beta).$$

Если числа α, \dots, β — алгебраические над K , то мы к тому же получаем алгебраическое конечно порожденное расширение, которое по теореме 3.15 будет конечным пространством над K , размерность которого помогает вычислить теорема о башне 3.14.

Поле алгебраических чисел

В дальнейшем нас в основном будут интересовать многочлены над полем \mathbb{Q} (с некоторыми поправками можно считать их многочленами над \mathbb{Z}) или его расширениями. Неприводимость для них становится еще более частым явлением.

Имеют место следующие свойства неприводимости многочленов:

- 1° Неприводимый над \mathbb{Z} многочлен неприводим и над \mathbb{Q} ;
- 2° (Критерий Эйзенштейна)⁴⁵ Пусть $f(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_0$ над \mathbb{Z} , и для некоторого простого p выполняются свойства:

- (a) a_n не делится на p ;
- (b) a_i ($i < n$) делятся на p ;
- (c) a_0 не делится на p^2 .

Тогда $f(x)$ неприводим над \mathbb{Z} ;

- 3° (следствие) $x^{p-1} + x^{p-2} + \dots + 1$ неприводим над \mathbb{Q} и \mathbb{Z} при любом простом числе p .

⁴⁵В [75] доказывается более общее утверждение. Вместо \mathbb{Z} рассматривается произвольное коммутативное кольцо с 1, вместо простого числа p — простой идеал P , а делимость (или неделимость) коэффициентов на p заменяется принадлежностью к соответствующему идеалу.

Размерность $\mathbb{Q}[\alpha]$ над \mathbb{Q} равна минимальной степени неприводимого над \mathbb{Q} многочлена, корнем которого является α .

В частности, $\sqrt[3]{2}$ является корнем $x^3 - 2$, неприводимого над \mathbb{Q} в силу критерия Эйзенштейна. Поэтому $[\mathbb{Q}[\sqrt[3]{2}] : \mathbb{Q}] = 3$ (см. ниже задачу об удвоении куба).

Рассматривая вложение $\mathbb{Q} \subset \mathbb{C}$, мы можем определить множество алгебраических чисел \mathbb{A} . Это все такие числа, которые являются корнями (в поле \mathbb{C}) многочленов с рациональными (целыми) коэффициентами.

Теорема 3.20.

- (1) \mathbb{A} является полем;
- (2) \mathbb{A} алгебраически замкнуто;
- (3) \mathbb{A} имеет счетную размерность над \mathbb{Q} .

Доказательство. Чтобы показать (1), необходимо убедиться, что для любых чисел $\alpha, \beta \in \mathbb{A}$ их сумма, разность, произведение и частное (при $\beta \neq 0$) также лежат в \mathbb{A} . Это достаточно легко сделать, рассмотрев поле $\mathbb{Q}[\alpha, \beta]$. Нужно показать, что $\mathbb{Q}[\alpha, \beta] \subset \mathbb{A}$. Как уже отмечалось ранее, $\mathbb{Q}[\alpha, \beta] = \mathbb{Q}[\alpha][\beta]$, т. е. это башня полей $\mathbb{Q} \subset \mathbb{Q}[\alpha] \subset \mathbb{Q}[\alpha][\beta]$, так что размерность $[\mathbb{Q}[\alpha][\beta] : \mathbb{Q}]$ конечная, следовательно, все элементы $\mathbb{Q}[\alpha, \beta]$ являются алгебраическими, т. е. принадлежат \mathbb{A} .

Чтобы доказать (2), нужно доказать, что все алгебраические над \mathbb{A} числа лежат в \mathbb{A} . Рассмотрим произвольный многочлен $A(x) = a_0 + a_1x + \dots + a_nx^n$ с коэффициентами из \mathbb{A} . Пусть x_0 — его корень в \mathbb{C} , т. е. алгебраическое над \mathbb{A} число (если корня нет, то это константа, которая не определяет никакое алгебраическое над \mathbb{A} число). Рассмотрим поле

$$K = \mathbb{Q}[a_0, a_1, \dots, a_n, x_0].$$

Это — конечное расширение \mathbb{Q} , содержащее x_0 (т. к. a_i — алгебраические над \mathbb{Q} , а x_0 — алгебраическое над $\mathbb{Q}[a_0, \dots, a_n]$). Значит, существует многочлен из $\mathbb{Q}[x]$, зануляющий x_0 . Поэтому $x_0 \in \mathbb{A}$.

Наконец, (3) следует из того, что многочленов с рациональными коэффициентами существует счетное множество, а значит, и корней они дают также счетное множество, так что само \mathbb{A} счетно и любой его базис не более чем счетен. С другой стороны, базис не может быть конечным, т. к. \mathbb{A} содержит сколь угодно высокие башни полевых расширений \mathbb{Q} . \square

Ранее мы уже отмечали тот факт, что конечно порожденное алгебраическое расширение и конечное расширение (размерность $< \omega$) суть одно и то же (теорема 3.15). Но еще более привлекательным является следующий факт. Мы его докажем только для полей, содержащих \mathbb{Q} и лежащих в \mathbb{C} , но на самом деле это свойство носит более общий характер (в частности, выполняется в конечных полях).

Теорема 3.21 (о примитивном элементе). *Любое конечно порожденное алгебраическое расширение является простым алгебраическим расширением, т. е. для любых алгебраических чисел $\alpha, \beta, \dots, \eta$ над K найдется такое алгебраическое θ , что*

$$K[\alpha, \beta, \dots, \eta] = K[\theta].$$

Доказательство. Рассмотрим теорему для случая только двух ступенек расширения: $K[\alpha, \beta]$. Ясно, что общий случай выводится отсюда по индукции.

Пусть P — минимальный многочлен над K , зануляющий α , и Q — минимальный многочлен над K , зануляющий β . Пусть $\alpha = \alpha_0, \alpha_1, \dots, \alpha_n$ — все различные корни P в поле \mathbb{C} , $\beta = \beta_0, \beta_1, \dots, \beta_m$ — все различные корни Q в поле \mathbb{C} . Эти корни (кроме самих α, β) могут быть кратными.

Возьмем такое число $\delta \neq 0$ из K , что для всех i, j, k, l будем иметь $\alpha_i + \delta\beta_j \neq \alpha_k + \delta\beta_l$. Это всегда возможно, т. к. количество отношений $(\alpha_i - \alpha_k)/(\beta_l - \beta_j)$ конечное, а поле K — бесконечное (здесь мы пользуемся предположением, что поле K расширяет \mathbb{Q} в \mathbb{C}). Пусть, кроме того, $\theta = \alpha + \delta\beta$. Ясно, что $\theta \in K[\alpha, \beta]$ и, следовательно, $K[\theta] \subseteq K[\alpha, \beta]$.

Докажем обратное вложение. Пусть $H(x) = P(\theta - \delta x)$. Отметим, что $H \in K[\theta][x]$. Тогда β — корень H и корень Q , причем в Q это некрратный корень (в силу минимальности Q). Это значит, что линейный многочлен $(x - \beta)$ является НОД H и Q (других общих корней, кроме β , у H и Q нет в силу выбора параметра δ). Но тогда, как во всяком евклидовом кольце, имеет место равенство

$$(x - \beta) = H(x)k_1(x) + Q(x)k_2(x),$$

где все многочлены принадлежат $K[\theta][x]$ (так как $P, Q \in K[x]$, а $H \in K[\theta][x]$), следовательно, и двучлен $(x - \beta)$ — тоже. Таким образом, $\beta \in K[\theta]$.

Отсюда следует, что и $\alpha = \theta - \delta\beta \in K[\theta]$. Вместе с предыдущим получается, что $K[\alpha, \beta] \subseteq K[\theta]$.

Далее по индукции легко доказать, что для произвольного конечного алгебраического расширения $K[\alpha, \dots, \beta]$ найдется такое $\theta = k\alpha + \dots + m\beta$ с коэффициентами $k, \dots, m \in K$, что $K[\alpha, \dots, \beta] = K[\theta]$. \square

Пример: $\mathbb{Q}[\sqrt{2}, \sqrt{3}] = \mathbb{Q}[\sqrt{2} + \sqrt{3}]$.

Доказанная теорема вместе с теоремой 3.15 показывает, что конечное расширение и простое алгебраическое расширение — суть одно и то же. Результат относится не только к кольцевым расширениям поля, но и к полевым, теперь уже в силу полученного выше равенства (3.18).

3.7.3 О построениях циркулем и линейкой

Одной из классических задач геометрии является изучение вопроса о том, какие геометрические построения можно произвести, имея только циркуль и

линейку. Первый позволяет строить окружности заданного радиуса, вторая — соединять любые две заданные точки и неограниченно продлевать отрезок за его границы. Все в точности с первыми тремя постулатами Евклида. При этом предполагается, что у нас есть некий мерный отрезок, задающий масштаб (единицу длины), а следовательно, вопрос о построениях циркулем и линейкой сводится к умению строить отрезки различной длины или, по-просту, строить числа.

Числа при этом получаются как длины отрезков, соединяющих получаемые при построениях точки пересечения линий — прямых и окружностей. Ясно, что такие пересечения могут давать только числа, являющиеся решениями линейных и квадратных уравнений. Иначе говоря, имея единицу, мы можем строить все рациональные числа (по теореме Фалéса), затем все рациональные комбинации различных корней из рациональных чисел, затем корней из этих корней и т.д. Речь идет, конечно же, о квадратных корнях.

В качестве упражнения предлагаем читателю самостоятельно — Упражнение 3.44. построить циркулем и линейкой отрезки длины $\sqrt{2}$ и $\sqrt{3}$. А чтобы задача не казалась сложной, скажем, что юный Гаусс в начале 19 века построил таким способом правильный 17-угольник впервые в истории математики. Это построение любопытно посмотреть в динамике, например, [тут](https://en.wikipedia.org/wiki/Constructible_polygon) (https://en.wikipedia.org/wiki/Constructible_polygon). Там же смотрите построение 15-, 257- и 65537-угольников.

Возникает вопрос: можно ли построить таким способом ребро куба, объемом которого равен 2, т.е. число $\sqrt[3]{2}$ (это античная задача об удвоении куба, известная наравне с задачей о квадратуре круга)?

Чтобы ответить на него, заметим, как уже говорилось выше, что построить мы можем только числа, попадающие в расширения поля \mathbb{Q} , получаемые при помощи последовательного присоединения корней из чисел, полученных на предыдущих шагах. Иначе говоря, для каждого конкретного числа, которое можно построить с помощью циркуля и линейки, можно указать башню расширений $\mathbb{Q} = P_0 \subset P_1 \subset \dots \subset P_n$, где каждое следующее поле P_{k+1} получается как $P_k[x_k]$, где x_k — корень квадратного уравнения с коэффициентами из P_k .

Теперь фокус: $[P_{k+1} : P_k] = 2$ для всех k (мы опускаем случай, когда размерность не повышается при присоединении x_k , т.к. в этом нет смысла). Но тогда по теореме 3.14 о башне имеем: $[P_n : \mathbb{Q}] = 2^n$.

С другой стороны, если бы число $\sqrt[3]{2}$ можно было построить циркулем и линейкой, то нашлась бы такая башня полей $\mathbb{Q} = P_0 \subset P_1 \subset \dots \subset P_n$, что $\mathbb{Q}[\sqrt[3]{2}]$ лежало бы в P_n , т.е. имела бы место башня $\mathbb{Q} \subset \mathbb{Q}[\sqrt[3]{2}] \subset P_n$ и, в силу теоремы о башне, $[P_n : \mathbb{Q}[\sqrt[3]{2}]] \cdot [\mathbb{Q}[\sqrt[3]{2}] : \mathbb{Q}] = [P_n : \mathbb{Q}] = 2^n$.

Но это последнее невозможно, поскольку размерность $[\mathbb{Q}[\sqrt[3]{2}] : \mathbb{Q}] = 3$, а 3 не делит 2^n . Следовательно, удвоить куб циркулем и линейкой невозможно.

Здесь мы предлагаем читателю в качестве упражнения самостоятельно показать, что $\mathbb{Q}[\sqrt[3]{2}] = \{a + b\sqrt[3]{2} + c\sqrt[3]{4} \mid a, b, c \in \mathbb{Q}\}$ и имеет размерность 3 над \mathbb{Q} , не используя теорему 3.19.

Упражнение
3.45.

3.7.4 Нормальные расширения

Комментарий 11. О нормальности :)

Здесь стило бы, наверное, отметить еще одну закономерность математики, которую, однако, не хочется возводить в ранг архетипа. А именно, употребление слова «нормальность», которое характеризует обычно такие свойства математических объектов, которые наиболее удобны в контексте определенной задачи или выражают наиболее комфортные, привычные для нас свойства.

Например, нормальные группы напоминают простые числа и вообще дают много чего интересного, скажем, через факторизацию. Нормализованные векторы лежат на сфере единичного радиуса и упрощают многие вычисления. Нормаль — перпендикулярный вектор, образует прямой угол с прямой или плоскостью. Нормальное распределение — то, что получается в больших однородных статистических выборках и часто встречается в природе. Норма — числовая величина, упорядочивающая изучаемые объекты и позволяющая сводить их свойства к свойствам действительных или даже натуральных чисел. Нормальная форма — выражение (алгебраическое или логическое), имеющее определенный канонический вид.

Здесь мы в очередной раз увидим нечто нормальное, а именно: расширение поля.

Скажем, что расширение $K \subset K[\alpha, \dots, \beta]$ является **нормальным**, если числа α, \dots, β представляют собой исчерпывающий список корней некоторого многочлена над K , т. е. если многочлен $(x - \alpha) \cdots (x - \beta)$ после раскрытия всех скобок представляет собой многочлен из $K[x]$. Например, расширение $\mathbb{Q}[\sqrt[3]{2}]$ не является нормальным, т. к. многочлен $x^3 - 2$ имеет комплексные корни, заведомо не принадлежащие этому расширению. В то же время, расширение $\mathbb{Q}[\sqrt{2}]$ нормально, т. к. многочлен $x^2 - 2$ имеет корни $\pm\sqrt{2}$, а они оба лежат в $\mathbb{Q}[\sqrt{2}]$, т. е. $\mathbb{Q}[\sqrt{2}] = \mathbb{Q}[\sqrt{2}, -\sqrt{2}]$.

Легко показать, что если имеет место башня расширений $K \subset E \subset F$ и расширение $K \subset F$ нормальное, то таково же и расширение $E \subset F$. При этом нижний этаж $K \subset E$ может не быть нормальным (тривиальный пример: $\mathbb{Q} \subset \mathbb{Q}[\sqrt[3]{2}] \subset \mathbb{C}$).

Будем называть башню $K \subset E \subset F$ **нормальной**, если все три расширения $K \subset E$, $E \subset F$ и $K \subset F$ нормальные.

Нормальность расширений играет очень важную роль при изучении авто-

морфизмов полей.⁴⁶

Дадим определение. **Автоморфизмом поля** F называется всякая биекция $f : F \leftrightarrow F$, сохраняющая операции поля (сложение, вычитание, умножение и деление). Автоморфизм 0 переводит в 0, 1 — в 1. Автоморфизмы поля F образуют группу с операцией композиции, обозначаемую $\text{Aut}(F)$. **Автоморфизмом поля F над полем K ($K \subset F$),** или K -автоморфизмом поля F , называется всякий автоморфизм $f : F \rightarrow F$ такой, что $f|_K = \text{id}$. То есть K -автоморфизм оставляет все точки K на месте.

В точности как $\text{Aut}(G)$ для группы G .

Нормальность алгебраического расширения $K \subset E$ эквивалентна каждому из двух свойств:

- (N1) каждый неприводимый многочлен $P \in K[x]$, имеющий хотя бы один корень в E , разлагается в E на линейные множители (иначе говоря, E является полем разложения некоторого множества многочленов из $K[x]$);
- (N2) любой мономорфизм $E \hookrightarrow K^*$ (т.е. инъекция, сохраняющая операции, или *изоморфное вложение*) поля E в алгебраическое замыкание K^* , сохраняющие точки K на месте, является K -автоморфизмом E .

При этом под *алгебраическим замыканием* поля K понимается минимальное поле K^* , в котором все многочлены из $K[x]$ разложимы (на линейные множители). Поскольку мы договорились считать, что у нас все поля находятся между \mathbb{Q} и \mathbb{C} , а точнее, все они являются алгебраическими расширениями поля \mathbb{Q} , то для них алгебраическим замыканием будет поле \mathbb{A} . Поэтому в нашем частном случае нормальность расширения $K \subset E$ означает, что любой мономорфизм $f : E \rightarrow \mathbb{A}$ является автоморфизмом E , т.е. $f \in \text{Aut}(E)$.

Группа K -автоморфизмов расширения $K \subset E$ называется **группой Галуа** и обозначается $\mathbf{G}(E/K)$.

3.7.5 Элементы классической теории Галуа

Здесь мы пойдем самым простым путем для объяснения теории Галуа, рассматривая задачу о нахождении радикальных формул для решения степенных уравнений.

⁴⁶ Существует также термин «расширение Галуа», которое представляет собой нормальное и сепарабельное расширение $K \subset E$, однако в нашем случае (расширения \mathbb{Q} внутри \mathbb{C}) сепарабельность расширения следует из того, что все рассматриваемые поля имеют характеристику 0, так что здесь расширение Галуа и нормальное расширение суть одно и то же.

Пусть нам дано уравнение $ax^2+bx+c=0$. Каждый школьник знает, как решить это уравнение по формуле дискриминанта.

Для уравнения третьей степени вида $x^3+px+q=0$ (к такому виду приводится любое уравнение третьей степени) также существует общая формула **Кардано** для выражения корня через коэффициенты p, q :



Эварист
Галуа

$$x = \sqrt[3]{-\frac{q}{2} + \sqrt{Q}} + \sqrt[3]{-\frac{q}{2} - \sqrt{Q}}, \quad Q = \left(\frac{p}{3}\right)^3 + \left(\frac{q}{2}\right)^2,$$

причем комплексные значения кубических корней здесь выбираются так, что их произведение равно $-p/3$. Найдя один корень, два другие можно найти, решая квадратное уравнение.

Наконец, для уравнения 4-ой степени также существуют формулы, которые выводятся сведением исходного уравнения к уравнению третьей степени, и носят имя **Феррари**.

А для 5-ой
степени —
Ламборгини?

Долгое время математики пытались получить общие формулы и для уравнений высших степеней, но ничего не получалось, пока в начале 19 века не пришел Эварист Галуа и не сказал, что это вообще невозможно!

Рассуждения Галуа чем-то схожи с нашими построениями циркулем и линейкой. А именно, анализируя задачу, мы поняли, что построить циркулем и линейкой можно только числа, удовлетворяющие определенным уравнениям — первой и второй степени, а значит, те, которые можно выразить с помощью алгебраических операций и квадратных корней. В случае решения уравнения $a_0 + a_1x + \dots + a_nx^n = 0$ с целыми (читай: рациональными) коэффициентами нахождение финитной алгебраической формулы для его корней означает умение построить конечную формулу с помощью операций: сложение, вычитание, умножение, деление, возведение в степень и извлечение корней целой степени.

Таким образом, взяв некоторое специальное уравнение 5-ой степени, нужно проверить, всегда ли возможно такое построение его корней. И если нет, то такое уравнение является контрпримером к общей формуле, а значит, и общей формулы не существует!

Осталось отыскать уравнение и методику проверки.

Теорема 3.22. Если расширение $K \subset E$ — нормальное, то группа автоморфизмов $G(E/K)$ имеет ровно столько элементов, какова степень расширения $[E : K]$.

Доказательство. Пусть $E = K[\alpha, \dots, \beta]$ — расширение, включающее корни многочлена $(x - \alpha) \dots (x - \beta)$, т.е. E является полем разложения данного многочлена и, следовательно, нормальным расширением K . По теореме 3.21

о примитивном элементе существует θ в виде линейной комбинации корней такой, что $K = K[\theta]$.

Пусть $[E : K] = n$. Ясно, что $\theta^0, \theta^1, \dots, \theta^{n-1}$ образуют базис K , т. к., с одной стороны, большее количество степеней будет линейно зависимым (их будет больше n), а с другой стороны, линейная зависимость перечисленных векторов будет означать наличие меньшего зануляющего многочлена для θ , что уже противоречит теореме 3.19.

Далее. Пространство $E \supset K$ состоит из значений всех многочленов из $K[x]$ степени не выше n в точке θ . Нетрудно показать, что если f — K -автоморфизм E , то для любого многочлена P в точке θ будем иметь $f(P(\theta)) = P(f(\theta))$ просто по свойствам K -автоморфизма. Но тогда θ является корнем многочлена P тогда и только тогда, когда $f(\theta)$ является корнем того же самого многочлена P . Таким образом, если P — минимальный многочлен, зануляющий θ , то любой K -автоморфизм просто переставляет его корни местами (или оставляет на месте).

Пусть далее x_1, \dots, x_n — корни такого многочлена P (минимального зануляющего θ). Этих корней ровно n в силу примитивности расширения $K[\theta]$. Предположим, что автоморфизм f_k переводит θ в x_k ($k = \overline{1, n}$). Это условие полностью определяет поведение автоморфизма в любой точке E , т. к. на базисных векторах имеем $f_k(\theta^m) = f(\theta)^m = x_k^m$, а далее действие f_k однозначно распространяется на все элементы E в силу их разложения по данному базису. Это же соображение доказывает не только единственность, но и существование каждого из автоморфизмов f_1, \dots, f_n .

Предположим далее, что есть еще какой-то K -автоморфизм g , который переводит θ в некоторую точку $x_0 \in E$. Но по доказанному выше x_0 должен быть корнем многочлена P , т. е. одним из x_1, \dots, x_n . Следовательно, g совпадает с одним из K -автоморфизмов f_1, \dots, f_n . То есть автоморфизмы f_k исчерпывают всю группу $G(E/K)$. \square

В ходе доказательства мы обнаружили, что любой K -автоморфизм корни произвольного многочлена $P \in K[x]$ переводит в его же корни, т. е. на корнях P он действует как перестановка. При этом, несложно показать, что если P раскладывается в произведение двух взаимно простых многочленов над K , то автоморфизм не может перепутывать их корни. В самом деле, пусть $P = QR$ и K -автоморфизм f переводит корень x_1 многочлена Q в корень y_1 многочлена R . Но тогда $Q(y_1) = Q(f(x_1)) = f(Q(x_1)) = 0$, т. е. y_1 является корнем Q в противоречии с предположением о взаимной простоте Q и R . Поэтому любой K -автоморфизм работает как перестановка на корнях только лишь каждой из неприводимых частей разложения многочлена P .

В то же время, если P неприводим (и нелинеен) над K , то любой его корень x_1 может быть переведен в любой другой его корень x_2 с помощью некоторого K -автоморфизма. Действительно, поля $K[x_1]$ и $K[x_2]$ изоморфны друг другу

(это следует из того, что все их элементы — суть многочлены над K в точке x_1 и x_2 , соответственно, а значит, сконструировать такой изоморфизм можно непосредственно), а изоморфизм этих двух полей можно расширить до мономорфизма (изоморфного вложения) $K[x_1, \dots, x_n] \hookrightarrow \mathbb{A}$, где x_1, \dots, x_n — все корни P . Но тогда в силу нормальности $K \subset K[x_1, \dots, x_n]$ и свойства (N2) нормального расширения такой мономорфизм окажется K -автоморфизмом поля $K[x_1, \dots, x_n]$. Это значит, что существует как минимум n автоморфизмов поля разложения многочлена P над K , причем каждый корень можно перевести в любой другой корень P каким-то автоморфизмом.

Из полученного, в частности, следует, что если $P = Q_{n_1} \dots Q_{n_j}$, где Q_{n_k} имеют степень k , неприводимы над K и не содержат кратных корней, то количество K -автоморфизмов не превышает числа $n_1! \dots n_j!$ и не меньше числа $n_1 \dots n_j$. Например, если $P = (x^2 - 2)(x^2 - 3)$, то количество \mathbb{Q} -автоморфизмов поля $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ равно $2!2! = 4$. То есть группа Галуа $G(\mathbb{Q}[\sqrt{2}, \sqrt{3}]/\mathbb{Q})$ имеет порядок 4. Более того, эта группа изоморфна четверной группе Клейна V_4 , т. к. она жонглирует корнями каждого из двучленов $x^2 - 2$ и $x^2 - 3$ независимо друг от друга. Так что мы нашли еще одно замечательное воплощение V_4 , которое можно добавить к ее свойствам на стр. 215.

Пусть теперь $P(x) = x^5 - 6x + 3$ и $E = \mathbb{Q}[x_1, x_2, x_3, x_4, x_5]$, где x_i — корни P . Очевидно, что E является нормальным расширением \mathbb{Q} . Обозначим $G = G(E/\mathbb{Q})$. Многочлен P неприводим над \mathbb{Q} в силу критерия Эйзенштейна. Следовательно, существует как минимум 5 и не более $5! = 120$ автоморфизмов в группе G . Покажем, что на самом деле $|G| = 120$.

Для начала заметим, что P имеет три вещественных корня (это легко проверить методами школьного Анализа) и два комплексных. Пусть x_1, x_2, x_3 — вещественные корни, а $x_4 = \bar{x}_5$ — комплексные. Наличие двух комплексно сопряженных корней позволяет сразу же найти простой автоморфизм, а именно: комплексное сопряжение. Оно оставляет первые три корня на месте, а 4-ый и 5-ый меняет местами.

Договоримся перестановки корней (которые осуществляют автоморфизмы группы G) записывать как обычные перестановки через номера этих корней. Таким образом, комплексное сопряжение будет соответствовать транспозиции (45). Кроме того, для простоты и сами корни будем обозначать их индексами. Это позволяет естественным образом построить вложение $G \hookrightarrow S_5$, так что в дальнейшем будем считать, что G есть подгруппа S_5 . Наша задача, таким образом, сводится к тому, чтобы показать, что $G = S_5$.

Далее, для любых двух корней i и j скажем, что они эквивалентны, $i \sim j$, если $(ij) \in G$. Легко видеть, что это отношение эквивалентности: $(i) = e$ (тождественный автоморфизм), если $(ij) \in G$, то $(ji) = (ij) \in G$ (обратный автоморфизм), и если $(ij), (jk) \in G$, то $(ij)(jk)(ij) = (ik) \in G$ (композиция автоморфизмов). Следовательно, множество корней разбивается на классы эквивалентности, которые называются **транзитивными классами**. В том

$|| \cdot ||$ — это
мощность
множества.

случае, если класс всего один, то множество корней называется **транзитивным**, а если их несколько — **интранзитивным**.

Поскольку существует автоморфизм (45), можно утверждать $4 \sim 5$, т. е. как минимум один транзитивный класс содержит как минимум 2 элемента. Предположим, что 1 находится в другом классе. Рассмотрим какой-нибудь автоморфизм σ_{14} , совершающий подстановку $1 \mapsto 4$ (про другие точки нам ничего неизвестно). Такой автоморфизм существует, как было отмечено ранее. Пусть, кроме того, этот автоморфизм переводит $i \mapsto 5$ (неизвестное $i \neq 1$). Тогда композиция $\sigma_{14}^{-1}(45)\sigma_{14}$ представляет собой транспозицию $(1i)$, и во втором классе находится не менее 2-х элементов.

Пусть в классе с элементами 4 и 5 есть еще элемент 3, т. е. существует автоморфизм (34). Тогда существует и автоморфизм $(345) = (34)(45)$. Взяв автоморфизм σ_{13} , переводящий $1 \mapsto 3$ и $j \mapsto 4$, $k \mapsto 5$ ($j \neq 1 \neq k$, $j \neq k$) построим транспозиции $\sigma_{13}^{-1}(34)\sigma_{13} = (1j)$ и $\sigma_{13}^{-1}(45)\sigma_{13} = (jk)$, т. е. в этом случае и во втором классе оказывается не менее 3-х элементов.

Таким образом, какой бы мы класс ни взяли, в остальных классах находится не меньше элементов, чем в исходном. Это по-просту означает, что все транзитивные классы равномощны. При этом они образуют в сумме множество из 5 элементов. Но 5 делится только на 5 и на 1. Поскольку у нас заведомо есть класс с элементами 4 и 5, т. е. в каждом классе не менее 2-х элементов, получаем, что существует только один транзитивный класс, включающий все корни многочлена P .

Тогда по определению транзитивного класса получаем, что в группе G есть автоморфизмы, соответствующие всем транспозициям на 5 элементах. Но тогда группа G содержит вообще все перестановки из группы S_5 , т. к. любая перестановка получается как произведение транспозиций (см. стр. 212).

Итак, $G(\mathbb{Q}[x_1, x_2, x_3, x_4, x_5]/\mathbb{Q}) \cong S_5$. На самом деле, справедливо и более общее утверждение, доказываемое аналогичным образом: если многочлен P простой степени p неприводим, то группа Галуа, порожденная всеми его корнями, изоморфна S_p .

Вернемся к нашей исходной задаче о разрешимости уравнения 5-ой степени в радикалах. Дальнейшая логика рассуждений сводится к следующим трем шагам:

Step1 Если хотя бы один корень уравнения $x^5 - 6x + 3$ можно получить операциями сложение, вычитание, умножение, деление, возведение в целую степень и взятие корня целой степени, то и все остальные корни таковы же, и все они лежат в некотором нормальном расширении $E \supset \mathbb{Q}$, причем имеет место башня полей:

$$\mathbb{Q} = K_0 \subset K_1 \subset K_2 \subset \dots \subset K_r = E,$$

где для каждого $i = \overline{0, r-1}$ имеют место расширения либо $K_{i+1} = K_i[\alpha_i]$, либо $K_{i+1} = K_i[\sqrt[p_i]{1}]$, где $\alpha_i^{p_i} \in K_i$ при некотором простом p_i .

То есть каждый этаж такой башни получается либо сразу присоединением всех комплексных корней p_i -ой степени некоторого числа α_i из предыдущего поля (если корни из 1 уже включены в K_i), либо присоединением всех комплексных корней p_i степени из 1, а корни α_i присоединяются уже на следующем шаге. Такие башни называются *башнями радикальных расширений* \mathbb{Q} , и все числа, получаемые указанными операциями из целых чисел, лежат в этих расширениях.

Заметим, что вовсе не обязательно $E = \mathbb{Q}[x_1, \dots, x_5]$, достаточно требовать вложения $\mathbb{Q}[x_1, \dots, x_5] \subset E$.

Step2 Все группы $G_i = \mathbf{G}(K_i/\mathbb{Q})$ являются группами Галуа такого же порядка, какова степень расширения $[K_i : \mathbb{Q}]$, и эти группы образуют субнормальный ряд с абелевыми факторами:

$$\{e\} \triangleleft G_1 \triangleleft G_2 \triangleleft \dots \triangleleft G_r,$$

т.е. G_i нормальна в G_{i+1} и фактор G_{i+1}/G_i является абелевой группой, $i = \overline{0, r-1}$. Для доказательства нормальности $G_i \triangleleft G_{i+1}$ достаточно взять произвольный автоморфизм $h \in G_i$, т.е. \mathbb{Q} -автоморфизм поля K_i , и произвольный автоморфизм $g \in G_{i+1}$, т.е. \mathbb{Q} -автоморфизм поля K_{i+1} , и проверить, что $g^{-1}hg \in G_i$. Это почти очевидно, поскольку $g|_{K_i} : K_i \hookrightarrow \mathbb{A}$ и расширение K_i нормально над \mathbb{Q} , т.е. по свойству нормальности (N2) $g|_{K_i} \in G_i$. Таким образом, $G_i \triangleleft G_{i+1}$.

То, что фактор G_{i+1}/G_i коммутативен, следует из того, что он изоморфен группе Галуа расширения $K_i \subset K_{i+1}$, поскольку автоморфизмы данной группы представляют собой в точности «остатки» автоморфизмов группы G_{i+1} по модулю группы G_i (факторизация уравнивает автоморфизмы K_{i+1} , отличающиеся внутри K_i и совпадающие снаружи). А группа $\mathbf{G}(K_{i+1}/K_i)$ изоморфна циклической группе \mathbb{Z}_{p_i} по построению (либо это присоединенные корни из 1, дающие цикличность по модулю p_i , либо это степени α_i^k , также дающие цикличность по модулю p_i).

Но тогда группа G_r является разрешимой!

Step3 Для уравнения $x^5 - 6x + 3$ группа G_r изоморфна S_5 , которая не является разрешимой. Противоречие.

Итак, никакой корень уравнения $x^5 - 6x + 3$ не может быть записан в виде конечной формулы в радикалах, отправляясь от рациональных (целых) чисел, подобно тому, как это делается для любого уравнения низших степеней по формулам Кардано и Феррари.

Следовательно, не существует общей формулы для решения уравнений 5-ой степени в радикалах, а значит, не существует общей формулы и для

уравнений всех более высоких степеней. Это есть основной результат классической теории Галуа.

Основная теорема Галуа у нас завуалирована между шагами Step1 и Step2, поскольку мы изучали один конкретный пример и не рассматривали эту теорему как самоцель наших мытарств. Эта теорема в более общем виде устанавливает жесткое взаимно однозначное соответствие между промежуточными полями нормальной башни полей и нормальными подгруппами группы Галуа верхнего поля башни (см., например, Miles Reid [75]).

Собственно, именно из этого соответствия и родилось понятие нормальной подгруппы, да и сам термин «нормальная подгруппа» взят именно из определения нормального расширения полей. Вот так, задом наперед, мы вновь возвращаемся к теории групп и видим, насколько плотно она вживается во все, что связано с числами.

Поэтому, в довершение глав о числах, будет уместным добавить здесь существующую на сегодняшний день классификацию групп.

3.8 Группы: завершающий аккорд

На протяжении нашего увлекательного путешествия в мире математики мы встретили огромное количество групп. Как видим, это изобретение, придуманное и освоенное в XIX веке Галуа, Кэли, Коши, Кронекером и многократно усиленное в XX веке работами Фробениуса и Бернсайда, Брауэра, Шура, Вейля, Картана, Шевалле и многих других, является мощным и удобным инструментом анализа (и даже языком описания) во многих областях математики. Наконец, во второй половине XX века было решено систематизировать «происхождение» групп, что привело к известной ныне классификации простых групп, в основном усилиями Горенштейна, Томпсона и Фейта.

Проект классификации групп превысил по сложности, объему и количеству вовлеченных ученых все предыдущие попытки полностью описать происхождение групп. Доказательство теоремы о классификации простых конечных групп насчитывает работы более 100 авторов, опубликованных в основном с 1955 по 2004 годы и содержащих в сумме тысячи страниц текста.

Текущие исследования направлены на упрощение классификации групп. Горенштейн, Соломон и Лайонс постепенно публикуют упрощенную и пересмотренную версию доказательства. К 2018 году опубликовано 7 томов [68–74] нового доказательства (классификация второго поколения) и предполагается выпуск еще 5 томов. Соломон [76] оценивает итоговое доказательство в 5000 страниц.

Интересно также обратиться к Атласу представлений конечных групп [51]. Кроме того, рекомендуем посмотреть видеолекцию С. Галкина «Происхождение групп» [1qoMfgeEMtw].

Наконец, стоит отметить, что до сих пор в теории групп не сложились устойчивые общепринятые обозначения. Например, ортогональные группы D_n обозначаются также $PSO(n, q)$, а группа $PSL(n, q)$ обычно не то же самое, что $PSL(n, F_q)$. Поэтому при изучении вопросов классификации групп стоит всегда помнить о дублирующих обозначениях.

Теорема 3.23 (Классификация простых конечных групп). *Любая конечная простая группа изоморфна одной из групп из следующих семейств:*

GF1 циклические группы \mathbb{Z}_p простого порядка;

GF2 знакопеременные группы A_n перестановок не менее 5 элементов;

GF3 группы Шевалле $PSL(n, F_q)$, $PSU(n, F_q)$, $PSp(n, F_q)$, $PSO(n, F_q)$;

GF4 исключительные и скрученные формы групп типа Ли (включая группу Титса);

GF5 26 спорадических групп.

Полный список 26 спорадических групп:

- Группы Матьё M_{11} , M_{12} , M_{22} , M_{23} , M_{24} ;
- Группы Янко J_1 , J_2 , J_3 , J_4 ;
- Группы Конвея Co_1 , Co_2 , Co_3 ;
- Группы Фишера Fi_{22} , Fi_{23} , Fi_{24} ;
- Группа Хигмана–Симса HS ;
- Группа МакЛафлина McL ;
- Группа Хельда He ;
- Группа Рюдвалиса Ru ;
- Группа Судзуки Suz ;
- Группа О'Нана $O'N$;
- Группа Харады–Нортонна HN ;
- Группа Лайонса Ly ;
- Группа Томпсона Th ;
- Группа «малый Монстр» B ;

- Группа «Монстр» Фишера–Грейса M .

20 из этих групп являются подгруппами (подфактор-группами) группы «Монстр» и образуют так называемое **счастливое семейство**. Остальные 6 групп ($J_1, J_2, J_3, O'N, Ru, Ly$) называются **париями**.

Порядок группы «Монстр» равен

$$24^6 \cdot 3^{20} \cdot 5^9 \cdot 7^6 \cdot 11^2 \cdot 13^3 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 41 \cdot 47 \cdot 59 \cdot 71 \approx 8 \cdot 10^{53}.$$

3.9 Числовые архетипы

В главах 2 и 3 мы рассмотрели огромное количество (число?) всевозможных числовых (и не совсем числовых) конструкций, которые не перестают удивлять своими свойствами не только математиков. Стоит, однако, отметить одну особенность, присущую всем определениям и свойствам чисел. Даже когда мы имеем дело с очень странными объектами вроде гипергеометрических или сюрреальных чисел, мы всегда стараемся элементы их структуры снабдить такими ярлыками, которые, будучи поставлены в одном предложении или формуле в определенном порядке, давали бы нам если не привычное арифметическое правило, то хотя бы его близкое подобие.

Действительно, мы называем нулем нейтральный элемент операции сложения, мы стараемся сложение и умножение связать по формуле $a + a + \dots + a = a \cdot n$, а степень и умножение — по формуле $a \cdot a \dots a = a^n$, причем самым тщательным образом выясняем, где и как именно эти операции дают сбой (в сравнении с обычной арифметикой), чтобы не запутаться и не наделать ошибок, действуя по арифметической интуиции.

При построении других геометрий, несмотря на высказывание Гильберта о «столах, стульях и пивных кружках» [7], мы все же вводим точки, прямые и плоскости, а потом ищем сходства и различия с их евклидовыми аналогами.

При построении альтернативных арифметик и анализа мы строим многочлены, ищем основную теорему алгебры, учимся раскладывать функции в ряды (снова суммы и степени!), придумываем наиболее удачное определение экспоненты, вводим классы «хороших» функций и изучаем их алгебраические и топологические свойства.

При рассмотрении других физик (например, в двойных числах — см. статью С. Кокарева и Д. Павлова в [85]) мы отыскиваем привычные по поведению дифференциальные операторы, уравнения для энергии и фундаментальные константы.

Возникает даже подозрение о том, что можно придумать некий универсальный математико-числовой язык, который смог бы описать все возможные числовые конструкции (не только алгебраические, но и функционально-топологические) в едином понятийном аппарате подобно тому, как язык мат-

логики единообразно описывает все математические доказательства, а язык теории множеств погружает в себя все методы создания математических объектов. Подозрение это подкрепляется и развитием собственно алгебры (включая ее ветви вроде алгебраической топологии), и возникновением теории типов и родов структур, и такой дисциплиной, как теория категорий. Возможна ли теория всего? — спросим мы математиков, как спрашивают друг друга физики. А может быть, эта теория всего будет общей для математики и физики?

Как бы то ни было, картина, которую мы нарисовали в разделе 1.1.3, пополнилась весьма детальными штрихами. Теперь мы видим не просто поле записей с функциями и рекурсиями, а вполне конкретные образы алгебраических структур: здесь и башни расширения полей, и деревья упорядоченных множеств, и вполне конкретные числовые поля и кольца с разнообразными надстройками в виде векторных пространств и модулей. Наконец, все это «хозяйство» удивительным образом находит свое место в универсумах множеств и на дереве сюрреальных чисел. А если немного приподняться над картиной и взглянуть за ее пределы, то мы увидим, как математика незаметно перерастает в физику и программирование, и далее уходит во все области человеческого знания. При этом в основе всего лежит арифметика! Читателям, которые дойдут до конца книги, предлагается примитивный скелет обрисованной картины в виде связанных друг с другом математических понятий и дисциплин, на который каждый может наращивать «мясо» в меру своих познаний и интересов.

Но вернемся от грёз к приведенным выше примерам. Говорят ли они об инерционности человеческого мышления, о том, что в мире хаоса наше сознание стремится расставить привычные флажки-термины и построить надежные дороги логики? Возможно. Говорит ли это об узости нашего мировоззрения и слабости методов познания? Вряд ли.

Скорее всего, при завоевании и освоении логико-математического пространства на протяжении почти трех тысяч лет человечество просто пользуется одними и теми же архетипами количественного миропонимания, заложенными, если хотите, в нашем подсознании почти так же явно, как в компьютер изначально заложена двоичная система счисления. Числовые архетипы, позволяющие познание превращать в знание путем вычислений (или *исчислений*) — это и есть краеугольный камень техногенной цивилизации.

Наконец, для того, чтобы зафиксировать некоторый целевой результат книги в данной главе, постараемся перечислить(!) **некоторые числовые архетипы**:

1. Ноль как точка отсчета.
2. Сложение \rightarrow Умножение \rightarrow Степень $\rightarrow \dots$ Башни (степеней и полей).
3. Числа над числами (матрицы, полиномы, функциональные простран-

ства и т.д.).

4. Порождение «новых» чисел из проблем со «старыми» числами.
5. Стремление к частичному или линейному упорядочению чисел, а также перенос порядка с чисел на другие объекты (например, мощность множеств, нумерация универсумов)

Здесь пункт 4 — это уже знакомый нам по первой главе архетип образующего элемента. Порождение новых чисел с помощью добавления нового порождающего числа, существование которого постулируется из алгебраической потребности (проще говоря, не хватает какого-то корня уравнения при том, что само уравнение можно записать, не выходя за рамки старой системы чисел).

При этом мы не сбрасываем со счетов изначальный архетип числа, о котором говорилось в начале второй главы, и те фундаментальные роли, которые он играет в нашем сознании. Перечисленные числовые архетипы лишь детализируют наше представление о первородном архетипе числа как такового.

Кроме того, в процессе работы с числами мы обнаружили еще несколько системообразующих архетипов (не числовых):

1. Архетип возможности неограниченного расширения числовых систем (149).
2. Архетип безграничной делимости чисел с помощью трансфинитной рекурсии (149).
3. Методологический архетип трансцендентного восприятия (191).
4. Архетип переноса свойств структуры на базовое множество (197).
5. Архетип инварианта (213).
6. Методологический архетип редукции (227).
7. Архетип вариативности представления математических объектов (237).
8. Архетип двойственности (290).

Итак, рассмотрев достаточно подробно понятие Числа, перейдем к тому, что, как мы упомянули выше, помогает познание превратить в знание, т.е. перейти от открытия или, если угодно, прозрения, к извлечению и накоплению знаний. Выражаясь использованным выше художественным языком, рассмотрим теперь нашу картину в движении.

*Да простят нас
пифагорейцы*

,B,

Матлогика. Исчисления I

4.1 Исчисление высказываний и предикатов

Здесь мы коротко пройдемся по терминологии и начальным сведениям из математической логики. Для более развернутого изучения данной тематики мы рекомендуем обратиться к книгам [12, 15, 23].

Теоремы этого раздела мы будем называть метатеоремами, чтобы отличать их от обычных теорем математики. Различие между теоремами и метатеоремами будет понятно из дальнейшего.

4.1.1 Исчисление высказываний

Высказыванием принято называть любое предложение, которое можно квалифицировать либо как ложное, либо как истинное. Например «Снег — белый» обычно является истинным высказыванием, «3 — четное число» есть ложное высказывание, «Совершенных чисел бесконечно много» — неизвестно, истинно такое высказывание или ложно (на 2019 год).

Выражение « x — простое число» не является высказыванием до тех пор, пока не определено конкретное x . Иначе говоря, высказывания не могут иметь параметров, в зависимости от значений которых может меняться их истинность. В то же время «все простые числа, которые больше 2, нечетные» — истинное высказывание. Оно не имеет параметров, хотя и утверждает нечто о многих числах.

Высказывания можно соединять логическими связками и получать новые высказывания. Обычно используются следующие **логические связки**: \wedge — логическое И (AND), \vee — логическое ИЛИ (OR), \neg — логическое отрицание (NOT). Связки порождают булевы функции от *пропозициональных переменных*, могущих принимать только истинное или ложное значение: $a \wedge b$, $a \vee b$, $\neg a$. Кроме перечисленных трех функций существуют: $a \rightarrow b$, $a \leftrightarrow b$, $a | b$ (NAND), $a \downarrow b$ (NOR), $a \oplus b$ (XOR). Такие функции задаются таблицами истинности:

a	b	$a \wedge b$	$a \vee b$	$\neg a$	$a \rightarrow b$	$a \leftrightarrow b$	$a b$	$a \downarrow b$	$a \oplus b$
0	0	0	0	1	1	1	1	1	0
1	0	0	1	0	0	0	1	0	1
0	1	0	1	1	1	0	1	0	1
1	1	1	1	0	1	1	0	0	0

Упражнение
4.1.
Сколько
всего
функций от
2 логических
переменных?

Здесь единица означает Истину, ноль — Ложь.

Некоторые наборы булевых функций образуют базис (**полную систему**) в пространстве всех булевых функций в том смысле,

что все остальные функции можно выразить через них. Например, $(a \rightarrow b) = (\neg a \vee b)$ и т.д. Существует необходимое и достаточное условие (критерий Поста) того, чтобы набор функций был базисом: он не должен целиком лежать ни в одном из пяти классов булевых функций: сохраняющие 0, сохраняющие 1, монотонные, линейные, самодвойственные.¹

Примеры базисов: стрелка Шеффера $a|b$ (NAND), $\{\neg, \wedge\}$, $\{\neg, \vee\}$. Как видим, в данном случае, в отличие от линейных пространств, размерность базиса может быть различной.

Поскольку все булевы функции можно выразить через базисные, т.е. все они являются выразимыми, каждой булевой функции соответствует формула, составленная из атомарных формул (т.е. формул базисных функций) и символов-связок. И обратно, каждой формуле соответствует функция. Поэтому в данном случае де факто нет различия между формулами и функциями, и в дальнейшем мы будем работать именно с формулами.

Тождественно истинные формулы называются **тавтологиями**. Например, $a \vee \neg a$ является тавтологией, что легко проверить с помощью таблиц истинности. Если формула допускает истинное значение хотя бы при одном наборе значений аргументов, то такая формула называется **выполнимой**. Формула выполнима тогда и только тогда, когда ее отрицание не является тавтологией.

Поскольку высказывания — термин из логики, логично говорить о том, как из одних высказываний можно вывести другие, пользуясь только истинностными умозаключениями. Например, если нам известно, что из a следует b (если x есть Сократ, то x есть человек), а из b следует c (если x есть человек, то x смертен), то мы считаем, что из a следует c (если x есть Сократ, то x смертен). Это называется силлогизмом.

Существует формальная схема, с помощью которой можно строить подобные высказывания или, как говорят, строить выводы. Это — схемы аксиом исчисления высказываний (далее — **ИВ**):

$$A1 \quad a \rightarrow (b \rightarrow a);$$

$$A2 \quad (a \rightarrow (b \rightarrow c)) \rightarrow ((a \rightarrow b) \rightarrow (a \rightarrow c));$$

¹Сохраняет 1, если $f(1, 1) = 1$; линейность: в алгебре операций $\{+, \wedge\}$; самодвойственность: $f(\neg a, \neg b) = \neg f(a, b)$.

A3 $(a \wedge b) \rightarrow a$;

A4 $(a \wedge b) \rightarrow b$;

A5 $a \rightarrow (b \rightarrow (a \wedge b))$;

A6 $a \rightarrow (a \vee b)$;

A7 $b \rightarrow (a \vee b)$;

A8 $(a \rightarrow c) \rightarrow ((b \rightarrow c) \rightarrow (a \vee b \rightarrow c))$;

A9 $\neg a \rightarrow (a \rightarrow b)$;

A10 $(a \rightarrow b) \rightarrow ((a \rightarrow \neg b) \rightarrow \neg a)$;

A11 $a \vee \neg a$;

и одно правило вывода (*modus ponens*, МР), позволяющее переходить от посылок к выводу:

$$\frac{a, a \rightarrow b}{b}$$

Выводы, построенные с помощью аксиом ИВ и правила МР, называются **теоремами ИВ**.

Отметим, что перечисленные аксиомы являются именно схемами аксиом, т. е. вместо любой переменной a, b, c в них может быть подставлена любая корректная формула ИВ (в том числе, формула любой из аксиом). То же самое относится к правилу вывода МР. Более правильно (с точки зрения грамматики) было бы ввести специальные переменные, обозначающие формулы ИВ, и сформулировать эти аксиомы в этих переменных. Но фишка в том, что в исчислении высказываний предметной областью являются сами высказывания (точнее, всего два символа, отвечающих за них — истина и ложь), поэтому переменные, пробегающие предметную область, одновременно могут олицетворять и формулы (их значения) без введения дополнительных переменных.

Нетрудно удостовериться (из таблиц истинности) в том, что все аксиомы являются тавтологиями. Более того, если правило вывода рассматривать как импликацию, то, имея тавтологию a в егосылке и установленную импликацию $(a \rightarrow b)$, мы имеем тавтологию b и в ее выводе (опять же — из таблиц истинности). Отсюда, в общем-то, очевидно следует, что всякая теорема ИВ является тавтологией.

Возникает вопрос: *верно ли обратное?* Всякая ли тавтология может быть выведена из аксиом ИВ. Ответ: да. И это свойство ИВ называется **полнотой**.

Метатеорема 4.1 (о полноте ИВ). *Всякая тавтология есть теорема исчисления высказываний.*

Строго говоря, необходимо также оговорить, что переменные в аксиомах могут заменяться любыми другими переменными и формулами (правило подстановки) так, чтобы вместо всех вхождений одной переменной подставлялась одна и та же переменная или формула.

Помимо аксиом ИВ в качестве отправной точки вывода может служить любой список формул Γ . Если из аксиом ИВ и формул Γ выводится формула a , то мы пишем $\Gamma \vdash a$. В случае пустого списка Γ пишем просто $\vdash a$. В силу полноты ИВ выводимость формулы a из формул Γ означает, что $\bigwedge \Gamma \rightarrow a$ является тавтологией. Здесь под $\bigwedge \Gamma$ мы понимаем список Γ , в котором все запятые заменены на знак \wedge , т.е. список формул из Γ превращен в одну большую конъюнкцию.

Про список формул Γ можно утверждать следующее:

1. Γ **противоречив**, если из него можно одновременно вывести a и $\neg a$ (если нельзя, то **непротиворечив**);
2. Γ **совместен**, если $\bigwedge \Gamma$ выполнима, т.е. при некоторых значениях переменных формул из Γ конъюнкция $\bigwedge \Gamma$ истинна (если это не так, то **несовместен**).

Понятно, что всякий совместный список Γ непротиворечив. Действительно, если это не так, то $\Gamma \vdash a$ и $\Gamma \vdash \neg a$, т.е. $(\bigwedge \Gamma \rightarrow a) \wedge (\bigwedge \Gamma \rightarrow \neg a)$ истинно, но данная формула равна (в силу таблиц истинности) формуле $\neg(\bigwedge \Gamma)$, откуда следует, что $\bigwedge \Gamma$ есть тождественная ложь, а значит, $\bigwedge \Gamma$ невыполнима, и список Γ несовместен.

Упражнение 4.2. | Данное рассуждение есть иллюстрация правила контрапозиции: если верно $\neg b \rightarrow \neg a$, то верно $a \rightarrow b$ (проверьте это через таблицы истинности).

Верно и обратное.

Метатеорема 4.2 (полнота ИВ, вторая форма). *Если список Γ непротиворечив, то он совместен.*

Выше мы предполагали, что Γ есть список формул, т.е. конечный набор формул, из которых можно построить конъюнкцию, которая также будет формулой. На самом деле, все рассуждения обобщаются на случай *ага, как в базисе Гамеля :)* бесконечного набора формул (поскольку в выводе участвует все равно только конечный набор формул. Более того, существует

Метатеорема 4.3 (компактности для ИВ). *Пусть Γ — набор формул, всякое конечное подмножество которого совместно. Тогда Γ совместен.*

Интересно, что доказательство этой теоремы для случая счетного Γ можно вывести из топологической теоремы о компактности канторовского пространства бинарных последовательностей $\{0, 1\}^\omega$. В этом пространстве точкой яв-

ляется последовательность $f : \omega \rightarrow \{0, 1\}^2$, а расстоянием между точками f и g является величина 2^{-n} , где n — номер расхождения f и g (т. е. до n эти функции совпадают, а в точке n принимают разные значения). Пространство последовательностей с такой метрикой является компактным.

Далее будем считать, что все переменные из счетного набора Γ пронумерованы натуральными числами, т. е. каждый элемент канторовского пространства — это конкретный набор значений переменных. Для любой формулы φ из набора Γ область ее истинности $V[\varphi] \subset \{0, 1\}^\omega$ есть открыто-замкнутое множество. Более того, каков бы ни был конечный набор формул $\varphi_1, \dots, \varphi_n \in \Gamma$, пересечение $V[\varphi_1] \cap \dots \cap V[\varphi_n]$ непусто, поскольку $\varphi_1 \wedge \dots \wedge \varphi_n$ выполнима. Следовательно, система множеств $\mathcal{U} \equiv \{V[\varphi] \mid \varphi \in \Gamma\}$ образует центрированную систему замкнутых множеств в компактном пространстве $\{0, 1\}^\omega$. В топологии есть теорема, которая утверждает, что тогда пересечение всей системы \mathcal{U} непусто. Это и означает совместность Γ .

Для более чем счетного набора формул Γ можно использовать уже теорему Тихонова о компактности $\{0, 1\}^\tau$ с определенной подходящей топологией.

Конечно, подобные рассуждения — кошмарный сон для тех, кто любит формальную логику. Например, вставка формулы φ в определение множества $V[\varphi]$ сразу уводит нас из языка теории множеств. Более того, для доказательства этой метатеоремы мы привлекаем аппарат теории множеств, считая ее тем самым еще более «высокопоставленной» теорией, что наводит на определенные мысли о порочном круге.

И если первую проблему довольно легко разрешить, вспомнив о том, что формулы ИВ мы можем просто подменить функциями вида $f : \{0, 1\}^k \rightarrow \{0, 1\}$, тем самым избавившись от формульной переменной в обозначениях множеств, то вторая проблема уводит нас чуть ли не в философскую плоскость размышлений.

Действительно, мы анализируем исчисление высказываний, которое изначально нами привлекалось как метаинструмент для изучения теории множеств (ведь теория множеств — это система аксиом ZF плюс аксиомы ИВ и правило вывода МР, иначе просто никакой теории не получится). И анализируем мы ИВ не как-нибудь, а именно в рамках самой теории множеств с помощью ее теорем и методов (для простоты считаем, что топология есть часть теории множеств). Тем самым мы выворачиваем мир логики и теории множеств наизнанку.

Мы можем наблюдать здесь следующую закономерность получения знаний. Сначала мы вводим «начальные» множества, которые по сути являются закорючками на бумаге и носят вполне себе финитный характер. Эти множества помогают нам строить простой вариант исчисления высказываний и получать о нем некоторые результаты. Далее, имея уверенность в корректно-

²Вспомним, что само \mathbb{R} можно определить с помощью таких же последовательностей!

сти и полноте ИВ, мы вводим аксиоматическую теорию множеств с конечным набором аксиом (точнее с набором аксиом и одной схемой аксиом, которая вполне описывается финитными методами), а также исчисление предикатов. Затем, имея на руках уже развитый арсенал средств теории множеств с ее понятиями бесконечного и моделями многих полезных теорий (групп, анализа, топологии и т.п.), мы пускаемся во все тяжкие, и доводим само ИВ до уровня теории с бесконечной моделью. И так далее.

Похожим образом строились сюрреальные числа, не правда ли?

В итоге исследования в области теории множеств и логики напоминают карабкание вверх по стенкам колодца, когда нужно попеременно поднимать то левую, то правую ногу, опираясь остальными конечностями на то, что уже достигнуто.

*Архетип
глубокого
колодца?*

Мы не будем здесь придумывать какой-либо архетип, поскольку это рассуждение выходит за рамки математики и больше относится к философии науки. Но отметим, что какого-то явного противоречия в таком подходе не наблюдается.

Описанное выше исчисление высказываний называется ИВ гильбертовского типа (по имени главного исследователя). Отметим, что существуют и другие типы ИВ, например, генценовское исчисление секвенций (оно используется для анализа синтаксической структуры вывода и изучается в теории доказательств) и интуиционистское исчисление высказываний, в котором исключен закон исключенного третьего: $a \vee \neg a$.

Подробнее об этих типах ИВ см. книгу [12].

4.1.2 Исчисление предикатов

При изучении высказываний мы не имели дела с какой-либо предметной областью, точнее, нашей предметной областью были логические константы И и Л. Даже если мы говорили, что «снег белый», нам было безразлично, что именно под этим понимается, важно только, что это высказывание истинное или ложное.

Если же мы ставим высказывание в зависимость от переменных, пробегающих некоторую предметную область (вообще говоря, отличную от множества $\{0, 1\}$), то мы имеем дело с **предикатами**. Формально говоря, предикаты — это истинностные функции, заданные на степени какого-то множества, симметризирующего предметную область исследования.

Наличие параметров у предикатов позволяет их «замыкать» кванторами (обычно их два: всеобщности и существования), превращая в высказывание. Так, выражение « $2x$ есть четное число» является предикатом с параметром x и зависит от области изменения переменной x . А выражение «для любого натурального x : $2x$ есть четное число» является истинным высказыванием. Здесь мы замкнули исходный предикат по переменной x , поставив квантор

всеобщности («для любого»), и поскольку других переменных в нем нет, мы получили предикат без параметров, т. е. высказывание.

Формулы исчисления высказываний можно рассматривать как предикаты от двоичных переменных, принимающих значения 0 или 1.

Пусть теперь у нас имеется некоторая алгебраическая **сигнатура**, т. е. набор значков для предикатов (отношений) и операций с указанием их арности (валентности, местности):

$$\sigma = \langle R, F \rangle,$$

где элементы R представляют собой обозначения предикатов (отношений), т. е. имеют вид $R(_, \dots, _)$, где количество мест для предметных переменных может быть любым натуральным числом, в том числе нулем. Значения отношений — всегда булевские, т. е. либо 0 (ложь), либо 1 (истина). Отношения нулевой арности — это логические константы 0 и 1. Элементы F — это операции на предметной области, которые ставят в соответствие набору элементов еще какой-то элемент этой же области: $F(_, \dots, _)$. Операции нулевой арности — это константы, их столько, сколько элементов в предметной области.

Элементы R принято называть предикатными символами, а элементы F — функциональными символами. В дальнейшем мы будем пользоваться терминологией, приближенной к алгебраической, т. е. отношениями и операторами, иногда переходя к указанным синонимам.

Например, в R могут быть включены значки \in , $=$, обозначающие бинарные отношения, используемые в теории множеств, и ни одного операторного символа (в нашей теории «начальных» множеств были операторные символы — это фигурные скобки, позволявшие собирать множество из списка множеств). Если мы говорим о теории групп, то там в сигнатуру может быть включен только один бинарный операторный символ для групповой операции.

Здесь мы неявно сталкиваемся с тем, что в программировании называется типизацией. И по-хорошему, следовало бы писать $R(Object, \dots, Object) = Logic$, где *Object* означает, что сюда можно подставлять только термы, пробегающие предметную область, а значение всегда будет строго логическим. Впрочем, некоторые современные языки программирования пытаются самостоятельно интерпретировать подставляемое значение как объект нужного типа. Например, 123 можно понимать и как строку, и как число, так что если это значение будет подставлено в аргумент типа *Int*, то оно будет проинтерпретировано как число (скажем, если вы пишете $a+b$ и вместо a и b подставляете 123 и 456), а если аргумент должен быть строкой (или получается таковым из контекста выполнения программы), то 123 будет интерпретировано как строка "123" (например, $123 + \text{«Hello»}$ выдаст «123Hello», поскольку второй аргумент оказался текстовым, и оператор $+$ был воспринят как оператор конкатенации вместо арифметического сложения). Это очень удобно,

но об этом нужно помнить при написании кода и его отладке.

В нашем случае разброс типов невелик (либо объектный, либо логический), поэтому нет нужды акцентировать внимание на типизации символов.

Тем не менее, поскольку именно у операторных символов тип аргументов и значений один и тот же (объекты предметной области), мы можем как угодно их комбинировать друг с другом и ввести понятие термина.

Итак, пусть у нас *терминальный алфавит* включает все символы из R и F , а также скобки, запятые, латинские буквы и символы логических связок. *Нетерминальный алфавит* включает следующие понятия: Список[k], Терм, Константа, Формула. Здесь число k у списка указывает на его длину. Строго говоря, для каждого натурального k мы создаем свой набор списков, и все такие списки включают ровно k элементов.

Сформулируем правила грамматики для языка в сигнатуре σ :

GL1 Терм \rightarrow <латинская буква>

GL2 Список[1] \rightarrow Терм

GL3 Список[$k + 1$] \rightarrow Список[k], Терм

GL4 Константа \rightarrow <Операторный символ из F арности 0>

GL5 Терм \rightarrow <Операторный символ из F арности k >(Список[k])

GL6 Формула \rightarrow <Предикатный символ из R арности k >(Список[k])

GL7 Формула $\rightarrow \neg$ Формула | Формула \wedge Формула | Формула \vee Формула |
Формула \rightarrow Формула

GL8 Формула $\rightarrow \forall$ <латинская буква> Формула | \exists <латинская буква> Формула

Здесь в угловых скобках указано название перечня альтернатив, который следует иметь виду при формировании структур языка.³ Например, терминами будут буквы a, b, c , из которых можно составить список a, b, c , затем взять операторный тернарный символ F и получить новый терм $F(a, b, c)$, после чего сформировать список $a, F(a, b, c)$ и при помощи бинарного предикатного символа R получить формулу $R(a, F(a, b, c))$, после чего можно добавить квантор: $\forall aR(a, F(a, b, c))$. И так далее.

Определение. Совокупность всех грамматических конструкций, полученных по правилам GL1–GL8, называется **языком**, порожденным сигнатурой

³Такой список не всегда может быть конечным, поэтому в дереве такой грамматики могут быть точки бесконечного ветвления.

σ , и обозначается $\mathcal{L}(\sigma)$. Такой язык является языком *первого порядка*, поскольку не включает в предметную область формулы (нельзя написать $\forall\varphi$, где φ является формульной переменной).

Отметим, что если перечень операторных символов пуст, а перечень предикатных символов состоит из единственного одноместного предиката без обозначения (т.е. он просто совпадает с тем объектом, который подставляется на его аргументное место), то мы получаем язык исчисления высказываний в чистом виде — только формулы и логические переменные. При этом у нас предметная область совпадает с множеством $\{0, 1\}$, т.е. с областью логических значений.

Более того, если подставлять все возможные термы языка $\mathcal{L}(\sigma)$ в предикатные символы, а полученные формулы — вместо переменных в формулах исчисления высказываний, то мы получим тот же самый язык $\mathcal{L}(\sigma)$.

Предикатами, или *атомарными формулами* (над сигнатурой σ) мы будем называть все, что получается по правилу GL6, т.е. предикаты — это формулы, не содержащие логических связок. Таким образом, язык $\mathcal{L}(\sigma)$ получается как композиция формул ИВ с предикатами над сигнатурой σ . Отсюда, в частности, следует, что все тавтологии ИВ порождают (путем подстановки вместо булевых переменных каких-либо предикатов или формул, в том числе содержащих кванторы) истинные формулы языка $\mathcal{L}(\sigma)$ независимо от значений параметров предикатов. Например, тавтология $a \vee \neg a$ порождает истинные формулы вида $\forall x R(x, y, z) \vee \neg(\forall x R(x, y, z))$, где x, y, z — предметные переменные или термы (возможно, тоже с параметрами).

Исчисление предикатов — это изучение языков первого порядка с произвольной сигнатурой, которое при этом включает аксиоматику ИВ (каждая аксиома ИВ рассматривается как схема аксиом, где вместо булевых переменных подставляются произвольные формулы языка $\mathcal{L}(\sigma)$) и **Modus Ponens**, а также новые аксиомы и правила, связанные с появлением кванторов:

$$\frac{\varphi, \varphi \rightarrow \psi}{\psi} \quad (\text{modus ponens});$$

$$\text{A12} \quad \forall x \varphi \rightarrow \varphi(x||t) \quad (\text{аксиома подстановки для } \forall);$$

$$\text{A13} \quad \varphi(x||t) \rightarrow \exists x \varphi \quad (\text{аксиома подстановки для } \exists);$$

$$\frac{\psi \rightarrow \varphi}{\psi \rightarrow \forall x \varphi} \quad (\text{правило Бернайса, введение } \forall);$$

$$\frac{\varphi \rightarrow \psi}{\exists x \varphi \rightarrow \psi} \quad (\text{правило Бернайса, удаление } \exists),$$

где $\varphi(x||t)$ — формула, полученная в результате подстановки терма t вместо каждого свободного вхождения переменной x в формулу φ (у самой формулы φ может быть несколько свободных переменных, может вообще не быть

свободных переменных, а также среди свободных переменных может не быть x , в этом случае подстановка тривиальна — она ничего не меняет), а формула ψ не содержит свободно переменной x .⁴

Комментарий 12.

В качестве еще одного важного штриха к портрету логических исчислений нужно отметить, что Гёдель и Лёб в свое время ввели еще одну одноместную логическую связку «бокс»: $\Box\varphi$, которая означает выводимость, или доказуемость формулы φ . Этим они как бы инкапсулировали понятие выводимости непосредственно в язык логики. Если о выводимости в языке мы рассуждаем, находясь в метатеории (например, ZF), то логическая связка «бокс» встроена уже в сам язык ИП. При этом было введено две системы аксиом, связывающих эту новую связку и остальное ИП, и полученные системы называли GL (не путать с группой GL!) и GLS (а это — с «мерседесом»). Эти две расширенные системы по сути определили вектор развития науки, называемой теорией доказательств.

Известная **теорема Лёба**: $\Box(\Box\varphi \rightarrow \varphi) \rightarrow \Box\varphi$. Следствием теоремы Лёба является то, что только в противоречивой теории высказывание «доказуемость φ влечёт φ » доказуемо для всех утверждений φ .

Заметим, что кванторы всеобщности и существования являются парными относительно отрицания и ведут себя аналогично конъюнкции и дизъюнкции. Формулу $\forall x \varphi(x)$ можно интерпретировать как конъюнкцию $\bigwedge_x \varphi(x)$ по всем значениям переменной x , а формулу $\exists x \varphi(x)$ — как дизъюнкцию $\bigvee_x \varphi(x)$. Отсюда легко усмотреть и правила двойственности, аналогичные законам де Моргана:

$$\neg(\forall x \varphi(x)) \leftrightarrow \exists x \neg\varphi(x), \quad \neg(\exists x \varphi(x)) \leftrightarrow \forall x \neg\varphi(x)$$

Формулы, в которых все переменные связаны кванторами, т. е. не содержащие свободных переменных, называются **замкнутыми**. Замкнутые формулы можно считать высказываниями, т. к. они имеют либо только истинное, либо только ложное значение. Отличие замкнутой формулы от высказывания только в том, что нам известна формальная внутренняя структура этого высказывания.

Это
«только»
перевесит
любые другие
плюшки.

Из первого правила Бернайса следует правило обобщения:

$$\frac{\varphi}{\forall x \varphi},$$

⁴Мы не заостряем внимание на понятии свободного вхождения переменной, но следует иметь ввиду, что при машинной обработке формулы сначала нужно устранять коллизии переменных, заменяя все связанные переменные новыми, опираясь от нижнего уровня формулы. Например, в формуле $\varphi(x) \wedge \forall x\psi(x)$ следует произвести замену связанного x : $\varphi(x) \wedge \forall y\psi(y)$. Это ничего не меняет в логике формулы, но избавляет от коллизий переменных.

позволяющее вводить квантор всеобщности в ранее полученную формулу.

Так же, как в ИВ, в ИП выводом формулы φ называется конечная последовательность формул, в которой на каждом шаге переход определяется правилами вывода, в начале стоят аксиомы ИП, а в конце — формула φ . В этом случае мы пишем: $\vdash \varphi$.

Если вывод осуществляется из какого-то набора посылок Γ , то пишут $\Gamma \vdash \varphi$.

Соответственно, можно говорить о «чистом ИП», где теоремами являются только формулы сигнатуры σ , выводимые из аксиом ИП, и о **теории** с аксиоматикой \mathcal{A} , где теоремами являются все *замкнутые* формулы, выводимые из набора \mathcal{A} замкнутых формул или схем аксиом. Чем больше (непротиворечивых, независимых) аксиом, тем больше получается теорем. Совокупность теорем, выводимых из \mathcal{A} , называется **теорией** с аксиоматикой \mathcal{A} . Пользуясь архетипом базового множества, мы такую теорию часто будем называть просто теорией \mathcal{A} .

Здесь мы усматриваем аналогию с векторным пространством. Действительно, мы выбираем некоторый базис (по умолчанию включающий набор аксиом ИП) и рассматриваем множество всех формул, которые из него выводимы. Получается некое логическое («выводимое») пространство с базисом из этих аксиом (или схем аксиом). Например, это могут быть аксиомы ZF.

Далее мы можем добавить какую-либо аксиому (скажем, аксиому выбора) или ее отрицание, и получить еще некоторый ряд теорем, причем не только прямых следствий именно этой аксиомы, а все, что выводится из совокупности аксиом. Это пополнение аксиоматики напоминает пополнение базиса векторного пространства.

В этом логическом пространстве существуют эквивалентные базисы — равносильные системы аксиом. Безусловно, о постоянстве мощности базиса (как и в случае булевых функций) тут речи не идет.

Отсюда мы приходим к еще одному методологическому архетипу математики — **архетипу подходящего базиса**. Под этим термином мы понимаем следующее: при изучении свойств какого-либо пространственного объекта (будь то фигура в векторном пространстве или система аксиом и теорем, или формализм какой-то теории) мы стараемся найти наиболее выгодный для работы базис, в котором изучаемые объекты предстают либо в упрощенном виде (как, например, матрица оператора в базисе из собственных векторов), либо в уже знакомом нам виде (например, если мы от изоморфизмов переходим к равенству или от отношения принадлежности в ординалах переходим к отношению линейного порядка). Повернув изучаемое пространство должным образом, мы можем увидеть новые нетривиальные результаты, а затем совершить обратный поворот, и эти новые результаты переложить на старые обозначения. Иногда после этого мы обнаруживаем, что и в исходном языке получить такой результат было бы несложно. Сложно было его увидеть!

Заметим также, что существуют эквивалентные сигнатуры и эквивалентные формулы в пределах одной сигнатуры. Сигнатуры можно считать эквивалентными, если между ними установлено взаимно однозначное соответствие такое, что отношению арности k первой сигнатуры соответствует отношение той же арности k второй сигнатуры, и функциональному символу арности k первой сигнатуры соответствует функциональный символ той же арности k второй сигнатуры. Например, ничто не мешает нам в сигнатуре теории множеств заменить \in на $=$, и наоборот, $=$ на \in , правда, читабельность такой формализации упадет в разы.

Эквивалентные формулы в рамках одной и той же сигнатуры — это, прежде всего, формулы, в которых между наборами связанных переменных установлено взаимно однозначное соответствие. Например, формула $\forall x \varphi(x)$ и формула $\forall y \varphi(y)$ будут эквивалентны, если только замена $x||y$ не привела к коллизии переменных (т.е. если внутри φ переменная y не входит). Мы не будем строго разбирать понятие связанных и свободных переменных, отметим лишь, что данное определение дается индуктивно по сложности формулы и терма.

Ну, и кроме того, формулы могут быть эквивалентными просто в силу аксиом ИВ, например, $\psi_1 \rightarrow (\psi_2 \rightarrow \varphi)$ эквивалентна $(\psi_1 \wedge \psi_2) \rightarrow \varphi$.

Таким образом, у одной и той же (по сути) формулы имеется множество имен. Поэтому для формул уместна еще одна аналогия: записи множеств. Одно и то же множество может быть записано по-разному, но от вида конкретной записи его свойства не меняются. Поэтому на протяжении первой главы мы учились отделять объекты-множества от их записей. Аналогично можно считать, что формулы существуют сами по себе, а их эквивалентные воплощения — не более чем символическое представление одного и того же объекта в некоем логическом пространстве.

Комментарий 13.

Напоминает файл в Linux: он существует сам по себе, но при этом имеет равноправные имена-ссылки в каталоге. Файл считается удаленным, когда у него нет ни одного имени (как человек без документов). В этом случае адресное пространство за ним не резервируется и может быть занято другими файлами.

Вместо обычной арифметики в пространстве выводов работает логическая арифметика — исчисление предикатов. И, как это часто бывает, мы видим архетип порождающего элемента (аксиоматика) и генерацию логического пространства на основе правил вывода.

На рис. 4.1 показана диаграмма вложения множеств формул в зависимости от уменьшения ограничений на них. Сюда же можно было бы включить так называемые пропозициональные формулы, т.е. формулы ИВ, в формировании которых не участвует сигнатура, они оказались бы самым маленьким

множеством и включились бы в любую такую схему независимо от выбранной сигнатуры. Проблема в том, что такое вложение не будет теоретико-множественным, поскольку пропозициональных формул нет в грамматике выбранной сигнатуры.

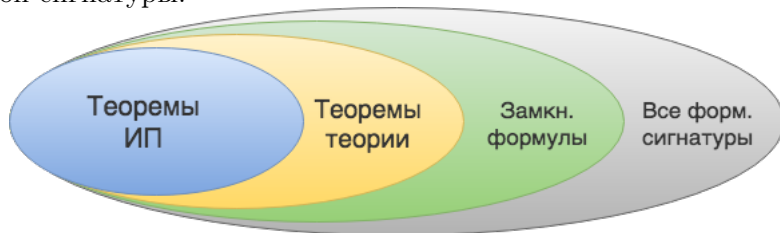


Рис. 4.1: Формулы языка $\mathcal{L}(\sigma)$.

С другой стороны, формулы «чистого» ИП хоть и зависят от выбранной сигнатуры, но их выводимость никак не привязана к этой сигнатуре, вследствие чего вся логическая арифметика «чистого» ИП — это манипуляции с формульными переменными и кванторами, и все теоремы «чистого» ИП остаются верными в любой теории первого порядка.

Приведем несколько примеров таких теорем:

- (1) если $\vdash \varphi$ и $\vdash \psi$, то $\vdash (\varphi \wedge \psi)$;
- (2) если $\vdash (\varphi \rightarrow \psi)$, то $\vdash (\neg\psi \rightarrow \neg\varphi)$;
- (3) если $\vdash (\varphi_1 \rightarrow \varphi_2)$ и $\vdash (\varphi_2 \rightarrow \varphi_3)$, то $\vdash (\varphi_1 \rightarrow \varphi_3)$;
- (4) $\vdash (\forall x \varphi) \rightarrow (\exists x \varphi)$;
- (5) $\vdash (\exists y \forall x \varphi) \rightarrow (\forall x \exists y \varphi)$;
- (6) если $\vdash (\varphi \rightarrow \psi)$, то $\vdash (\forall x \varphi) \rightarrow (\forall x \psi)$;
- (7) если $\vdash (\varphi \rightarrow \psi)$, то $\vdash (\exists x \varphi) \rightarrow (\exists x \psi)$;
- (8) $\vdash (\forall x \varphi) \leftrightarrow (\neg \exists x \neg \varphi)$;
- (9) $\vdash (\exists x \varphi) \leftrightarrow (\neg \forall x \neg \varphi)$;

Из этих же свойств можно получить и такое: замена подформулы на эквивалентную подформулу приводит к эквивалентной формуле. Например, если у нас $\varphi_1 \leftrightarrow \varphi_2$, то $(\varphi_1 \rightarrow \psi) \leftrightarrow (\varphi_2 \rightarrow \psi)$, и т.д.

Появление аксиоматики обусловлено требованием «обвязать» предикаты сигнатуры σ некоторыми зависимостями. Например, в теории ZF мы видели, что отношения принадлежности и равенства очень тесно связаны (взять хотя бы аксиому объемности). Таким образом, зная свойства предикатов, мы можем получать и новые выводы о них, т. е. теоремы теории.

В то же время, ценность любой теории будет ничтожна, если в ней можно будет вывести два противоречащих друг другу суждения. Назовем замкнутую формулу φ **независимой** от аксиоматики \mathcal{A} , если ни φ , ни $\neg\varphi$ не выводимы из \mathcal{A} . Если из \mathcal{A} можно одновременно вывести и φ и $\neg\varphi$ для некоторой замкнутой формулы φ , то теория \mathcal{A} называется **противоречивой**. Если независимых от \mathcal{A} формул в языке $\mathcal{L}(\sigma)$ не существует и \mathcal{A} непротиворечива, то аксиоматика (теория) \mathcal{A} называется **полной**. На схеме 4.2 изображены эти понятия.

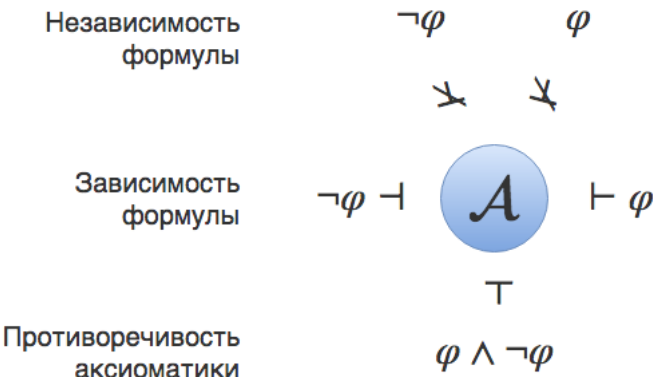


Рис. 4.2: Зависимость формул и аксиоматики.

Заметим, что наличие формулы φ , независимой от аксиоматики \mathcal{A} , вообще говоря, свидетельствует о непротиворечивости этой теории. Но проблема в том, что независимость формулы обычно доказывается построением таких моделей, что в одной из них выполняется $\mathcal{A} + \varphi$, а в другой $\mathcal{A} + \neg\varphi$. И вот тут мы упираемся в то, что само по себе построение моделей тоже происходит в какой-то теории (обычно — в **ZF**), а это значит, что непротиворечивость таких моделей и сам метод доказательства существенно опирается на непротиворечивость базовой теории (в которой строится модель, т.е. на непротиворечивость **ZF**). Таким образом, хотя рассуждение «если есть независимая формула, то теория непротиворечива» логически верно, но истинность его посылки может быть под вопросом.

Полноту теории не следует путать с полнотой исчисления (высказываний или предикатов), это радикально отличающиеся понятия. Полнота аксиоматики напоминает полноту системы векторов в линейном пространстве. Полная аксиоматика порождает все замкнутые формулы языка $\mathcal{L}(\sigma)$ в том смысле, что она делит пространство замкнутых формул ровно на две части — выводимые в данной аксиоматике и их отрицания.

Полнота исчисления говорит о том, что понятия истинности и выводимости совпадают. Это все равно что мы задали бы в линейном пространстве

некоторую рекурсивную процедуру (выводимость, построение), которая бы из базисных векторов построила бы всю их линейную оболочку (определение которой неконструктивно). В этом случае можно сказать, что такая процедура является полной.

4.1.3 Модели

Наконец, нужно определиться с тем, что такое истинность в исчислении предикатов. Когда мы рассматривали ИВ, для нас вопрос истинности сводился к тому, что формула должна быть тождественно истинной как функция булевых переменных. В случае ИП мы оторвались от земли, определив произвольную чисто символическую сигнатуру.

Поэтому с понятием сигнатуры неразрывно связано понятие **интерпретации**.

Под интерпретацией сигнатуры σ (и языка $\mathcal{L}(\sigma)$) мы будем понимать в дальнейшем произвольную алгебраическую структуру с изоморфной сигнатурой. Точнее, если у нас, с одной стороны, имеется сигнатура $\sigma = \langle R, F \rangle$ с набором отношений R и набором функциональных символов F , а с другой стороны, имеется алгебраическая структура (M, R', F') с носителем M такая, что между R и R' , а также между F и F' можно установить взаимно однозначное соответствие, сохраняющее арность отношений и операций, то структуру (M, R', F') мы и будем называть интерпретацией сигнатуры σ .

Проще говоря, мы находим множество с отношениями и операциями, которые по своей арности в точности соответствуют предикатным и функциональным символам нашей сигнатуры, тем самым мы овеществляем нашу чисто символическую конструкцию с помощью теоретико-множественной конструкции.

Произведя такое сопоставление, мы можем для любой формулы φ сигнатуры σ построить ровно по тем же грамматическим правилам соответствующую формулу в структуре модели M , просто заменяя всюду значки исходной сигнатуры на значки из интерпретации. Полученное выражение для формулы называют ее **релятивизацией** и обозначают φ^M .

Существует родственное определение. Предположим, что у нас имеется (незамкнутая) формула φ с параметрами x_1, \dots, x_k . Очевидно, что ее релятивизация φ^M является функцией из $M^k \rightarrow \{0, 1\}$, или предикатом. Говорят, что данный предикат **выражается** формулой φ . Все такие предикаты (а они еще являются и отношениями на M), а также их области истинности называются **выразимыми**.

Например, пусть задана сигнатура с единственным бинарным отношением $=$, тогда мы можем взять любое множество с *любым* бинарным отношением, и оно будет интерпретировать данную сигнатуру. Почему с *любым*? Потому что пока кроме сигнатуры нет никаких ограничений (кроме аксиом ИП), кроме

арности отношения, от нас ничего не требуется.

Предположим, что формула φ истинна в любой интерпретации (например, формула $R(a, b) \vee \neg R(a, b)$). Тогда она называется **общезначимой**. В этом случае пишут $\models \varphi$.

Для исчисления предикатов справедлива теорема, аналогичная теореме о полноте исчисления высказываний.

Метатеорема 4.4 (Гёделя о полноте ИП). *Замкнутая формула выводима в ИП тогда и только тогда, когда она истинна в любой интерпретации сигнатуры (общезначима):*

$$\vdash \varphi \Leftrightarrow \models \varphi$$

Для исчисления предикатов, как и для ИВ, справедлива

Метатеорема 4.5 (о компактности ИП). *Пусть Γ — произвольное множество замкнутых формул в некоторой сигнатуре. Если любое его конечное подмножество имеет модель, тогда Γ имеет модель.*

Если некоторая формула или набор формул Γ сигнатуры σ истинны в некоторой интерпретации M данной сигнатуры, то пишут $M \models \Gamma$ или (чем мы будем пользоваться) $\models_M \Gamma$. В этом случае говорят о **совместности** набора формул Γ . В случае, если под Γ подразумевается какая-то аксиоматика, то пишут также $\text{Con}(\Gamma)$, что означает совместность соответствующей теории.

Естественно, гораздо интереснее изучать интерпретации сигнатур с теориями (или проще — интерпретации теорий). В этом случае от интерпретации требуется не только попадание в форму отношений и функциональных символов сигнатуры (т.е. соблюдение арности), но и в логические связи между ними, т.е. чтобы внутри интерпретации выполнялись аксиомы теории \mathcal{A} или, что то же самое, чтобы было $\models_M \mathcal{A}$.

Если интерпретация такова, что в ней истинны аксиомы \mathcal{A} , то такая интерпретация называется **моделью** теории \mathcal{A} . Как обычно, если мы говорим, что множество M является моделью теории \mathcal{A} , то мы подразумеваем наличие алгебраической структуры с носителем M и наборами отношений и операций, сигнатура которых изоморфна сигнатуре теории \mathcal{A} .

Итак, совместность теории означает существование у нее модели (в предположении что ZF непротиворечива).

Метатеорема 4.6 (о полноте ИП в сильной форме). *Теория совместна тогда и только тогда, когда она непротиворечива.*

Эта теорема является следствием предыдущих двух теорем (Гёделя о полноте и компактности), а полное ее доказательство можно посмотреть в [12].

Подумайте почему. | Теперь становится понятным, что независимость формулы φ от аксиоматики \mathcal{A} означает как $\text{Con}(\mathcal{A} + \varphi)$, так и $\text{Con}(\mathcal{A} + \neg \varphi)$. Иначе

говоря, если мы можем построить (в ZF) две модели, в одной из которых истинна формула φ , а в другой ее отрицание, то мы тем самым докажем ее независимость от \mathcal{A} . На рис. 4.3 это изображено схематически.

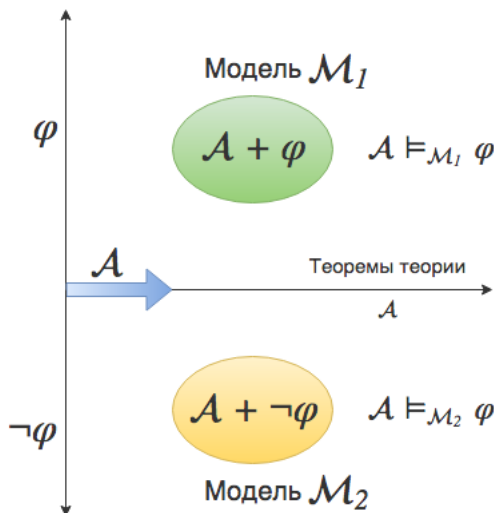


Рис. 4.3: Модели и независимость.

В то же время, говорят, что замкнутая формула φ **семантически следует** из \mathcal{A} , если она истинна в любой модели теории \mathcal{A} . Обозначение: $\mathcal{A} \models \varphi$.

Обозначение $\mathcal{A} \models_M \varphi$ мы зарезервируем за тем случаем, когда формула φ истинна в модели M , в которой также истинны аксиомы \mathcal{A} . Последнее не означает, что из \mathcal{A} можно вывести формулу φ .

С другой стороны, формула φ выводима из \mathcal{A} тогда и только тогда, когда φ семантически следует из \mathcal{A} (т.е. истинна в любой модели \mathcal{A}). Этот факт следует из теоремы Гёделя о полноте ИП (подробнее см. [12]).

В таблице 4.1 кратко собраны основные понятия о выводимости, непротиворечивости, совместности для случая исчисления высказываний, исчисления предикатов (чистого) и произвольной теории.

Среди формальных теорий в языке первого порядка важное место занимают *теории с равенством*. В этих теориях по умолчанию в довесок к аксиомам ИП добавляются, во-первых, символ отношения равенства ($=$), а во-вторых, следующие аксиомы:

EQ1 $\forall a \ a = a$ (рефлексивность равенства);

EQ2 $\forall a \forall b \ a = b \rightarrow b = a$ (симметричность равенства);

EQ3 $\forall a \forall b \forall c \ (a = b) \wedge (b = c) \rightarrow a = c$ (транзитивность равенства);

Таблица 4.1: Понятия теории моделей.

Свойство	ИВ	ИП (сигнатура)	Теория
Язык	Формулы и булевы переменные	Формулы, Предикаты, Термы, переменные и константы	=ИП
Аксиоматика	Аксиомы ИВ + modus ponens	ИВ + кванторы + обобщение	ИП + \mathcal{A} (доп. аксиомы)
Истинность формулы φ	\equiv 1. Тавтология (тожд. истинная формула)	$\models \varphi$. Общезначимая формула (истинная во всех интерпретациях)	$\mathcal{A} \models \varphi$. Истинна во всех моделях, в которых истинна аксиоматика
Выводимость φ	$\vdash \varphi$ из аксиом ИВ	$\vdash \varphi$ из аксиом ИП	$\mathcal{A} \vdash \varphi$ из аксиом + ИП
Полнота исчисления (1 форма)	Выводимость равносильна тавтологичности ($\vdash \Leftrightarrow \equiv 1$)	Выводимость равносильна общезначимости ($\vdash \Leftrightarrow \models$)	Выводимость равносильна истинности во всех моделях аксиоматики ($\mathcal{A} \vdash \Leftrightarrow \mathcal{A} \models$)
Непротиворечивость набора формул	Невозможно $\Gamma \vdash \varphi$ и $\Gamma \vdash \neg\varphi$	то же	Невозможно $\mathcal{A} \vdash \varphi$ и $\mathcal{A} \vdash \neg\varphi$
Совместимость набора формул	Выполнимость (все формулы Γ истинны при некоторых значениях переменных)	$\models_M \Gamma$ (все формулы Γ истинны в некоторой интерпретации M)	$\text{Con}(\mathcal{A})$ (все аксиомы теории истинны в некоторой интерпретации M)
Полнота исчисления (2 форма)	Совместимость равносильна непротиворечивости	Совместимость равносильна непротиворечивости	Совместимость равносильна непротиворечивости
Полнота набора формул	$\Gamma \vdash \varphi$ или $\Gamma \vdash \neg\varphi$ (φ — произвольная формула)	$\Gamma \vdash \varphi$ или $\Gamma \vdash \neg\varphi$ (φ — произвольная замкнутая формула)	либо φ , либо $\neg\varphi$ является теоремой в теории \mathcal{A}

EQ4 $\forall a\forall b\forall c\forall d (a = b) \wedge (c = d) \rightarrow (R(a, c) \rightarrow R(b, d))$ (сохранение предикатов)
для всякого отношения R сигнатуры σ ;

EQ5 $\forall a\forall b\forall c\forall d (a = b) \wedge (c = d) \rightarrow (F(a, c) = F(b, d))$ (сохранение операторов)
для всякого функционального символа F сигнатуры σ .

В последних двух требованиях для простоты указано по 2 аргумента у отношений и операторов, на самом деле эти аксиомы нужно экстраполировать на произвольную арность отношений и операторов сигнатуры σ . Здесь мы видим полный аналог определения равенства в теории множеств, где мы требовали сохранения отношения принадлежности при замене его аргументов на равные (аксиомы AV1 и AE1, см. раздел 1.2.1, AV1 при этом еще задает остальные свойства равенства множеств, так что в ZF аксиомы EQ1–EQ5 равносильны AV1+AE1).

Как и в разделе 1.2.1 из аксиом для отношений и операторов нетрудно доказать индукцией по сложности формулы и терма более сильную форму аксиомы равенства: равенство сохраняет любой предикат и любую формулу языка $\mathcal{L}(\sigma)$.

Если теория с равенством имеет модель, в которой равенство теории интерпретируется равенством в модели (т. е. равенством в смысле теории ZF), то такая модель называется **нормальной**. Если же равенство теории моделируется отношением эквивалентности, то всегда можно перейти к фактор-модели, где равенство классов эквивалентности будет моделировать равенство теории, и, таким образом, фактор-модель будет нормальной.

Архетип факторизации.

Отметим, что на протяжении последних абзацев мы излагали некоторые факты и определения относительно языков и формул. По сути, мы занимались теорией, предметом изучения которой являются формальные теории. Такие рассуждения и даже строгие доказательства о формальных языках и теориях принято называть **метатеорией**. Именно поэтому в данном разделе все теоремы мы именуем метатеоремами. Интересно, что само исчисление высказываний и предикатов является формализацией метатеории и тоже может быть предметом изучения вышестоящей метатеории. Это напоминает переход от множеств к классам, а затем к суперклассам и т.д. Важно не заблудиться в этом разнообразии.

4.1.4 Ультрастепени

Существуют, однако, методы, позволяющие не просто прыгать между эквивалентными языками и моделями, занимаясь лишь переобозначениями термов и формул, но и конструировать новые языки и формулы. Тут, правда, снова не обойтись без солидной помощи теории множеств.

Важную роль при построении моделей в теории множеств играет конструкция, именуемая **ультрапроизведением**. Мы рассмотрим более простой случай — **ультрастепень**. Пусть X — непустое множество, на котором заданы операции $F_t(x_1, \dots, x_{n_t})$, где $F_t : X^{n_t} \rightarrow X$, и предикаты $\varphi_t(x_1, \dots, x_{m_t})$, отвечающие за отношения на X ,⁵ т. е. некоторая сигнатура. Таким образом, мы имеем алгебраическую структуру с каким-то набором операций и отношений (заданных предикатами φ_t).

Пусть I — некоторое непустое множество индексов (например, ординал).

Пусть \mathcal{U} — ультрафильтр (в данном случае неважно, главный или неглавный, так что зависимость от аксиомы выбора здесь не требуется) на I (определение см. в разделе 2.4.5). Рассмотрим множество всех функций из I в X

⁵Индекс t , вообще говоря, пробегает произвольное множество, которое может быть и пустым, и конечным, и бесконечным.

и определим на них отношение эквивалентности:

$$f \sim g, \text{ если } \{i \in I \mid f(i) = g(i)\} \in \mathcal{U}.$$

Обозначим

$$U \equiv X^I / \sim$$

фактор-множество по данному отношению. Множество U называется **ультрастепенью** X (относительно ультрафильтра \mathcal{U}).

Наконец, для операции F_t определим операцию F_t^U по правилу:

$$F_t^U([f_1], \dots, [f_{n_t}]) = [g], \text{ если } \{i \in I \mid F_t(f_1(i), \dots, f_{n_t}(i)) = g(i)\} \in \mathcal{U},$$

для предиката φ_t определим предикат φ_t^U по правилу:

$$\varphi_t^U([f_1], \dots, [f_{m_t}]) \leftrightarrow \{i \in I \mid \varphi_t(f_1(i), \dots, f_{m_t}(i))\} \in \mathcal{U}.$$

Таким образом, мы осуществили следующие переходы:

$$\begin{aligned} X &\rightsquigarrow U = X^I / \sim \\ x \in X &\rightsquigarrow [f : I \rightarrow X] \\ F_t : X^k &\rightarrow X \rightsquigarrow F_t^U : U^k \rightarrow U \\ \varphi_t &\rightsquigarrow \varphi_t^U \end{aligned}$$

Наконец, рассмотрим произвольную формулу φ в *сигнатуре алгебраической структуры* X , т. е. формулу, построенную по правилам построения формул исключительно из предикатов φ_t и атомарных функциональных термов, соответствующих операциям F_t . Затем произведем в этой формуле замену переменных $x_i \in X$ на переменные $u_i = [f_i] \in U$, замену атомарных формул φ_t на их релятивизации φ_t^U , замену функциональных символов F_t на функциональные символы F_t^U . В итоге получим релятивизацию φ^U формулы φ , заданной в сигнатуре X , которая будет формулой в сигнатуре ультрастепени U .

Основное свойство ультрастепени заключается в том, что истинность исходной формулы эквивалентна истинности ее релятивизации. Более точно, верна

Метатеорема 4.7 (Лося об ультрастепени).

$$U \models \varphi^U([f_1], \dots, [f_m]) \leftrightarrow \{i \in I \mid X \models \varphi(f_1(i), \dots, f_m(i))\} \in \mathcal{U}$$

Здесь выражение $X \models \varphi(f_1(i), \dots, f_m(i))$ означает, что в модели X формула φ истинна, но поскольку мы изначально предполагали, что φ построена именно в сигнатуре модели X , то запись $X \models$ перед формулой можно пропускать. В общем же случае, если φ рассматривается как формула некоего

внешнего языка, то здесь следует подставлять именно ее релятивизацию в сигнатуре X и записывать выводимость в X , как в модели.

Доказательство теоремы Лося производится индукцией по сложности формулы.

Комментарий 14.

Какова мощность ультрастепени конечного множества X , если $||X|| = n$?

Казалось бы, ответ зависит от выбора ультрафильтра. Или, по крайней мере, нужно рассмотреть все возможные ультрафильтры на множестве мощности n (кстати, сколько их?) и посмотреть, как ведет себя отношение эквивалентности, чтобы оценить количество классов эквивалентности.

Но в данном случае все гораздо проще. Нам достаточно написать формулу φ , которая выражала бы высказывание о том, что мощность X не превосходит n . Например,

$$\varphi \Rightarrow \forall x_1 \dots \forall x_{n+1} (x_1 = x_2) \vee (x_1 = x_3) \vee \dots \vee (x_n = x_{n+1}),$$

т. е. среди $n + 1$ элементов всегда найдутся совпадающие.

Затем эту формулу нужно переписать в терминах ультрафильтра (отличие будет только в области действия кванторов). По теореме Лося она также будет истинной, а это и означает, что в ультрастепени не более n элементов. Аналогично, взяв отрицание φ для n переменных (вместо $n + 1$), доказывается, что в ультрастепени не менее n элементов.

На этом шаге теории моделей мы, наконец, подошли к обоснованию модели гипердействительных чисел, построенной как ультрастепень \mathbb{R} и описанной в разделе 2.4.5, и теоремы Робинсона (2.8).

Более того, наконец, мы можем совершенно точно сформулировать, что мы подразумеваем под «понятием» в теории множеств и чем оно отличается от «естественного» объекта теории множеств (см. рассуждения на стр. 133). Естественный объект — это элемент универсума множеств, а понятие — это общее название (элементов) предметной области произвольной теории, имеющей модель в теории множеств. Например, такое понятие как группа есть название всей предметной области, изучаемой теорией групп. Соответственно, любая модель теории групп является группой. А действительное число — это любой элемент из предметной области в теории действительных чисел. Соответственно, если мы фиксируем какую-либо модель теории действительных чисел, то элементы такой модели мы называем действительными числами.

Одновременно мы стали лучше понимать, какие высказывания, записанные на языке теории множеств можно применять для изучения того или иного понятия на модели, а какие — нет. А именно, все высказывания ZF , которые построены на функциях и отношениях, соответствующих сигнатуре данной

теории, будут применимы для изучения такой теории, прочие же — нет. Это отличие мы подчеркивали на стр. 142 без прямой отсылки к теории моделей. Например, если мы рассматриваем модель \mathbb{Z} с сигнатурой $\sigma = \langle +, *, <, 0, 1 \rangle$, построив в \mathbf{ZF} моделирующее множество Z и функции и отношение, соответствующие операциям, предикату и константам сигнатуры σ , то для изучения свойств целых чисел (с данной сигнатурой) можно пользоваться только замкнутыми формулами, построенными из этих функций и отношения по правилам грамматики. Привлечение любых других сведений о множестве Z и его внутренней структуре будет ошибочным.

Дадим еще несколько определений.

Через $\text{Th}_\sigma(M)$ обозначим совокупность всех истинных в интерпретации M замкнутых формул сигнатуры σ . $\text{Th}_\sigma(M)$ называется **элементарной теорией интерпретации** M . Такая теория, очевидно, полна. Если рассматривается только одна сигнатура, то принято опускать ее обозначение: $\text{Th}(M)$.

Если $\text{Th}(M_1) = \text{Th}(M_2)$, то интерпретации M_1 и M_2 называются **элементарно эквивалентными**.

Далее, теория с конечной сигнатурой называется **разрешимой**, если существует алгоритм, который по произвольной замкнутой формуле за конечное число шагов определяет, выводима ли она в этой теории или нет (т.е. алгоритм распознает теоремы теории).

Смежное определение: теория с конечной сигнатурой называется **эффективно аксиоматизируемой**, если существует алгоритм, распознающий ее аксиомы. Заметим, что теория разрешима тогда и только тогда, когда она полна и эффективно аксиоматизируема.

Метатеорема 4.8 (Чёрча о неразрешимости исчисления предикатов). *Не существует алгоритма, проверяющего общезначимость формул первого порядка.*

Ниже мы увидим, что добавление аксиом в теорию может сделать ее разрешимой.

Теория с равенством, имеющая конечную или счётную сигнатуру, называется **категоричной в счётной мощности**, если все её счётные нормальные модели изоморфны.

4.1.5 Примеры формальных теорий

Приведем теперь несколько примеров теорий и обсудим некоторые их свойства.

Исчисление высказываний и предикатов

ИВ является непротиворечивым, полным (и как исчисление, и как теория) и разрешимым.

ИП является непротиворечивым и полным как исчисление, но не является ни полным как теория, ни разрешимым.

Здесь нужно отметить, что полнота ИП как теории вообще нечто неопределенное. Если мы рассматриваем ИП без сигнатуры, то мы по сути имеем дело с ИВ, которое полно. Но если мы добавляем какой-то предикатный символ R и не связываем его аксиомами, то высказывание вида $\forall a \exists b R(a, b)$ в разных интерпретациях может быть как истинным, так и ложным. Соответственно, ни о какой полноте ИП с непустой сигнатурой речи быть не может. Неразрешимость ИП есть содержание теоремы Чёрча.

Чистая теория равенства

Рассмотрим сигнатуру с единственным предикатным символом $=$. Добавим три аксиомы равенства: рефлексивность, симметричность и транзитивность. Аксиомы EQ4 и EQ5 для такой тривиальной сигнатуры не актуальны, поскольку в них вместо R можно подставить только $=$, а вместо F подставить и вовсе нечего.

Моделями данной теории будут все множества с отношениями эквивалентности, нормальными моделями — множества произвольной мощности. Теоремами этой теории будут все формулы с равенством, истинные в множествах любой мощности. В зависимости от мощности модели в дополнение к теоремам теории равенства могут появляться истинные в данной модели формулы. Например, если модель есть двухточечное множество $\{0, 1\}$, то формула $\exists x \exists y x \neq y$ истинна в этой модели, но ложна в любом одноэлементном множестве.

Известно, что чистая теория равенства разрешима, но не полна. Последнее следует, например, из того, что приведенная выше формула не зависит от аксиом равенства (она истинна в одной модели и ложна в другой).

Полугруппы

Теория полугрупп — это теория с сигнатурой $\langle =, \circ \rangle$, где \circ обозначает полугрупповую операцию. Ее аксиомы — это аксиомы чистой теории равенства вместе с аксиомой EQ5, где вместо F нужно подставлять \circ , а также предметная аксиома ассоциативности операции \circ .

Нормальные модели этой теории называются полугруппами (см. раздел 2.4.1).

Известно, что теория полугрупп (т. е. множество всех замкнутых формул,

истинных во всех полугруппах) неразрешима. И она неполна по той же причине, что и теория равенства.

Если помимо аксиомы ассоциативности в теорию полугрупп добавить аксиомы нейтрального и обратных элементов, то мы получим теорию групп. Нормальные модели теории групп называются группами.

Эта теория также неразрешима и неполна.

Интересно, что если мы рассмотрим теорию абелевых групп, т. е. добавим к аксиоматике теории групп еще и аксиому коммутативности операции \circ , то мы получим разрешимую теорию (но по-прежнему неполную).

Плотные линейно упорядоченные множества

Рассмотрим сигнатуру $\langle =, < \rangle$ и примем аксиомы равенства EQ1–EQ4 (вместо R подставляем знак $<$), а также аксиомы плотного линейно упорядоченного множества, не ограниченного ни сверху, ни снизу (или, как говорят, без первого и последнего элементов) — см. раздел 1.3.1.

Любая модель рациональных чисел образует счетную модель этой теории, а действительных — несчетную.

Эта теория категорична в счетной мощности (теорема 1.12), полна и разрешима!

Прям-таки синтаксический сахар, а не теория!

Целые числа

Рассмотрим сигнатуру $\langle =, S(_), 0 \rangle$ с равенством, одноместным функциональным символом S и константой 0 . Добавим сюда аксиомы равенства EQ1, EQ2, EQ3, EQ5,⁶ а также следующие аксиомы:

$$S1 \quad \forall x \forall y (S(x) = S(y)) \rightarrow (x = y);$$

$$S2 \quad \forall x \exists y (S(y) = x);$$

$$S3 \quad \forall x (x \neq S(x));$$

$$S4 \quad \forall x (x \neq S(S(x)));$$

$$S5 \quad \forall x (x \neq S(S(S(x))));$$

$$S6 \quad \dots$$

Таким образом, в нашей теории счетный набор аксиом.

Здесь унарный оператор $S(x)$ интерпретируется как $x + 1$, а простейшей моделью является любая модель \mathbb{Z} . Данная теория полна и разрешима. В отличие от предыдущего примера эта теория не является категоричной в

⁶Напоминаем, что константа — это нульместный функциональный символ, поэтому для него также необходима аксиома EQ5.

счетной мощности. Например, $\mathbb{Z} + \mathbb{Z}$ (в смысле прямой суммы множеств) является ее моделью, но не изоморфна \mathbb{Z} .

Данная теория совпадает с $\text{Th}(\mathbb{Z}, =, S, 0)$.

Целые числа с порядком

Добавим к предыдущей сигнатуре отношение $<$ с аксиомами линейного порядка, причем потребуем, чтобы у каждого элемента был последующий и предыдущий:

$$\forall x \exists y (x < y) \wedge \neg(\exists z x < z < y),$$

$$\forall x \exists y (y < x) \wedge \neg(\exists z y < z < x).$$

При этом, если положить, что $S(x)$ есть следующий за x по порядку элемент, то нам не потребуется счетный набор аксиом $S1, S2$ и т.д. из предыдущей теории, они станут теоремами.

Данная теория совпадает с $\text{Th}(\mathbb{Z}, =, <, S, 0)$, причем она является конечно аксиоматизируемой. Кроме того, она полна, разрешима, но не категорична в счетной мощности.

Формальная арифметика

Рассмотрим теорию $\text{Th}(\omega, =, +, \cdot, S, 0)$, которая содержит все истинные на ω замкнутые формулы сигнатуры $\langle =, +, \cdot, S, 0 \rangle$ со свойствами сложения и умножения натуральных чисел. Эта теория полна (как всякая элементарная теория модели), но неразрешима. Кроме того, она не является конечно аксиоматизируемой (т. е. ее аксиоматический базис не может содержать конечное число аксиом). Более того, у этой теории не существует разрешимого множества теорем, порождающего (в смысле выводимости) все теоремы данной теории, т. е. не существует и разрешимой аксиоматики данной теории.

Тем не менее, мы укажем в явном виде ту аксиоматику, которая является классической аксиоматикой арифметики. Итак, **аксиоматика Пеано** включает аксиомы равенства (EQ4 можно не включать, т. к. в сигнатуре нет предикатных символов, кроме $=$), а также следующие аксиомы:

PA1 $\forall x S(x) \neq 0$ (особая точка S);

PA2 $\forall x \forall y (S(x) = S(y)) \rightarrow x = y$ (инъективность S);

PA3 $\forall x x + 0 = x$ (нейтральность нуля по $+$);

PA4 $\forall x \forall y x + S(y) = S(x + y)$ (согласованность S и $+$);

PA5 $\forall x x \cdot 0 = 0$ (мультипликативное свойство нуля);

PA6 $\forall x \forall y x \cdot S(y) = x \cdot y + x$ (согласованность S и \cdot);

РА7 $(\varphi(0) \wedge (\forall n \varphi(n) \rightarrow \varphi(S(n)))) \rightarrow \forall n \varphi(n)$ для любой формулы φ данной сигнатуры (схема аксиом индукции).

Из этих аксиом выводятся, в частности, такие теоремы:

РАТ1 $\forall x \forall y \forall z (x + y) + z = x + (y + z)$ (ассоциативность $+$);

РАТ2 $\forall x \forall y \forall z x(y + z) = xy + xz$ (левая дистрибутивность);

РАТ3 $\forall x \forall y \forall z (xy)z = x(yz)$ (ассоциативность \cdot);

РАТ4 $\forall x xS(0) = x \wedge S(x) = x + S(0)$ (свойства единицы);

РАТ5 $\forall x \forall y x + y = y + x$ (коммутативность $+$);

РАТ6 $\forall x \forall y xy = yx$ (коммутативность \cdot);

РАТ7 $\forall x (x \neq 0) \rightarrow (\exists y x = S(y))$ (предшествующий элемент).

Упражнение
4.3.
Докажите,
пользуясь
вложенной
индукцией.

Заметим, что аксиомы сложения и умножения (РА3–РА6) можно рассматривать как рекурсивные определения соответствующих операций в сигнатуре $\langle =, S, 0 \rangle$, правда, тогда потребуется обоснование рекурсивного определения. А оно требует не только аксиомы индукции, но и либо некоторых теоретико-множественных конструкций (например, мы не можем в РА определить понятие функции), либо специальной аксиомы рекурсии, позволяющей строить рекурсивные определения функциональных символов. Подробнее об этом см. в [15].

Такие определения, как аксиомы РА3–РА6 называются неявными, они напоминают определения функций с помощью уравнений. В разделе 2.1 мы строили операции сложения и умножения именно рекурсивным способом, имея в своем арсенале мощный аппарат ZF.

Явные же определения представляют собой переобозначение какого-либо терма или предиката и не требуют усиления формализма. Так, отношение $<$ мы можем ввести в нашей аксиоматике следующим образом:

$$(x < y) \Leftrightarrow \exists z (z \neq 0) \wedge (y = x + z).$$

Упражнение 4.4. Для этого предиката можно вывести все стандартные свойства линейного порядка: антисимметричность, транзитивность, связность, а также его согласованность с арифметическими операциями. Для доказательства потребуется только индукция и уже полученные выше теоремы.

Мы видим, что арифметика Пеано, несмотря на свою «продвинутость» в математике и логике, все-таки не дает в полном объеме тот инструментарий, который очевиден на интуитивном уровне. Тем не менее, даже у нее имеется

ряд «неудобных» проблем. А именно, аксиоматика Пеано неполна (это, собственно, и есть содержание знаменитой **теоремы Гёделя о неполноте**). Это означает, что существуют утверждения, формулируемые в сигнатуре арифметики, но независимые от ее аксиоматики. Наиболее простой и яркий пример такого утверждения — теорема Гудстейна (см. раздел 1.1.10), которую мы доказывали, выходя за рамки арифметики натуральных чисел.

Непротиворечивость арифметики была доказана в 1936 году Генценом с помощью трансфинитной индукции до ординала ε_0 , т. е. примерно тем же инструментарием, которым мы пользовались при доказательстве теоремы Гудстейна. Собственно, известно, что теорема Гудстейна эквивалентна арифметическому утверждению, означающему совместность арифметики Пеано [37].

Стоит отметить, что на сегодняшний день известно несколько различных подходов к аксиоматизации арифметики. Например, арифметика Пресбургера [12] для целых чисел, имеющая моделью структуру $(\mathbb{Z}, =, <, +, 0, 1)$ (без умножения), разрешима и полна(!), но не является конечно аксиоматизируемой (все из-за той же схемы аксиом индукции, что и в арифметике Пеано).

Кроме того, известно, что аксиоматика Пеано имеет и нестандартные модели, в которых существует число, отличное от всех натуральных чисел, получаемых из 0 последовательным прибавлением 1. И это снова отсылает нас к нестандартному анализу и бесконечно большим числам.

Вещественно замкнутые поля

В предыдущем примере мы видели, казалось бы, простейшую арифметическую аксиоматику, которая, тем не менее, оказывается весьма сложной своими логическими свойствами (неполнота теории и бедность инструментария). Тем удивительнее будет следующий пример.

Рассмотрим сигнатуру $\langle +, \cdot, <, 0, 1 \rangle$ и предположим, что над ней задана аксиоматика упорядоченного поля (см. раздел 2.4.1), т. е. операции $+$ и \cdot ассоциативны, коммутативны, дистрибутивны, 0 является нейтральным элементом по сложению, 1 — по умножению, для каждого элемента существует противоположный, а для каждого ненулевого — обратный, кроме того, отношение $<$ является отношением линейного порядка и согласовано с операциями сложения и умножения.

Упорядоченное поле называется **вещественно замкнутым**, если любой многочлен, имеющий на концах отрезка разные знаки, имеет корень на этом отрезке.⁷

Теория, состоящая из аксиом равенства, аксиом линейно упорядоченного поля и аксиомы вещественной замкнутости, называется теорией вещественно

⁷Это одно из нескольких эквивалентных определений вещественно замкнутого поля. В алгебре [55] можно встретить, например, такое: поле вещественно замкнуто, если в нем -1 невозможно представить в виде суммы квадратов.

замкнутых полей. В этой теории выполняются основные факты о многочленах и производных (определяемых алгебраически, а не через пределы), а также содержится модель целых чисел.

Кроме того, такая теория совпадает с элементарной теорией вещественных чисел (с той же сигнатурой), полна и разрешима, и, кроме того, все вещественно замкнутые упорядоченные поля элементарно эквивалентны. Тем не менее, она не категорична в счетной мощности, поскольку конечные расширения $\mathbb{Q}[x_1, \dots, x_n]$ не изоморфны (можно сравнить, например, $\mathbb{Q}[e]$ и $\mathbb{Q}[\pi]$).

Единственное, чего здесь нет — это индукции.

Аксиоматика Тарского для планиметрии

Ранее мы уже касались различных видов геометрии, в том числе неевклидовых (см. раздел 3.6.6), однако не заостряли внимание на формальной стороне дела, рассматривая различные модели в рамках теории множеств (точнее, модели внутри моделей \mathbb{R}^n и \mathbb{C}^n).

Здесь мы приведем наиболее современный вариант аксиоматики планиметрии (геометрии плоскости) — **аксиоматику Тарского**. Данная аксиоматика удобна тем, что излагается полностью в языке первого порядка с довольно простой грамматикой и одним видом сущностей — *точками*. Точки обозначаются строчными латинскими буквами. Заметим, что мы не используем ни понятие множества точек, ни понятие принадлежности. Вместо этого вводятся три атомарных предиката отношений:



Альфред
Тарский

- бинарное: $x = y$, означающее *равенство* точек;
- тернарное:⁸ $x \leq y \leq z$, означающее, что *y лежит между x и z* (допускается, что $y = x$ или $y = z$);
- тетрарное: $xy \equiv uv$, означающее *конгруэнтность* отрезков xy и uv , т. е. возможность совместить их движением.

Несмотря на то, что мы вложили определенный смысл в приведенные обозначения, формально это всего лишь три атомарных отношения: первое от двух аргументов, второе — от трех, третье — от четырех, и никакой формальной нагрузки символы $=$, \leq и \equiv не несут. Так же, как запись xy не является записью объекта (отрезка).

⁸В оригинале и современных работах используется обозначение $Bxyz$ от англ. *Betweenness*. Мы здесь нарушили эту традицию для того, чтобы более привычно записать это отношение. Не следует только воспринимать такую запись как конъюнкцию двух отношений порядка!

Все остальные формулы языка геометрии Тарского строятся по тем же грамматическим правилам, что и язык ZF , т. е. атомарные формулы заключаются в скобки, соединяются логическими связками и кванторами. Причем в этом языке нет символов-констант, нет функциональных термов и нет термов-кванторов.

Далее необходимо сформулировать аксиомы, которые зададут правила оперирования с данными символами.

Прежде всего, должны выполняться **аксиомы равенства** EQ1–EQ4.⁹

Наконец, переходим к содержательной части теории. Аксиоматика Тарского представлена на основе [8] и [107].

Tar1 $ab \equiv ba$ (рефлексивность конгруэнтности).

Tar2 $(ab \equiv pq) \wedge (ab \equiv rs) \rightarrow (pq \equiv rs)$ (транзитивность конгруэнтности).

Tar3 $(ab \equiv cc) \rightarrow (a = b)$ (равенство вырожденному отрезку влечет равенство начала и конца).

Tar4 $\exists x ((q \leq a \leq x) \wedge (ax = bc))$ (от любой точки a можно отложить любой отрезок bc в любом направлении qa).

Tar5 если $(a \neq b) \wedge (a \leq b \leq c) \wedge (a' \leq b' \leq c') \wedge (ab \equiv a'b') \wedge (bc \equiv b'c') \wedge (ad \equiv a'd') \wedge (bd \equiv b'd')$, то $cd \equiv c'd'$ (равенство пятых отрезков, если равны первые, вторые, третьи и четвертые в связной треугольной конструкции).¹⁰

Эта аксиома заменяет признак равенства треугольников по двум сторонам и углу между ними, не используя понятие «угол».

Tar6 $(a \leq b \leq a) \rightarrow (a = b)$ (тождественность отношения *лежать между*).

Tar7 Аксиома Паша: $(a \leq p \leq c) \wedge (q \leq c \leq b) \rightarrow \exists x (a \leq x \leq q) \wedge (b \leq p \leq x)$ (две диагонали четырехугольника $abcs$ пересекаются в некоторой точке p).

Tar8 $\text{Dim} \geq 2$: $\exists x \exists y \exists z \neg(x \leq y \leq z) \wedge \neg(y \leq z \leq x) \wedge \neg(z \leq x \leq y)$ (т. е. существуют три точки, не лежащие на одной прямой, аксиома плоскости).

Tar9 $\text{Dim} \leq 2$: если $(ap_1 \equiv ap_2) \wedge (bp_1 \equiv bp_2) \wedge (cp_1 \equiv cp_2) \wedge (p_1 \neq p_2)$, то $(a \leq b \leq c) \vee (b \leq c \leq a) \vee (c \leq a \leq b)$ (любые три точки, равноудаленные от двух других, лежат на одной прямой, аксиома плоскости).

⁹В аксиоме EQ4 вместо предикатного символа можно подставлять равенство, «лежит между» и конгруэнтность с учетом асимметрии этих отношений. Аксиома EQ5 здесь не требуется, поскольку в сигнатуре нет функциональных символов.

¹⁰В работе [107] показано, что если в этой аксиоме вывод $cd \equiv c'd'$ заменить на $dc \equiv c'd'$ (переставить местами d и c), то аксиому рефлексивности можно доказать, т. е. ее можно исключить из списка аксиом.

Tar10 Аксиома Евклида: $(a \leq d \leq t) \wedge (b \leq d \leq c) \wedge (a \neq d) \rightarrow \exists x \exists y ((a \leq b \leq x) \wedge (a \leq c \leq y) \wedge (y \leq t \leq x))$.

Эта аксиома является одной из эквивалентных форм аксиомы о параллельных (пятый постулат).

Tar11 Непрерывность: если $\exists u \forall x \forall y (\varphi(x) \wedge \psi(y) \rightarrow (u \leq x \leq y))$, то $\exists v \forall x \forall y (\varphi(x) \wedge \psi(y) \rightarrow (x \leq v \leq y))$ (если на луче с вершиной u точки, заданные $\varphi(x)$ лежат левее точек, заданных $\psi(y)$, то между этими точками есть разделяющая точка v).¹¹

Эта аксиома постулирует существование сечения двух определяемых формулами множеств точек на прямой. В полном объеме, то есть для произвольных подмножеств прямой, эта аксиома не выразима в логике первого порядка (если не использовать ZF).

В качестве упражнения докажите следующие теоремы:

Упражнение
4.5.

1. $ab \equiv ab$;
2. $(ab \equiv cd) \rightarrow (cd \equiv ab)$;
3. $a \leq b \leq b$;
4. $(a \leq b \leq c) \rightarrow (c \leq b \leq a)$ (используется аксиома Паша);
5. $(a \leq b \leq d) \wedge (b \leq c \leq d) \rightarrow (a \leq b \leq c)$ (используется аксиома Паша).

В аксиомах Тарского кое-где можно усмотреть первоначальный смысл постулатов Евклида, например, Tar4 означает, что любой отрезок можно неограниченно продлевать в заданном направлении. Проблема весьма неявного соответствия известных со школы аксиом Евклида современным аксиомам планиметрии кроется в том, что в 19–20 веках они прошли несколько стадий уточнений и переформулировок, пока не были оформлены Д. Гильбертом и А. Тарским в формализме по канонам гильбертовской программы формализации математики. Тем не менее, при достаточной сноровке постулаты Евклида можно сформулировать и доказать в аксиоматике Тарского.

Кроме того, если выйти за пределы языка первого порядка, то систему аксиом Тарского можно усилить так, что аксиома непрерывности будет сформулирована в общем виде — для произвольных подмножеств прямой. Это делает аксиоматику Тарского полностью эквивалентной аксиоматике Гильберта для евклидовой плоскости. Кроме того, такая усиленная аксиоматика

¹¹Дополнительно нужно указать, что u и v не входят свободно в формулы φ и ψ , кроме того, x не входит свободно в ψ , y не входит свободно в φ .

планиметрии оказывается категоричной в том смысле, что любая ее модель будет изоморфна евклидовой плоскости.

Тарским доказано, что приведенная выше система аксиом (как первого порядка, так и усиленная) полна и, более того, разрешима.

Теория множеств Цермело-Френкеля

Последний штрих к портрету: собственно, сама теория множеств. Обычно вопрос о ее совместности сводится к опыту. За 100 лет математики не смогли найти противоречие, следовательно, теория скорее совместна, чем противоречива. Символическое построение теории «начальных» множеств, проделанное нами в первой главе, заставляет укрепиться в этом опыте.

Если мы предполагаем совместность ZF, то наличие независимой замкнутой формулы в языке ZF означает ее неполноту. Такие утверждения есть, например, аксиома выбора. Так что ZF неполна.

Неразрешимость ZF следует из неразрешимости арифметики.

В конце соберем в таблицу сравнения (см. таб. 4.2) свойства приведенных выше примеров теорий.

Таблица 4.2: Сравнение теорий.

	Полнота	Разрешимость	Конечная аксиоматика	Категоричность в счетной мощности
Равенство	✗	✓	✓	✓
Полугруппы и группы	✗	✗	✓	✗
Абелевы группы	✗	✓	✓	✗
Плотные ЛУМ (<)	✓	✓	✓	✓
Целые числа без операций	✓	✓	✗	✗
Целые числа без операций с порядком	✓	✓	✓	✗
Формальная арифметика	✗	✗	✗	✗
Арифметика Пресбургера	✓	✓	✗	✗
Вещественно замкнутые поля	✓	✓	✓	✗
Геометрия Тарского	✓	✓	✗	✗
Теория множеств	✗	✗	✗	✗

Как уже отмечалось, вместо аксиоматик первого порядка для теории множеств и геометрии можно задать более «продвинутые» системы аксиом, эквивалентные нашим и содержащие только конечный набор аксиом. Поэтому существование конечной аксиоматики для ZF и геометрии можно также считать положительным.

4.2 Аксиома выбора: полезная и странная

На протяжении всей книги мы неоднократно анонсировали появление на сцене этой исторической аксиомы, и вот, наконец, пробил ее час. Встречаем:

$$\text{AC} : \quad \forall a \exists f (f : a \rightarrow \cup a) \wedge ((\forall x \neq \emptyset) f(x) \in x).$$

Мы намеренно не стали разворачивать определение функции в этом высказывании, чтобы сохранить удобочитаемость. Функция f , существование которой для произвольного множества a постулирует данная аксиома, называется **функцией выбора** для элементов множества a .

Формулировкой этой аксиомы, и вообще тем, что она была замечена как необходимое логическое звено в доказательствах ряда теорем анализа, мы обязаны Дж. Пеано (1890) и Б. Леви (1902), а боевым крещением этой аксиомы — Э. Цермело (1904), с работ которого начался такой непростой путь новоявленной аксиомы выбора через всю математику.

Прежде всего, приведем утверждения, которые эквивалентны AC:

ZT **Теорема Цермело**: всякое непустое множество может быть вполне упорядочено;

ZL **Лемма Цорна**: если в ч.у.м. любая цепь ограничена, то в нем существует максимальный элемент;

HMP **Принцип максимума Хаусдорфа**: в любом ч.у.м. существует максимальное (по вложению) л.у.м.

I. Покажем что из AC следует теорема Цермело. Пусть X — непустое множество. Положим $f(0) = x \in X$, где 0 — наименьший ординал. Дальнейшая нумерация элементов X строится следующим образом: выбрасываем из X все ранее пронумерованные элементы, и в оставшемся множестве, если оно не пустое, снова выбираем один элемент и нумеруем его наименьшим из неиспользованных ординалов.

Точнее, пусть $c : \mathcal{P}(X) \rightarrow X$ есть функция выбора, и $c(\emptyset) = X$. Тогда построим трансфинитной рекурсей функцию

$$f(\alpha) = c(X \setminus \{f(\beta) \mid \beta < \alpha\}).$$

Нетрудно показать, что до некоторого ординала γ эта функция является инъекцией, а на ординале γ — биекцией в множество X . А $f(\gamma)$ и всех последующих ординалов равно X . Таким образом, γ индуцирует вполне упорядочение на X .

II. Покажем, что из теоремы Цермело следует лемма Цорна. Пусть $(X, <)$ — непустое ч.у.м., в котором каждая цепь ограничена. Под **цепью** понимается подмножество $Y \subseteq X$, такое, что $(Y, <)$ — л.у.м.

Введем на X вполне упорядочение \prec в соответствии с теоремой Цермело. Построим в X цепь рекурсивно, выбирая на каждом шаге несобственную верхнюю грань уже построенного участка цепи с помощью порядка \prec .

$$f(\alpha) = \min_{\prec} \{a \in X \mid \forall \beta < \alpha : f(\beta) < a\},$$

если это множество непустое, и $f(\alpha) = X$ в противном случае. Так, $f(0) = \min_{\prec} X$. Нетрудно показать, что до некоторого ординала γ эта функция является инъекцией, а на ординале γ — биекцией в некоторую цепь $C \subseteq X$. А $f(\gamma)$ и всех последующих ординалов равно X . Таким образом, в X выстраивается цепь C , у которой нет несобственной верхней грани (такая цепь называется сквозной). А поскольку в X все цепи ограничены, то существует $\max_{\prec} C$, который и является максимальным элементом X .

III. Покажем, что из леммы Цорна следует принцип максимума Хаусдорфа. Пусть $(X, <)$ — непустое ч.у.м. Очевидно, в нем есть хотя бы одно одноточечное л.у.м. Тогда множество всех линейно упорядоченных подмножеств X непусто. Обозначим его $L(X)$ и рассмотрим его как ч.у.м. с отношением вложения \subseteq .

Пусть C — какая-нибудь цепь в $L(X)$. **Пределом** цепи (в случае отношения порядка \subseteq) называется $\cup C$. Легко проверить, что $\cup C$ также является линейно упорядоченным (по $<$) подмножеством X , т.е. $\cup C \in L(X)$. Кроме того, $\cup C$ ограничивает сверху цепь C .

Таким образом, в $L(X)$ любая цепь ограничена, а значит, по лемме Цорна, существует максимальный элемент $L \in L(X)$. А это и есть максимальное по вложению линейно упорядоченное подмножество X .

IV. Покажем, что из принципа максимума Хаусдорфа следует аксиома выбора. Пусть X — непустое множество с непустым элементом y (для удобства можно считать вообще, что $\emptyset \notin X$). Тогда на множестве $\{y\} \subseteq X$ существует функция выбора $f_{\{y\}}$, которая ставит в соответствие точке y какой-то его элемент.¹²

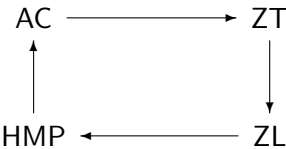
Таким образом, мы можем рассмотреть множество всех функций выбора, определенных на подмножествах X , и оно заведомо будет непустым. Обозначим его $F(X)$ и рассмотрим его как ч.у.м. с отношением вложения \subseteq . Вложение функций означает, что одна (бóльшая) продолжает вторую, а на пересечении их областей определения они совпадают.

¹²Заметим, что существование функции выбора для конечных множеств выводится из аксиом ZF, поскольку мы можем конечное число раз записать квантор существования для элемента непустого множества и, тем самым, явно вывести существование функции выбора. Для бесконечных множеств конечной формулой, увы, не отделаешься.

Тогда из принципа максимума Хаусдорфа следует, что в $F(X)$ существует максимальное по вложению л.у.м. (линейный порядок тот же — вложение функций), которое мы обозначим \mathcal{F} . В этом множестве функций выбора каждая функция содержит в себе все меньшие ее, а значит, нетрудно показать, что предел цепи \mathcal{F} , который мы обозначим $f = \cup \mathcal{F}$, является а) функцией, б) функцией выбора на подмножестве X и в) функцией выбора на всем X .

Последнее следует из того, что если бы f не была определена в какой-то точке X , то можно было бы доопределить ее до более широкой (хотя бы на одну точку шире) функции выбора, а это противоречит максимальности \mathcal{F} .

Итак, мы показали следующие импликации:



Отсюда нетрудно понять, что все четыре утверждения эквивалентны.

Как видим, аксиома выбора дает достаточно мощный инструмент теории множеств, прежде всего, связанный с тем, что любое множество можно вполне упорядочить. В то же время, надо понимать, что эта аксиома неконструктивна — она не дает алгоритм построения такого упорядочения. И, как будет видно из дальнейшего, такой алгоритм в общем случае невозможен.

В каком-то смысле можно сказать, что аксиома выбора — это тот водораздел в логике, который отделяет конструктивизм от чистой математики. Конструктивизм требует от нас явного (предикативного, рекурсивного) построения объектов, существование которых мы заявляем. В чистой же математике нам достаточно доказать формулу $\exists x$ такой, что...

Насколько безопасно ее включение в теорию ZF? Ведь если AC противоречит ZF, то вместе с полезными свойствами из нее можно вывести вообще все что угодно (и ниже мы увидим ряд странностей, которые можно извлечь из этой аксиомы).

С другой стороны, если бы функцию выбора можно было построить по алгоритму, то это означало бы ее формализацию и докузуемость в языке ZF. Поэтому, если уж она не противоречит ZF, то она должна быть независимой от ZF. Именно это и доказывают следующие утверждения матлогики.

Метатеорема 4.9 (Курт Гёдель, 1936). *Невозможно опровергнуть аксиому выбора:*



Курт Гёдель

$$\text{Con}(\text{ZF}) \rightarrow \text{Con}(\text{ZF} + \text{AC})$$

Метатеорема 4.10 (Поль Коэн, 1963). *Невозможно доказать аксиому выбора:*

$$\text{Con}(\text{ZF}) \rightarrow \text{Con}(\text{ZF} + \neg \text{AC})$$

Таким образом, в предположении совместности ZF мы получаем пример независимого от ZF утверждения. Теорию ZF+AC принято обозначать ZFC.

Интересно, что первая теорема (о непротиворечивости ZFC) доказана Гёделем при помощи построения модели конструктивных множеств [24] внутри ZF. Гёдель назвал *конструктивными* такие множества, которые могут быть построены как результат трансфинитной последовательности предикативных определений множеств. Его открытие состоит в том, что конструктивные множества (их универсум) образуют модель для аксиом ZF и AC, а также для аксиомы GCH.¹³

Более точно, контруктивность задается следующим образом. Для начала положим, что обозначение X' для любого множества обозначает

$$X' \equiv X \cup \{\text{dom}(\varphi^X) \mid \varphi^X : X \rightarrow \{0, 1\}\}, \quad (4.1)$$

при этом φ^X является не любой функцией на X , а такой, которая получается из произвольной формулы φ языка ZF путем ее релятивизации на множестве X . Иначе говоря, в X мы определяем предикаты (булево-значные функции) тем же способом, каким мы задаем формулы языка ZF, затем берем их область истинности (например, по первой свободной переменной, считая остальные переменные произвольными параметрами, принадлежащими X) в качестве новоиспеченных элементов множества X' , и присовокупляем к ним сами элементы X , поскольку они участвуют в построении предиката φ^X . Таким образом, $X' \subseteq X \cup \mathcal{P}(X)$.

Строго говоря, определение (4.1) выходит за рамки языка ZF, оно больше претендует на определение из языка второго порядка, допуская перебор по всем формулам ZF. Однако, можно показать, что такое определение формализуется в ZF [24], поскольку оно формализуется в теории с классами Гёделя–Бернайса.

Далее X' рассматривается как рекурсивный генератор для построения универсумов конструктивных множеств (точно так же, как $\mathcal{P}(X)$ мы использовали как рекурсивный генератор трансфинитной последовательности обычных универсумов), а именно:

$$\begin{aligned} K_0 &= \emptyset \\ K_\alpha &= (\cup \{K_\beta \mid \beta < \alpha\})', \end{aligned}$$

где α — положительный ординал. В итоге мы получаем возрастающую трансфинитную последовательность конструктивных универсумов.

¹³Обобщенная континуум-гипотеза. К ней мы еще вернемся ниже.

Множество x назовем **конструктивным**, если оно принадлежит какому-то K_α . На этом определении сторится т.н. **аксиома конструктивности**: *всякое множество конструктивно*. Если бы мы пользовались теорией Гёделя–Бернайса (с классами), то аксиома конструктивности приняла бы вид: класс всех множеств совпадает с классом всех конструктивных множеств.

И вот тут кроется самое интересное. Если мы теперь релятивизируем все аксиомы ZF в класс конструктивных множеств (т. е. просто будем считать, что все переменные в этих аксиомах обозначают только конструктивные множества), полученные предикаты окажутся выводимыми в ZF. Интуитивно это означает, что класс конструктивных множеств является моделью ZF. В строгом смысле это не так, поскольку модель должна быть множеством, а класс конструктивных множеств не есть множество. Поэтому ограничиваемся такой мягкой теоремой матлогики: все релятивизованные в конструктивных переменных аксиомы ZF выводимы в ZF.

Больше того, если релятивизировать таким же образом построение обычных универсумов, то окажется, что мы построим все конструктивные универсумы. В классах это выражается так: утверждение о равенстве класса множеств и класса всех конструктивных множеств, будучи релятивизованным в конструктивных множествах, доказуемо в ZF.

Наконец, в предположении аксиомы конструктивности доказывается аксиома выбора. А это и означает, что мы нашли такую модель (определяемую аксиомой конструктивности), где выполняются все аксиомы ZFC. То есть мы показали $\text{Con}(\text{ZF}) \rightarrow \text{Con}(\text{ZFC})$.

Каким же образом доказывается, что аксиома выбора выводится из аксиомы конструктивности?

На самом деле, нам здесь достаточно показать, что если X вполне упорядочено, то можно вполне упорядочить и X' . Но дело в том, что довесок к X в множестве X' — это элементы y , заданные формулами вида $\varphi^X(z, t_1, \dots, t_k)$, где z пробегает область истинности φ^X и определяет содержимое множества y , а t_1, \dots, t_k — параметры, принадлежащие X .

Далее, все формулы φ^X можно занумеровать натуральными числами (как это сделать строго, мы здесь опустим). Стало быть, все наши множества y определяются параметром n (счетчик формул) и параметрами t_1, \dots, t_k , где k пробегает натуральный ряд, а t_i пробегает вполне упорядоченное множество X .

Короче говоря, множество $X' \setminus X$ параметризуется множеством

$$\bigcup_{k < \omega} X^k \times \omega,$$

которое, в свою очередь ставится во взаимно однозначное соответствие с ординалом

$$\omega + \alpha\omega + \alpha^2\omega + \dots + \alpha^k\omega + \dots$$

(сумма счетная по ω), где α — порядковый тип X . Но из арифметики ординалов (а также и «в лоб») получается, что данная сумма есть вполне упорядоченное множество, индуцирующее порядок на $X' \setminus X$.¹⁴

Итак, X' может быть вполне упорядочено, если X вполне упорядочено.

Завершает доказательство индукция по номерам конструктивных универсумов, показывающая, что каждый конструктивный универсум может быть вполне упорядочен, причем некоторым эффективным способом. А отсюда уже строится вполне упорядочение всего класса конструктивных множеств. Тем самым мы доказываем теорему Цермело в конструктивных множествах. А она, как мы видели выше, эквивалентна аксиоме выбора.¹⁵

Как уже отмечалось, класс конструктивных множеств не есть множество, и поэтому не подпадает под наше определение модели, данное в разделе 4.1.3, однако известно, что теория с классами (Гёделя–Бернайса) эквивалентна ZF, а в ней класс конструктивных множеств является объектом теории и может рассматриваться как модель. Переход от теории с классами к обычной ZF состоит в том, что мы подменяем утверждение о принадлежности множества классу некоторым предикатом, и в этом смысле мы можем переносить все рассуждения о моделях из одной теории в другую и обратно, не теряя их смысла и истинности.

Приведенные нами рассуждения дают лишь краткий обзор выдающегося результата Гёделя, не претендуя на логическую точность, но демонстрируя сам метод доказательства такого рода метатеорем. Подробности можно найти в [14, 18, 24].

Теорему Коэна о независимости AC от аксиом ZF мы не будем здесь раскрывать ввиду чрезвычайной (по мнению автора) сложности доказательства методом вынуждения (форсинга). Заметим только, что и она напрямую связана с доказательством независимости континуум–гипотезы. Точнее, поскольку Гёдель доказал, что из аксиомы конструктивности следует GCH, а из GCH, как известно, следует AC (см. ниже), то если у нас имеется модель, в которой AC ложно, то в ней же будет ложным и GCH. Кроме того, Коэн отдельно показал, что и простая CH ложна в некоторой модели. Таким образом, аксиомы AC, CH и GCH независимы от ZF.

Теперь пара слов о континуум–гипотезе.

Мы знаем, что множества ω и $\mathcal{P}(\omega)$ неравномощны (теорема Кантора). Поэтому возникает вопрос: а насколько далеко друг от друга отстоят эти мощности. Можно ли найти такое множество, которое будет мощнее ω , но при этом менее мощное, чем $\mathcal{P}(\omega)$?

¹⁴Нужно еще заметить, что иногда разные формулы могут задавать одинаковые множества y , в этом случае достаточно присваивать множеству y первый из ординальных номеров порождающих его формул.

¹⁵На самом деле, мы доказываем здесь даже более сильное утверждение — возможность вполне упорядочить весь класс множеств.

Ответ на этот вопрос пытался найти еще сам Кантор, а Гильберт вынес его под номером 1 в списке своих знаменитых проблем математики на II Международном конгрессе математиков в Париже 1900 года (см., например, [6]).

В алеф-нумерации кардиналов континуум-гипотеза формулируется очень просто: $\aleph_1 = 2^{\aleph_0}$ и обозначается СН.

Обобщенная континуум-гипотеза (алеф-версия): $\aleph_{\alpha+1} = 2^{\aleph_\alpha}$. Обозначается GCH.

Если же мы хотим сформулировать GCH вне зависимости от существования кардинальных чисел нужного размера, то нам потребуется дать ряд определений: для множеств A и B положим

- $A \ll B$, если существует инъекция $f : A \rightarrow B$,
- $A \leftrightarrow B$, если $A \ll B$ и $B \ll A$,

последнее по теореме Кантора-Бернштейна-Шредера означает, что A и B равномощны, т. е. между ними существует биекция.

Обобщенная континуум-гипотеза (основная версия): если $A \ll B \ll \mathcal{P}(A)$, то либо $B \leftrightarrow A$, либо $B \leftrightarrow \mathcal{P}(A)$.

Соответственно, основная континуум-гипотезы (в формулировке Кантора): если $\omega \ll B \ll \mathcal{P}(\omega)$, то либо $B \leftrightarrow \omega$, либо $B \leftrightarrow \mathcal{P}(\omega)$.

Основная версия формулировки континуум-гипотезы хороша тем, что не требует существования кардинала мощности континуума, т. е. $\|\mathcal{P}(\omega)\|$, обычно обозначаемого \mathfrak{c} , тем самым не требуя и существования его полного упорядочения. Эта разница между двумя формулировками оказывается существенной, если мы рассматриваем так называемую аксиому детерминированности, о которой речь пойдет ниже.

Найденное П. Козеном доказательство того факта, что из обобщенной континуум-гипотезы следует аксиома выбора, мы приведем здесь полностью, т. к. эта теорема редко встречается в литературе.

Теорема 4.1. $\text{GCH} \rightarrow \text{AC}$.

Доказательство. Для начала заметим, что для любых множеств X и W

$$(X \cup W \leftrightarrow \mathcal{P}(2X)) \rightarrow (\mathcal{P}(X) \ll W). \quad (4.2)$$

Здесь символом $2X$ мы обозначили дизъюнктивную сумму $X_1 \cup X_2$ двух экземпляров X , например, можно считать, что $2X = X \cup (X \times \{X\})$.

Возьмем биекцию $f : X \cup W \rightarrow \mathcal{P}(2X)$ и биекцию $g : \mathcal{P}(2X) \rightarrow \mathcal{P}(X_1) \times \mathcal{P}(X_2)$. Ясно, что образ $(g \circ f)X$ слабее по мощности, чем $\mathcal{P}(X)$, поэтому и его проекция $\text{Pr}_1(g \circ f)X$ не может покрывать весь $\mathcal{P}(X_1)$.

Упражнение 4.6.
Проверьте, что $\mathcal{P}(2X)$ равномощно $\mathcal{P}(X) \times \mathcal{P}(X)$.

Пусть теперь $c \in \mathcal{P}(X_1) \setminus \text{Pr1}(g \circ f)X$. Тогда множество $\{c\} \times \mathcal{P}(X_2)$ целиком находится в образе $(g \circ f)W$. А поскольку $g \circ f$ биективно, мы получаем, что $\mathcal{P}(X) \ll W$.

Теперь, пусть A — некоторое множество, обозначим:

$$B \equiv \mathcal{P}(A \cup (\omega \times \{A\})),$$

т. е. вместо исходного A мы рассматриваем степень множества со счетным вполне упорядоченным «хвостом». Умножение на синглетон $\{A\}$ производится исключительно для того, чтобы сумма с A была дизъюнктивной. Для такого множества легко видеть, что

$$2B \leftrightarrow \mathcal{P}(A \cup ((\omega + 1) \times \{A\})) \leftrightarrow \mathcal{P}(A \cup (\omega \times \{A\})),$$

последнее вытекает из того, что $\omega + 1 \leftrightarrow \omega$, что легко проверить непосредственно. Таким образом, $2B \leftrightarrow B$.

Далее, $B \ll B \cup \{B\} \ll 2B$, так что $B \cup \{B\} \leftrightarrow B$. Отсюда следует, что $\mathcal{P}(B) \leftrightarrow \mathcal{P}(B \cup \{B\}) \leftrightarrow 2\mathcal{P}(B)$. Аналогично рассуждая, приходим к соотношению

$$2\mathcal{P}^k(B) \leftrightarrow \mathcal{P}^k(B), \quad (4.3)$$

где $k = 0, 1, 2, \dots$

Наконец, пусть

$$R \equiv \{< \mid X \subseteq B, (X, <) \text{ — вполне упорядоченное множество}\}$$

То есть, R содержит все полные упорядочения всех подмножеств B , какие только существуют. Очевидно, что R не пусто, т. к. в нем есть хотя бы упорядочения конечных подмножеств B .

Напомним, что отношение есть совокупность упорядоченных пар, поэтому нетрудно получить, что $R \subseteq \mathcal{P}^3(B)$.

Пусть теперь $r \sim q$ для $r, q \in R$, если они изоморфны, т. е. их порядковый тип — один и тот же ординал. Ясно, что отношение \sim является отношением эквивалентности, поэтому мы можем рассмотреть фактор-множество $W = R/\sim$.

Ясно также, что $W \subseteq \mathcal{P}^4(B)$. Кроме того, нетрудно показать, что W вполне упорядочен тем порядком, который в него индуцируется с ординалов, поскольку элементы W взаимно однозначно соответствуют ординалам. Кроме того, порядковый тип $|W|$ не может быть задан ни для какого подмножества $X \subseteq B$, т. к. иначе мы бы получили, что $|W| < |W|$.

В итоге мы имеем:

$$\neg(W \ll B) \quad (4.4)$$

$$W \ll \mathcal{P}^4(B) \quad (4.5)$$

Далее наши рассуждения строятся следующим образом. Предположим, что

$$\mathcal{P}^k(B) \ll \mathcal{P}^k(B) \cup W \ll \mathcal{P}^{k+1}(B) \leftrightarrow \mathcal{P}(2\mathcal{P}^k(B)),$$

последнее вытекает из (4.3). Собственно, левое неравенство всегда верно, а вот правое мы предполагаем индуктивно, отправляясь от установленного случая при $k = 3$. Тогда в силу обобщенной континуум-гипотезы у нас есть два варианта: либо равенство мощностей достигается справа, либо слева.

В случае, если оно достигается справа, в силу (4.2) мы получаем $W \leftrightarrow \mathcal{P}^{k+1}(B)$ (при $X = \mathcal{P}^k(B)$), что при $k = 0, 1, 2, 3$ означает возможность вполне упорядочить само множество B (т. к. оно естественным образом инъективно вкладывается во все свои степени), а значит, и исходное множество A , т. к. и оно вкладывается естественным образом в B .

В случае, если оно достигается слева, это означает, что $W \ll \mathcal{P}^k(B)$, т. е. мы получаем неравенство (4.5), только с пониженной степенью. Здесь мы снова встречаем **метод бесконечного спуска**, известный нам еще по теореме Ферма. Правда, в данном случае спуск состоит всего из 3 шагов, на каждом из которых в случае равенства справа мы будем получать возможность вполне упорядочить B .

Спуск стартует со значения $k = 3$, при котором правое неравенство нами уже установлено в (4.5).

Слева же мы в конце концов упрямся в равенство $B \leftrightarrow B \cup W$ (при $k = 0$), но это невозможно в силу (4.4).

Таким образом, при всех допустимых k мы получаем возможность вполне упорядочить A . Это значит, что из обобщенной континуум-гипотезы следует теорема Цермело, которая эквивалентна аксиоме выбора. \square

Отметим, что для нужд анализа, как правило, вполне достаточно так называемой *счетной аксиомы выбора*, обозначаемой AC_ω и утверждающей, что для всякого счетного множества существует функция выбора, либо чуть более сильной аксиомы зависимого выбора:

$$DC: \forall a : \forall I : a \rightarrow \mathcal{P}(a) \exists f : \omega \rightarrow a : \forall n < \omega : f(n+1) \in I(f(n)).$$

Эта аксиома чем-то напоминает рекурсивное определение, поскольку каждое следующее значение функции выбора f ограничено множеством, заданным ее предыдущим значением.

Упражнение 4.7. | Несложно показать, что аксиома зависимого выбора влечет аксиому счетного выбора, а сама следует из общей аксиомы выбора:

$$AC \rightarrow DC \rightarrow AC_\omega.$$

Для того, чтобы выделить теоремы, использующие какие-либо дополнительные аксиомы, помимо аксиом ZF, мы будем явно указывать обозначение этих дополнительных аксиом в наименовании теорем.

4.2.1 Полезные следствия аксиомы выбора

Для начала вернем долг по отсылкам к аксиоме выбора в наших предыдущих выкладках.

Теорема 4.2 (АС). *В предположении общей аксиомы выбора справедливы следующие утверждения:*

- 1) $\prod_{\lambda \in \Lambda} X_\lambda$ не пусто, если не пусты все X_λ ;
- 2) для каждого кардинала существует последующий кардинал (следствие: не существует наибольшего кардинала);
- 3) каждое множество имеет мощность–кардинал;
- 4) квадрат бесконечного множества равномощен этому множеству;
- 5) в любом ненулевом модуле над кольцом существует базис Гамеля;
- 6) все базисы Гамеля в модуле над коммутативным кольцом равномощны;
- 7) (теорема Тарского) каждый фильтр можно вложить в ультрафильтр на том же множестве.

Заметим, что свойство 4), именуемое также теоремой о квадрате, легко выводится из свойства умножения кардиналов $CW9$, т. к. АС гарантирует нам взаимно однозначное соответствие между произвольным множеством и некоторым кардиналом.

Существование базиса Гамеля напрямую выводится из леммы Цорна. Действительно, если модуль ненулевой, то в нем есть хотя бы один независимый набор векторов (например, из одного элемента). Кроме того, предел цепи независимых наборов также является независимым набором. Стало быть, независимые наборы образуют ч.у.м. с ограниченными цепями, и по лемме Цорна, существует максимальный независимый набор, т. е. базис Гамеля.

Равномощность базисов мы уже показывали ранее в соответствующей врезке.

Теорема Тарского об ультрафильтрах также следует из леммы Цорна.

Приведем без доказательства еще ряд полезных и достаточно простых следствий разных вариантов аксиомы выбора. Предлагаем читателю доказать следующие шесть теорем самостоятельно.

Упражнение
4.8.

Теорема 4.3 (АС_ω). *Объединение не более чем счетного множества не более чем счетных множеств не более чем счетно.*

Аксиома AC_ω здесь нужна лишь для того, чтобы для каждого из счетных множеств зафиксировать какую-то их нумерацию натуральными числами (это нужно сделать счетное число раз), после чего можно применять канторовскую диагональную нумерацию.

Теорема 4.4 (AC_ω). *Ко всякой предельной точке множества действительных чисел сходится последовательность, лежащая в этом множестве.*

Теорема 4.5 (AC_ω). *Всякое бесконечное множество содержит счетное подмножество.*

Теорема 4.6 (AC_ω). *Мера Лебега счетно-аддитивна, т. е. мера счетной дизъюнктивной суммы измеримых по Лебегу множеств равна сумме их мер.*

Здесь требуется искусно построить счетные покрытия так, чтобы их суммарная мера мало (на произвольное $\varepsilon > 0$) отличалась от суммы мер исходных множеств (достаточно работать только с множествами меры ноль).

Предполагая знакомство читателя с основами математического анализа, напомним, что **множествами первой категории** (Бэра) называются счетные объединения нигде не плотных множеств.¹⁶

Теорема 4.7 (AC_ω). *Объединение счетного множества множеств первой категории является множеством первой категории.*

Отсюда, в частности, следует, что \mathbb{R} не является множеством первой категории. Однако, этот факт можно доказать и без участия счетной формы аксиомы выбора, оперируя только интервалами с рациональными границами (напомним, что рациональные числа мы умеем эффективно нумеровать натуральными, предъявив явную алгебраическую формулу такой нумерации).

Это лишний раз показывает, что некоторые следствия неконструктивных аксиом на самом деле могут быть получены конструктивными методами.

Теорема 4.8 (DC). *Множество вполне упорядочено тогда и только тогда, когда в нем не существует счетной строго убывающей последовательности.*

Здесь стоит отметить, что доказательство необходимости не требует участия аксиомы выбора. Если множество вполне упорядочено, то в нем не существует бесконечно убывающей последовательности просто по определению. И этот факт использовался нами при доказательстве теоремы Гудстейна и теоремы Ферма при $n = 3$ и $n = 4$ (метод бесконечного спуска). А вот достаточность требует построения некоторой счетной рекурсии на основе аксиомы DC, чтобы придти к противоречию.

¹⁶Множество A нигде не плотно в \mathbb{R} , если для любого непустого интервала $(a; b)$ найдется непустой интервал $(c; d)$ такой, что $(c; d) \subseteq (a; b) \setminus A$.

Последними белыми шарами на чашу весов полезности АС положим следующие две теоремы, позволяющие связать частичный и линейный порядки произвольных множеств.

Теорема 4.9 (Шпильрайн, АС). *Любой частичный порядок может быть продолжен до линейного.*

Данная теорема была получена польским математиком Эдвардом Шпильрайном в 1930-м году. Ее доказательство основано на применении леммы Цорна. Позднее (1941) Бен Душник и Б. У. Миллер доказали еще одно утверждение о порядках.

Теорема 4.10 (Душник, Миллер, АС). *Каждое отношение частичного порядка является пересечением содержащих его отношений линейного порядка.*

Существует также понятие размерности частичного порядка, введенное Душником и Миллером в статье [31]. Размерность Душника–Миллера есть наименьшее количество (в кардинальном смысле) всех линейных порядков, пересечение которых совпадает с данным частичным порядком.

Комментарий 15.

Интересно, что алгоритмическая задача распознавания того, превосходит ли размерность данного *конечного* частичного порядка заданное число k , принадлежит классу Р при $k < 3$, но является NP-полной при $k \geq 3$.

4.2.2 Странные следствия аксиомы выбора

Перейдем теперь к черным шарам на весах полезности аксиомы выбора.

Прежде всего заметим, что базис Гамеля в \mathbb{R} как модуле над кольцом рациональных чисел позволяет построить очень неприятный пример. Напомним, что аддитивная функция — та, которая сохраняет сложение, т. е. $f(a + b) = f(a) + f(b)$. Отсюда, например, следует, что $f(p/q) = f(1)p/q$, т. е. в рациональных точках наша функция принимает вид линейной: $f(x) = kx$. Беда в том что в иррациональных точках она может быть совсем другой.

Теорема 4.11 (АС). *Существует вещественная аддитивная не непрерывная функция.*

Доказательство. Возьмем какой-нибудь базис Гамеля \mathbb{R} над кольцом \mathbb{Q} , обозначив его S . Тогда для любого $x \in \mathbb{R}$ имеем единственное разложение $x = \sum_{k=1}^n r_k s_k$, где $s_k \in S$, $r_k \in \mathbb{Q} \setminus \{0\}$, а выбор n , s_k и r_k однозначно определя-

ется числом x .¹⁷ Пусть также $\hat{s} \in S$ — некоторое выделенное число из базиса. Число \hat{s} может входить в набор s_k для каждого конкретного x , а может не входить. Тогда положим

$$f(x) = \begin{cases} r_k, & \text{если } \exists k \ s_k = \hat{s}, \\ 0, & \text{иначе} \end{cases}$$

Упражнение 4.9. Легко проверить, что f аддитивна, но в то же время не является непрерывной, поскольку принимает только рациональные значения (различные). \square

Отметим, что для доказательства этой теоремы требуется общая аксиома выбора, а точнее, нужна функция выбора на $\mathcal{P}^2(\mathbb{R})$, которое равномощно $\mathcal{P}^3(\omega)$ (это показывает анализ доказательства леммы Цорна), т. е. построить данный контрпример таким путем в рамках только AC_ω не получится. Здесь мы уже начинаем подмечать ту особенную роль, которую играет счетная аксиома выбора AC_ω , доставляя нам, с одной стороны, необходимый инструментарий для мат.анализа теоремами 4.3–4.7, а с другой стороны, не доставляя неудобный контрпример. В дальнейшем мы увидим еще несколько аргументов в пользу ограничения области действия аксиомы выбора.

Теперь построим известный пример Витали неизмеримого множества.

Теорема 4.12 (АС).

- (1) Существует неизмеримое по Лебегу множество в \mathbb{R} .
- (2) Существует множество, не обладающее свойством Бэра.

Доказательство. Для чисел $x, y \in \mathbb{R}$ положим $x \sim y$, если $x - y \in \mathbb{Q}$. Это — отношение эквивалентности, так что мы можем рассмотреть фактор-множество \mathbb{R}/\sim , элементами которого будут счетные множества, откуда следует, что фактор \mathbb{R}/\sim не счетен (по крайней мере, в предположении АС).

Выберем теперь по одной точке из каждого класса эквивалентности (здесь потребуются несчетная форма аксиомы выбора), причем только такие, которые лежат в $[0; 1]$. Полученное множество обозначим M . Если теперь сдвигать M на произвольное рациональное число $r \in [-1; 1]$, то мы будем получать всякий раз новое множество, не пересекающееся с другими аналогичными множествами (иначе мы бы имели как минимум 2 эквивалентных точки в M).

Взяв сумму всех таких множеств (а это счетная дизъюнктивная сумма), мы получим множество A , мера которого ограничена сверху мерой отрезка $[-1; 2]$, а снизу — мерой отрезка $[0; 1]$. Но тогда, если мера M равна нулю, это противоречит счетной аддитивности меры и тому, что мера A не меньше

¹⁷Отметим, что нумерация s_k элементов базиса S не «сквозная», она возникает всякий раз заново при выборе нового x .

1, а если мера M положительна, то это противоречит счетной аддитивности меры и тому, что мера A не больше 3. Противоречие.

Следовательно, M неизмеримо по Лебегу.

Это же множество доставляет и пример множества, не обладающего свойством Бэра. \square

Или имеет бесконечно малую меру? :)

Отметим, что и здесь нам потребовалась несчетная форма аксиомы выбора, что дает еще +1 очко в пользу AC_ω .

Чтобы построить следующий замечательный пример, определим свободное произведение групп. В отличие от внешнего прямого произведения, определенного равенством (3.2), свободное произведение конечных групп порождает счетную группу. Пусть G и H — группы, тогда все возможные несократимые записи вида $g_1 g_2 \dots g_n$, где $g_k \in G \cup H$, составляют группу $G * H$, которая и называется **свободным произведением групп** G и H . Под несократимостью понимается невозможность найти отрезок $g_k g_{k+1} \dots g_{k+j}$, в котором все элементы принадлежат одной группе и при умножении в указанном порядке дают элемент e соответствующей группы. Умножение элементов $G * H$ — это приписывание второго сомножителя к первому справа, например, $aabbb \circ cccdf f f = aabbbcccd f f f$, с последующим сокращением отрезков, равных единице (если $bc = e$, то указанное произведение сократится до $aadf f f$). Единичные элементы обеих групп являются единицами $G * H$.

Возьмем в качестве группы G подгруппу группы вращений мнимой сферы S^2 , состоящую из поворотов на 180° вокруг некоторой оси, например, это может быть группа, порожденная преобразованием $R_i(h) = -ihi$, где i — мнимая единица, h — кватернион единичной длины, пробегающий поверхность мнимой сферы (см. раздел 3.6.2). R_i реализует поворот на 180° вокруг оси i . Ясно, что группа $G = \langle R_i \rangle$ является циклической группой второго порядка, изоморфной \mathbb{Z}_2 .

В качестве группы H возьмем еще одну подгруппу группы вращений сферы, реализующую повороты на 120° вокруг другой оси. Например, преобразование $R_{\frac{1+\bar{\xi}\sqrt{3}}{2}}$ реализует поворот вокруг оси $\bar{\xi}$ на данный угол согласно формуле (3.16), где $\bar{\xi} = ai + bi + ck$ — мнимый вектор единичной длины. Ясно, что группа $H = \langle R_{\frac{1+\bar{\xi}\sqrt{3}}{2}} \rangle$ является циклической группой третьего порядка, изоморфной \mathbb{Z}_3 .

Однако, вторую ось $\bar{\xi}$ нам следует выбрать так, чтобы группа $G * H$ получилась действительно свободным произведением. Например, в случае выбора $\bar{\xi} = j$ в качестве второй оси мы получаем зависимость между элементами разных групп, поскольку уже комбинация $R_i \circ R_{\frac{1+j\sqrt{3}}{2}} \circ R_i \circ R_{\frac{1+j\sqrt{3}}{2}}$ дает тождественное преобразование.

Тем не менее, легко показать, что поскольку всех возможных элементов в группе $G * H$ лишь счетный набор, а различных углов между осями i и $\bar{\xi}$ — континуум, обязательно найдется такая ось вращения $\bar{\xi}$, что произведение

$G * H$ будет доставлять только нетривиальные (кроме единичного) и попарно различные вращения. Очевидно, что порождающий группу H кватернион $\bar{\xi}$ должен иметь хотя бы один трансцендентный коэффициент при одной из мнимых единиц.

Обозначим $p = R_i$, $q = R_{\frac{1+\bar{\xi}\sqrt{3}}{2}}$. При этом мы знаем, что $p^2 = 1$ и $q^3 = 1$ (в силу выбора углов). Тогда все элементы $G * H$ описываются последовательностями следующих четырех видов:

$$pq^{\pm}pq^{\pm}\dots q^{\pm}p, \quad q^{\pm}pq^{\pm}\dots q^{\pm}p, \quad pq^{\pm}pq^{\pm}\dots pq^{\pm}, \quad q^{\pm}pq^{\pm}p\dots pq^{\pm}.$$

Разобьем группу $G * H$ на три непересекающихся подмножества A, B, C таких, что в A входят элементы, запись которых начинается с q , в B — начинающиеся на q^{-1} , в C — начинающиеся с p , а также элемент e . Ясно, что $A = qC$, $B = q^{-1}C$, $A \cup B \cup \{e, p\} = pC$.

При каждом вращении сферы (кроме тождественного) неподвижными остаются только 2 точки, лежащие на оси вращения, поэтому все нетривиальные вращения группы $G * H$ могут оставлять на месте лишь счетный набор точек сферы. Обозначим через Q множество всех этих точек. Множество Q счетно.

Рассмотрим теперь множество M , содержащее ровно по одной точке с каждой орбиты действия группы $G * H$ на «колотовой» сфере $S^2 \setminus Q$.¹⁸ Существование такого множества M гарантирует аксиома выбора в несчетной форме.

Далее применим такой же подход, как в примере Витали. Сдвинем множество M вдоль орбит группы $G * H$ ее подмножествами A, B, C , получим:

$$X = AM, \quad Y = BM, \quad Z = CM,$$

полученные множества попарно не пересекаются, т. к. разные элементы $G * H$ дают разные значения на всех элементах M .

Используя определение A, B, C , имеем:

$$X = qZ, \quad Y = q^{-1}Z, \quad X \cup Y \cup M \cup pM = pZ.$$

Теперь, поскольку p, q, q^{-1} — изометрии, то множества X, Y, Z попарно конгруэнтны (переходят друг в друга вращением сферы) друг другу, а множество $X \cup Y$ конгруэнтно некоторой части Z . Иначе говоря, с помощью движений сферы мы вкладываем в Z сумму двух множеств, конгруэнтных исходному Z . При этом мы видим, что все множества X, Y, Z, Q попарно не пересекаются и в сумме дают всю сферу S^2 .

Чуть сложнее доказывается более точное утверждение, опубликованное Хаусдорфом в 1914 году.

¹⁸Напомним, что орбитой действия группы называется класс эквивалентности по отношению \sim , где $x \sim y$, если $x = g(y)$, где g — элемент группы преобразований.

Теорема 4.13 (Хаусдорф, АС). *Сферу S^2 можно разбить на четыре попарно непересекающихся множества X, Y, Z, Q таких, что $X, Y, Z, X \cup Y$ попарно конгруэнтны, а Q — счетно.*

Заметим, что и эта теорема использует несчетную форму аксиомы выбора.

Еще +1 очко в пользу $АС_\omega$?

Еще один результат (см. [36]), отчасти основанный на теореме Хаусдорфа и использующий несчетную форму аксиомы выбора:

Теорема 4.14 (Банах, Тарский, АС, 1924). *Замкнутый шар B^3 можно разбить на два непересекающихся подмножества $B^3 = X \cup Y$ таких, что B^3, X, Y попарно конгруэнтны.*

Иначе говоря, шар можно «распилить» на части так, что после некоторой их перестановки можно получить два шара того же диаметра без наложений и пустот. Естественно, эти части будут неизмеримы по Лебегу.

4.2.3 О сверхбольших кардиналах

Как мы уже видели выше, следствием аксиомы выбора являются два замечательных факта: во-первых, каждое множество имеет алеф-мощность, т. е. его можно взаимно однозначно сопоставить с некоторым (единственным) кардиналом, а во-вторых, не существует наибольшего кардинала, т. е. шкалу кардиналов можно продолжать столь же долго, сколько и шкалу ординалов. Это чем-то напоминает нумерацию простых чисел натуральными, где натуральные числа играют роль ординалов, а простые числа — кардиналов.

В таком случае довольно просто взять и написать, например, кардинал \aleph_ω . Это просто предел (в смысле вложения множеств) кардиналов \aleph_n . А в случае принятия GCH так это и вовсе предел мощностей $\|\mathcal{P}^n(\omega)\|$. При этом заметим, что в качестве индекса в алеф-нумерации мы пока задействовали только конечные числа и ω . Ничто не мешает нам (при аксиоме выбора) строить такие кардиналы, как $\aleph_{\omega+\omega}$ или \aleph_{ε_0} и т.д. Делая сравнительно небольшие шаги по шкале ординалов (пока мы находимся только среди счетных), мы уже очень далеко забрались по шкале кардиналов. Так, в случае принятия CH мы уже давно перешагнули через континуум (правда, без CH этого нельзя утверждать). Тем не менее, мы видим, что все эти большие предельные(!) кардиналы получаются как предел счетной возрастающей последовательности меньших кардиналов. О последующих кардиналах такое, конечно, сказать нельзя.

Если принять $C_2H_5OH \dots$ нет, только CH.

Напомним, что кардинал τ называется **последующим**, если $\tau = \xi^+$ при некотором кардинале ξ , т. е. между ξ и τ на шкале кардиналов нет промежуточных точек. Ненулевой кардинал называется **предельным**, если он не последующий. В алеф-нумерации последующему несчетному кардиналу соот-

ветствует последующий ординал, предельному несчетному кардиналу — предельный ординал. Кардинал ω также является предельным, но его обычно принято исключать в определениях, связанными с кардиналами.

На возможности представления кардинала как предела цепи меньших кардиналов основано понятие конфинальности. Скажем, что кардинал ξ является **конфинальным характером кардинала** τ и обозначается $\text{cf}(\tau)$, если ξ — минимальная мощность цепи кардиналов $< \tau$, дающая в пределе τ . Тначе говоря, внутри кардинала τ можно построить лестницу из $\text{cf}(\tau)$ ступенек, достающую до его вершины. Более короткую такую лестницу построить не удастся.

Бесконечный кардинал τ называется **регулярным**, если $\tau = \text{cf}(\tau)$, в противном случае он называется **сингулярным**. Легко видеть, что ω и все \aleph_n регулярны, в то время как $\aleph_\omega, \aleph_{\varepsilon_0}$ — сингулярны. Вообще, любой бесконечный последующий кардинал регулярен. Кроме того, сам кардинал $\text{cf}(\tau)$ является регулярным. Известно также следующее соотношение, выводимое в предположении существования 2^τ : $\tau < \text{cf}(2^\tau)$ для любого кардинала $\tau \geq \omega$.

В частности, $\omega < \text{cf}(\mathfrak{c})$, т. е. множество действительных чисел нельзя представить как счетное объединение множеств мощности меньше \mathfrak{c} (в предположении, что мощность \mathfrak{c} существует, т. е. \mathbb{R} можно вполне упорядочить).

Поскольку регулярность последующих кардиналов тривиальна, интересное рассмотреть случай предельного кардинала. Мы видели, что даже \aleph_{ε_0} является сингулярным, т. е. его вершины можно достичь очень более короткой лестницей — счетной. Назовем несчетный кардинал **слабо недостижимым**, если он предельный и регулярен.

Известно [28], что ZF не противоречат следующие утверждения:

- а) \mathfrak{c} — последующий кардинал;
- б) \mathfrak{c} — сингулярный кардинал;
- с) \mathfrak{c} — слабо недостижимый кардинал.

Итак, при желании даже \mathfrak{c} можно считать слабо недостижимым, при этом он будет больше любого кардинала \aleph_α , где α — не только счетный, но и любой менее чем континуальный ординал. Согласитесь, это уже довольно большое кардинальное число.¹⁹

Однако мы на этом не остановимся и спросим себя, как ведут себя мощности булеанов меньших, чем τ , кардиналов. Ведь если рассмотреть кардинал ω , то вместе с неравенством $n < \omega$ мы также имеем и $2^n < \omega$, т. е. переход к булеану оставляет нас внутри ω . На этом основано следующее определение: кардинал τ называется **доминантным** (или сильно предельным), если для

¹⁹Заметим, что если в определении слабо недостижимого кардинала снять требование несчетности, то ω окажется слабо недостижимым.

всех кардиналов $\xi < \tau$ имеем также $2^\xi < \tau$. Очевидно, ω — доминантный кардинал. Наконец, несчетный кардинал называется **сильно недостижимым**, если он предельный и доминантный.²⁰

Известно, что если σ — первый строго недостижимый кардинал, то он является σ -м в списке всех кардиналов, т. е. $\sigma = \aleph_\sigma$. Это несколько напоминает нам определение ординала $\varepsilon_0 = \omega^{\varepsilon_0}$. Если в случае ординалов ε_0 является некой точкой равновесия бесконечного возведения ω в ординальные степени, то σ является точкой равновесия шкалы алефов. Причем в самом начале функция \aleph_α растет взрывоподобно («большой взрыв»?) — первые алефы сразу и безоговорочно далеко отрываются от своих ординальных номеров. Но затем происходит насыщение, и взрыв стабилизируется. Кстати, если мы определим $\omega!$ как предел $n!$ по традиции трансфинитных определений, то и тут мы получим, что $\omega = \omega!$, т. е. мы приходим в точку равновесия, хотя рост факториала нам кажется колоссальным. Таковы странности бесконечности.

В этом месте нам следует пополнить наш список архетипов **архетипом недостижимости**. Этот архетип в различных формах принадлежит к наиболее важным концепциям культурных традиций и состоит в представлении об экстраординарно огромном феномене, который не может быть достигнут менее мощными средствами, чем сам этот феномен.²¹

Утверждение о существовании сильно недостижимого кардинала обозначается **SI** и называется аксиомой по той причине, что не установлено, зависимо ли оно или его отрицание от какой-либо теории множеств (ZF или ZFC). Известна следующая

Метатеорема 4.11. *Невозможно доказать, что невозможно опровергнуть аксиому SI:*

$$\text{Con}(\text{ZF}) \not\vdash \text{Con}(\text{ZF} + \text{SI}).$$

Первый строго недостижимый кардинал принято обозначать \aleph_0 , а последующие нумеровать аналогично алефам.

Замечательный факт, связанный с первым строго недостижимым кардиналом, показывает следующая

Метатеорема 4.12. *Если τ — первое строго недостижимое кардинальное число, то универсум V_τ является моделью теории ZFC.*

Напомним, что универсумы мы нумеровали ординалами, но каждый кардинал является также и ординалом, поэтому универсум V_τ имеет право на определение.

С точки зрения исчисления высказываний эта теорема довольно бессмысленная, т. к. она утверждает, что из совместности более широкой теории

²⁰Заметим, что если в определении сильно недостижимого кардинала снять требование несчетности, то ω окажется сильно недостижимым.

²¹Это почти дословная цитата из [19].

$ZFC + SI$ следует совместность ZFC . Но это и так ясно, поскольку противоречивость любой теории T автоматически влечет противоречивость и любой более широкой теории.

Тем не менее, это хороший пример построения полной модели теории множеств на множестве (а не на классе!), что полностью вписывается в теорию моделей.²² Кроме того, это еще и модель самой теории GB (Гёделя–Бернайса) с классами и с аксиомой выбора, причем с наиболее сильной аксиомой выбора, утверждающей, что не только на множестве, но и на всяком классе существует функция выбора. То есть непротиворечивость SI влечет непротиворечивость усиленной формы аксиомы выбора. Наконец, поскольку в V_τ нет кардинала τ , то в модели V_τ выполняется отрицание аксиомы SI . Таким образом, если SI не противоречит ZFC , то она и вовсе не зависит от ZFC , т. к. $\neg SI$ тоже не противоречит ZFC .

Упражнение | Мы предлагаем читателю самостоятельно доказать данную теорему. Для доказательства нужно просто показать, что все аксиомы ZFC выполняются, если предположить, что переменные, пробегающие множества, пробегают только элементы V_τ , т. е. нужно проверить истинность релятивизации этих аксиом внутри универсума V_τ .

Возможность построения модели ZF в универсуме с номером, равным строго недостижимому кардиналу, говорит о колоссальной величине этого кардинала относительно нижестоящих его кардиналов (то же самое, кстати, можно сказать о сравнении ω с натуральными числами). Тем не менее, существуют определения кардиналов, доставляющих (в ZFC) и еще бо́льшие величины. К таким кардиналам относятся, например, измеримые по Уламу кардиналы. Известно, что первый измеримый кардинал χ является χ -м на шкале строго недостижимых кардиналов.

Интересно отметить, что изначально придумав лестницу ординалов, потенциально безграничную в мире множеств, мы тут же расставили на ней вехи — кардинальные числа, тем самым сильно сократив подъем по этой лестнице, ведь переходя к следующему кардиналу, мы пропускаем сразу столько же ординалов, какова мощность этого следующего кардинала. Затем мы уже на лестнице кардиналов нашли новые вехи — строго недостижимые кардиналы, и научились пропускать сразу столько же кардиналов, какова мощность строго недостижимого кардинала. И вот теперь мы вводим понятие измеримого кардинала, тем самым перешагивая сразу же через столько строго недостижимых кардиналов, какова мощность измеримого кардинала. Получается своего рода фрактальная рекурсия, шаг увеличения масштаба в которой связан с введенным выше архетипом недостижимости. Получается некий переход по шкале самоподобных миров. Правда, каждый новый шаг укрупнения

²²Вспомните, что моделью теории конструктивным множеств был класс, а не множество, в связи с чем мы были вынуждены ссылаться на эквивалентность теории ZF и теории с классами Гёделя–Бернайса.

связан с необходимостью вводить новые аксиомы, безопасность присоединения которых к ZF становится все более сомнительной, особенно, если учесть работы [19, 20].

4.2.4 Конкурент АС: аксиома детерминированности

Представим себе простую игру. Мы взяли некоторое непустое подмножество A в пространстве векторов ω^n , т. е. некоторый набор цепочек натуральных чисел. Для простоты можно даже считать, что это цепочки состоящие из цифр $1, 2, \dots, 9$. Эти цифры могут олицетворять что угодно, например, некоторые комбинации положений фигур на шахматной доске (коих огромное количество, сильно больше 9, но все же конечное).

Игрок I пишет первую цифру, Игрок II — вторую, затем Игрок I — третью, Игрок II — четвертую. Игра продолжается до тех пор, пока не будут написаны n цифр. В итоге получится некоторый вектор из нашего пространства, который может либо принадлежать множеству A , либо нет (третьего не дано).

Если полученный в ходе игры вектор попал в A , то считается, что выиграл Игрок I, а если нет — Игрок II. В шахматах, правда, бывает ничья, но здесь для простоты картины мы просто считаем эту ситуацию выигрышем Игрока II. В этом случае множество A должно состоять из всех векторов, которые заканчиваются на цифры, соответствующие одному из выигрышных для Игрока I положений на доске.

Заметим, что пока мы предполагаем, что игра длится ровно n шагов. Если мы хотим избавиться от этого ограничения, нам следует в качестве базового пространства рассмотреть объединение $\bigcup_{n < \omega} \omega^n$. В этом случае мы получаем возможность закодировать сколько угодно состояний игры и сколько угодно шагов игры до ее завершения (терминального состояния).

Введем теперь понятие выигрышной стратегии. Под этим термином принято понимать наличие такой функции f от истории игры (т. е. от построенного на каждом шаге игры вектора), что значение $a_n = f(a_0, \dots, a_{n-1})$ диктует в случае четного n Игроку I такой ход, что в конце игры Игрок I выигрывает. Функция f при этом называется выигрышной стратегией Игрока I. Аналогично определяется выигрышная стратегия Игрока II.

Существование выигрышной стратегии Игрока I для игры в $2n + 1$ шагов можно выразить следующим способом:

Σ - и
П-формулы,
см. раздел
4.3.3.

$$\exists a_0 \forall a_1 \exists a_2 \dots \forall a_{2n-1} \exists a_{2n} : (a_0, \dots, a_{2n}) \in A,$$

отрицанием этого утверждения является

$$\forall a_0 \exists a_1 \forall a_2 \dots \exists a_{2n-1} \forall a_{2n} : (a_0, \dots, a_{2n}) \notin A,$$

что постулирует существование выигрышной стратегии Игрока II для игры в $2n + 1$ шагов.

Таким образом, для каждого n существуют выигрышная стратегия либо Игрока I, либо Игрока II. Мы описали полностью детерминированную игру в том смысле, что в случае ее осуществления за конечное число шагов для каждого из игроков существует стратегия выигрыша, выбор которой, правда, зависит от n , и это вновь возвращает нас к счетной форме аксиомы выбора и подозрению, что на практике не всегда такую стратегию можно указать.²³

Но мы не остановимся на этом и предположим, что игра может длиться ω шагов, так что пространством игровых историй у нас будет уже $\mathcal{N} = \omega^\omega$ (конечные игры можно условиться обозначать хвостами нулей), именуемое также **пространством Бэра**.

В этом случае наличие выигрышной стратегии Игрока I в кванторной форме потребует бесконечной формулы

$$\exists a_0 \forall a_1 \exists a_2 \dots \forall a_{2n-1} \exists a_{2n} \dots : (a_0, \dots, a_{2n}, \dots) \in A,$$

что, конечно, не допустимо в формализме ZF. Поэтому для бесконечной игры понятие выигрышной стратегии вводится функционально, как мы это указывали выше: существует функция f , определенная на конечных векторах, подписывающая игроку следующий ход.

Таким образом, для каждого множества $A \subseteq \mathcal{N}$ можно говорить об игре G_A и о наличии/отсутствии выигрышных стратегий для одного из игроков в этой игре. Если для игры G_A существует выигрышная стратегия у одного из игроков, то множество A и игра G_A называются **детерминированными**.

Стоит отметить, что если A счетно, то легко построить выигрышную стратегию для Игрока II, так что счетные подмножества \mathcal{N} заведомо детерминированы.

Множество \mathcal{N} принято называть пространством, когда на нем задана естественная топология, порожденная системой подмножеств, которые строятся как все последовательности, имеющие общее конечное начало.²⁴ Такие подмножества называются *бэровскими интервалами*. Они порождают топологию на \mathcal{N} , соответственно, мы можем говорить об открытых и замкнутых множествах в \mathcal{N} .

Существует теорема Гейла–Стьюарта, показывающая, что все открытые множества в \mathcal{N} детерминированы. Это теорема доказывается в чистой ZF. Существует и более сильный ее вариант, показывающий, что все борелевские

²³Стоит также вспомнить пример последовательности Гудстейна, где мы, заранее зная, в какое число шагов она уложится, можем доказать ее сходимости к 1, но беда в том, что заранее мы-таки ничего не знаем.

²⁴Мы уже имели дело с похожим пространством, а именно — с канторовским пространством бинарных последовательностей 2^ω , когда говорили о свойстве компактности Исчисления высказываний.

множества \mathcal{N} детерминированы, однако такая теорема доказывается уже в теории $\mathbf{ZF} + \mathbf{DC}$, т. е. включает как минимум счетную аксиому выбора (напомним, что \mathbf{AC}_ω следует из \mathbf{DC}). Этот результат интересен тем, что он использует аксиоматику $\mathbf{ZF} + \mathbf{DC}$ в полном объеме — нельзя исключить ни одну из аксиом для его получения.

С другой стороны, используя \mathbf{AC} (в общей версии) можно доказать существование недетерминированного множества A . Разумеется, оно окажется неборелевским точно также, как пример Витали и разбиение шара дают неизмеримые множества. Вообще аксиома выбора славится тем, что доставляет неудобные неконструктивные примеры множеств, привнося ложку дегтя в любую, сколь угодно большую и красивую, бочку мёда.

Аксиома детерминированности формулируется следующим образом:

AD: каждое $A \subseteq \mathcal{N}$ детерминировано.

Из написанного чуть выше сразу же следует, что аксиомы **AD** и **AC** несовместимы. Это значит, что в теории $\mathbf{ZF} + \mathbf{AD}$, которую мы обозначим **ZFD**, неверны ни теорема Цермело, ни лемма Цорна.

С другой стороны, известно, что **AD** влечет $\mathbf{AC}_\omega(\mathbb{R})$ — счетную форму аксиомы выбора для подмножеств \mathbb{R} . А это, в свою очередь, означает, что большинство теорем анализа остаются справедливыми, и настоящее сражение **AC** vs **AD** разворачивается за пределами вещественного континуума — в теории множеств и топологии.

Кроме того, **AD** вносит ясность (детерминированность) в вопрос формулировки континуум-гипотезы. Напомним, что существует два ее вида: алеф-версия и общая версия. Так вот, из **AD** следует общая версия **CH**, запрещающая наличие промежуточной мощности между ω и $\mathcal{P}(\omega)$, однако, **AD** противоречит формулировка **CH** в алеф-версии ($\aleph_1 = 2^{\aleph_0}$), поскольку это означало бы возможность вполне упорядочить континуум (а это означало бы возможность построить пример Витали, что невозможно — см. ниже). Отсюда же следует, что в теории **ZFD** вообще не существует модели гипердействительных чисел, т. к. (как уже отмечалось выше) единственной (с точностью до изоморфизма) моделью \mathbb{HR} при выполнении **CH** (но только в альфа-версии) является поле \mathbf{No}_{\aleph_1} .

На схеме 4.4 мы показываем логические зависимости между обсуждаемыми в этом разделе аксиомами.

Дальше — еще интереснее. Дело в том, что аксиома детерминированности исключает известные примеры неудобных множеств, доставляемых аксиомой выбора (теоремы Банаха, Мычельского и Сверчковского). А именно, в рамках **ZFD** все множества действительных чисел измеримы по Лебегу и обладают свойством Бэра (пример Витали исчезает). В рамках **ZFD** невозможен парадокс Банаха-Тарского об удвоении шара. В рамках **ZFD** не существует неглавного ультрафильтра на ω (что подрывает схему построения нестандартного анализа на основе ультрастепени ω).

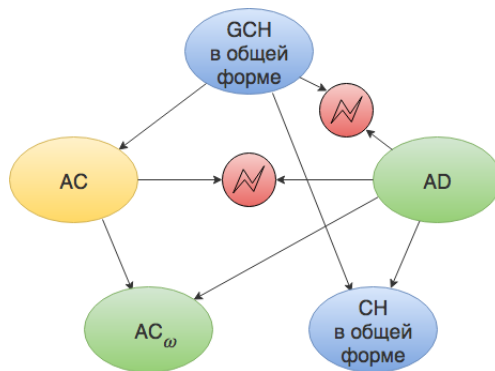


Рис. 4.4: Связь аксиом.

Больше того, в ZFD первый несчетный измеримый по Уламу кардинал существует и совпадает с \aleph_1 . Напомним, что измеримые по Уламу кардиналы — одни из самых гигантских кардиналов в теории ZFC (если допустить их существование), причем их существование не доказывается в ZFC. В ZFD, как видим, все гораздо проще. При этом \aleph_1 не есть мощность континуума.

В таблице 4.3 мы приводим сравнение выполнимости некоторых рассмотренных нами утверждений при различных аксиомах.

Таблица 4.3: Сравнение аксиом.

	AC	AC _ω	AD	ZF
Существование кардинальных мощностей множеств	✓	счетные	счетные	счетные
Неизмеримое и небэровское множество	существует	не зависит	нет	не зависит
Счетная сумма счетных мн-в счетна	✓	✓	✓	не зависит
Счетная посл-ть к предельной точке	✓	✓	✓	не зависит
Счетная аддитивность меры Лебега	✓	✓	✓	не зависит
Беск. мн-во содержит счетное подмн-во	✓	✓	✓	не зависит
Фильтр вкладывается в ультрафильтр	да	не зависит	не зависит	не зависит
Неглавный УФ на ω	существует	не зависит	нет	не зависит
Удвоение шара	существует	не зависит	исключено	не зависит
CH в общей форме	не зависит	не зависит	✓	не зависит

Наконец, добавим ложку дёгтя и в концепцию принятия аксиомы детерминированности. В последнее время (2011) появились результаты (см. [19, 20]), показывающие (с помощью конструктивного универсума), что аксиома SI

несовместна с ZF. Это автоматически отвергает аксиому детерминированности в ее общей формулировке, поскольку из нее следует существование измеримого кардинала (который является и строго недостижимым тоже). Если эти результаты верны, они могут оказаться революционными в теории множеств и потребовать нового поиска компромисса между странностью мира ZFC и чрезвычайным совершенством мира множеств в теории ZFD.

Пока таким компромиссом нам видится AC_ω или DC, либо какие-то специализированные варианты AD, допускающие в ряде случаев недетерминированные игры.

4.2.5 Промежуточный итог

Подводя локальный итог двум предыдущим разделам — об исчислении высказываний и предикатов, и об аксиомах теории множеств, хочется отметить следующее. Попытка поставить математику на компьютерные рельсы, полностью формализовать ее и сделать любой математический результат эффективно (т. е. за конечное время и с помощью детерминированного алгоритма) вычислимым, по всей видимости, не осуществима в полном объеме.

Как наличие теоремы Гёделя о неполноте, так и оперирование с конкретными примерами аксиом, доставляющими примеры этой теореме, поставили перед математиками еще больше проблем, чем решили ранее существовавших. Поведение аксиом выбора и детерминированности демонстрирует нам еще один пример фундаментальной неопределенности знаний. Первый такой пример — это пятый постулат Евклида. Оказывается, что даже фундамент математики можно выбирать по своему вкусу в зависимости от красоты и богатства получаемых в нем теорем. Так, тополог предпочитает сегодня работать в ZFC несмотря на все неконструктивные и явно противоречащие интуиции следствия аксиомы выбора. В то же время, тем, кто занимается анализом и математической физикой вообще нет дела до противоречий между следствиями AC и AD. Его скорее будет мучить вопрос о выборе подходящей геометрии для описания окружающего мира. Логик же находит удовольствие в том, чтобы выявлять контраст между различными формализмами, пытаясь тем самым понять, насколько они непротиворечивы и соответствуют интуиции, и можно ли вообще хоть как-то обосновать непротиворечивость ZF, а вместе с ней и математики в целом.

Тем не менее, стоит отметить один общий архетип в этих исканиях: **создание исчислений**. На каком бы уровне абстракции мы ни находились — будь то метаматематика, топология, анализ или теория вероятностей, мы всегда старательно расчищаем себе некоторое детерминированное пространство, в котором вместо нас могут работать роботы, т. е. компьютерные алгоритмы, формулы, формальные языки. Чем шире и искуснее становится это пространство, тем легче и богаче становится применение математики в других обла-

стях знаний и, стало быть, комфортнее жизнь в науке.

Исчисления, таким образом, на нешей общей математической картинке создают инфраструктуру. Если раньше (см. раздел 3.9) мы видели деревья, башни, поля и кольца с разнообразными надстройками, то с появлением исчислений наша картина снабжается сетями коммуникаций, службами обеспечения и помощи, иначе говоря, математическая картина постепенно превращается в мегаполис.

Дальнейшая наша задача будет состоять в том, чтобы более детально рассмотреть предоставленную инфраструктуру и выявить новые, а также подтвердить ранее найденные архетипы математики.

4.3 Вычислимость и доказуемость

4.3.1 Нестрогое введение в теорему Гёделя о неполноте

Ранее мы уже встречались с понятием алгоритмической вычислимости, когда говорили о разрешимой теории. Более точно, если у нас есть конечный набор аксиом в некотором формализме с конечной сигнатурой, то мы можем любую замкнутую формулу языка этой теории записать конечным числом символов. Возникает вопрос — выводима эта формула в данной теории или нет? И если существует такой конечный набор предписаний, выполняя которые компьютер сможет дать детерминированный ответ (да или нет) для любой замкнутой формулы, то мы говорим, что теория разрешима. Отчасти (и по терминологии тоже) это напоминает разрешимость алгебраических уравнений в радикалах (см. раздел 3.7.5). Уравнение разрешимо, если компьютер за конечное число шагов с помощью арифметических операций²⁵ найдет корень уравнения, отправляясь от коэффициентов уравнения.

Собственно, первая теорема Гёделя о неполноте и утверждает то, что в арифметике Пеано не каждую формулу можно доказать или опровергнуть (не для каждой существует эффективный доказывающий алгоритм), даже если она истинна или ложна в конкретной интерпретации. Пройдемся коротко по доказательству этой теоремы, пропуская некоторые существенные, но технически сложные моменты.

Итак, рассмотрим арифметику Пеано и совокупность всех формул, которые строятся в ее грамматике. Заметим, что построение формул по правилам грамматики алгоритмично, т.е. для каждой формулы можно указать конкретный финитный алгоритм ее построения. Среди этих формул будут

²⁵К ним, правда, относятся и корни, а их вычисление происходит приближенными методами, но тоже за конечное число шагов, и в этом смысле для компьютерных вычислений нет особой разницы между алгебраическими и вещественными числами, поскольку трансцендентные корни тоже ищутся за конечное число шагов с заданной точностью с помощью алгоритмов приближенных вычислений, которые, правда, иногда могут заикливаться :)

предикаты с одним параметром вида $\varphi(x)$. Ясно, что как все формулы вообще, так и формулы с одним параметром можно эффективно перенумеровать натуральными числами (например, сначала нумеруем формулы, состоящие из одного символа, коих конечное множество, затем — из двух, и т.д.). Пусть далее

$$\mathcal{F} = \{\varphi_0(x), \varphi_1(x), \dots\}$$

есть множество всех формул с одним параметром x .²⁶

Ключевой момент: предположим, что все формулы в \mathcal{F} имеют проверяющий их выводимость алгоритм (т. е. в арифметике либо φ_n , либо $\neg\varphi_n$ можно вывести с помощью исчисления предикатов).

Далее рассмотрим формулу $\psi(x) = \neg\varphi_x(x)$. Здесь мы видим работу **архетипа трансцендентного объективизма** (стр. 95), который мы уже встречали при доказательстве теоремы Кантора, причем именно в его «диагональной» форме (пересекаем все формулы по диагонали, приравнивая номер и аргумент). Ясно, что эта формула, с одной стороны, должна иметь проверяющий алгоритм (для этого нужно немного модифицировать алгоритм для φ_n), а с другой стороны, не может принадлежать множеству \mathcal{F} , т. к. равенство $\psi(x) \leftrightarrow \varphi_n(x)$ при всех x , означало бы и равенство $\psi(n) \leftrightarrow \varphi_n(n)$, а это не верно. Мы получили противоречие.

Можно предположить, что формула $\neg\psi(x)$ доказуема или опровержима, но тогда этот же самый алгоритм будет, соответственно, опровергать или доказывать саму формулу $\psi(x)$ в противоречии с полученным.

Следовательно, не все формулы с одним аргументом (произвольным натуральным числом) доказуемы или опровержимы в арифметике.

И тут мы снова можем вспомнить теорему Гудстейна с параметром n , являющимся стартовым числом для алгоритма Гудстейна.

4.3.2 О вычислимости, разрешимости, перечислимости

Теория вычислимости, также известная как теория рекурсивных функций, — это раздел современной математики, лежащий на стыке математической логики, теории алгоритмов и информатики, возникшей в результате изучения понятий вычислимости и невычислимости. Изначально теория была посвящена вычислимым и невычислимым функциям и сравнению различных моделей вычислений. Сейчас поле исследования теории вычислимости расширилось — появляются новые определения понятия вычислимости и идёт слияние с математической логикой, где вместо вы-

**Архетип
рекурсии!**

²⁶Обращаем внимание читателя на то, что сейчас мы «препарируем» арифметику, находясь в более мощной метатеории, например, в \mathbf{ZF} , поскольку оперируем понятием множества. На самом деле, этого можно избежать, если вместо множества \mathcal{F} использовать формулу $U(n, x)$ (см. ниже), которая при каждом n принимает ровно то же истинностное значение, что и формула φ_n .

числимости и невычислимости идёт речь о разрешимости и неразрешимости формальных теорий.

На самом деле, если читатель вновь обратится к первой главе, то может увидеть, что мы постоянно ходим вокруг более-менее похожих друг на друга финитных конструкций, которые на том или ином языке описывают различные математические задачи. При этом, немного отстраняясь от синтаксиса и углубляясь в семантику, мы начинаем ставить вопросы общего характера вроде «а всегда ли работает та или иная формула или алгоритм?». Тем самым, мы как бы поднимаемся над множеством всех этих финитных конструкций, воспринимая его как актуальное множество, как вещь в себе. А ведь оно бесконечное! И методы, необходимые для получения ответов на такие вопросы, требуют уже различных трансфинитных теорий, описывающих бесконечные структуры как есть, как отдельно взятый объект. Но сами эти теории вновь оказываются написанными на конечном языке конечным набором формул!

Таким образом, какие бы страшные гигантские бесконечности нам ни попадались в математике, наш мозг все равно оперирует ими как компьютер магниной лентой или файлом на конечном носителе, и потому всегда и везде с нами явно или неявно пребывает понятие вычислимости или разрешимости.

Рассмотрим теорию RA, аксиомы которой приведены на стр. 359. В грамматике этой теории допустимо строить различные высказывания о натуральных числах (т. е. функции вида $\mathbb{N}^k \rightarrow \{0, 1\}$ или, что то же самое, отношения вида $R \subseteq \mathbb{N}^k$). Кроме того, когда мы будем анализировать эту теорию, находясь во внешней метатеории ZF, мы будем говорить о стандартной модели $\langle \mathbb{N}, =, +, \cdot, S, 0 \rangle$,²⁷ которую в дальнейшем для краткости будем обозначать символом ее носителя, т. е. \mathbb{N} .

Дадим следующие определения. Функция из \mathbb{N} в \mathbb{N} называется **вычислимой**, если существует алгоритм, который, получив на входе некоторый набор чисел, за конечное число операций выдает на выходе значение этой функции на данных числах, либо же заикливается, если данная функция не определена на этом наборе.

Мы не уточняем здесь понятие алгоритма, поскольку наш главный герой компьютер знает о нем не понаслышке, как и любой квалифицированный читатель, а кроме того, рассматриваем так называемые частичные функции (т. е. не всюду определенные), так что алгоритм корректно работает только в области их определения, а вне ее — заикливается. Конечно, на практике иногда бывает трудно понять, заиклился алгоритм или же просто машина «долго думает».

Архетип
базового
множества.

²⁷В данном случае мы не выбираем конкретную модель \mathbb{N} , поэтому используем теоретико-групповой значок для множества натуральных чисел.

Множество $X \subseteq \mathbb{N}$ называется **разрешимым**, если его функция-индикатор ($i_X(n) = 1$ if $n \in X$ else 0) вычислима. Поскольку функцию-индикатор можно понимать как формулу с параметром, разрешимое множество есть область истинности некоторой формулы, проверку истинности или ложности которой можно осуществить алгоритмически. Заметим, что функция-индикатор *тотальна*, т. е. является всюду определенной.

Упражнение 4.11.
Подумайте, что может этому помешать в произвольной формуле.

Наконец, множество $X \subseteq \mathbb{N}$ **перечислимо**, если существует алгоритм, печатающий все без исключения элементы X , и только их. Иначе говоря, некоторая программа за конечное число итераций доходит до команды **print** и печатает какой-то элемент множества X , затем еще через какое-то конечное число итераций печатает какой-то (отличный от первого) элемент X , и т. д. Этот алгоритм может длиться вечно (если X — счетное), но любой элемент X через конечное число шагов обязательно будет напечатан, и никакие другие натуральные числа никогда не появятся в его выдаче.

Ясно, что все три определения можно расширить на случай нескольких числовых параметров (т. е. на случай \mathbb{N}^k).

Известны следующие свойства, связывающие эти три понятия:

Упражнение 4.12.
Найдите (в любом смысле) их доказательства.

Count1 Если частичная функция $f : \mathbb{N} \rightarrow \mathbb{N}$ вычислима, то ее область определения $\text{dom}(f)$ и область значений $\text{ran}(f)$ перечислимы.

Count2 Частичная функция $f : \mathbb{N} \rightarrow \mathbb{N}$ вычислима тогда и только тогда, когда ее график, т. е. множество $\{(x, y) \mid y = f(x)\}$, т. е., собственно говоря, само множество f , перечислимо.

Count3 Если множество X перечислимо, то функция, определяемая условием $z_X(n) = 0$ if $n \in X$ else *None*, является (частичной) вычислимой функцией.

Count4 Если X разрешимо, то X перечислимо.

Count5 (Обратно) Если X перечислимо и $\mathbb{N} \setminus X$ перечислимо, то X разрешимо.

Count6 Множество $X \subseteq \mathbb{N}$ перечислимо тогда и только тогда, когда оно является проекцией разрешимого множества $B \subseteq \mathbb{N} \times \mathbb{N}$.

Теорема 4.15. *Существует перечислимое множество X такое, что его дополнение неперечислимо. То есть, понятие разрешимости является существенно более сильным, чем понятие перечислимости.*

Рассмотрим некоторое семейство \mathcal{F} одноместных (т.е. с одним аргументом) вычислимых функций. Предположим также, что существует функция $U(n, x)$ такая, что при каждом n функция $U_n(x) = U(n, x)$ вычислима и принадлежит семейству \mathcal{F} , и всякая вычислимая функция \mathcal{F} является одной из U_n при некотором n . Иначе говоря, U перечисляет семейство \mathcal{F} . В этом случае говорят, что U является **универсальной функцией** (или **нумерацией**) семейства \mathcal{F} .

Известна теорема о том, что существует *вычислимая* универсальная функция двух аргументов, перечисляющая все унарные вычислимые функции от 1 аргумента. Используя перечисленные выше свойства, легко от функций перейти к множествам. А именно, пусть имеется некоторое семейство множеств натуральных чисел, тогда множество W называется универсальным для него, если всякое множество этого семейства (и только его) получается как правая проекция W при фиксированном левом аргументе: $X_n = \{x \mid (n, x) \in W\}$. Соответственно, существует перечислимое множество W , являющееся универсальным для класса всех перечислимых множеств натуральных чисел.

Оказывается, что не для всякого семейства \mathcal{F} универсальная функция является вычислимой. Допустим, что это не так и возьмем в \mathcal{F} все вычислимые *тотальные* (всюду определенные) функции, предполагая, что U также вычислима. Далее рассмотрим функцию, определяемую равенством $d(x) = U(x, x) + 1$. Ясно, что в силу вычислимости U функция d также вычислима и тотальна. Однако, она не может совпадать ни с одной из функций семейства \mathcal{F} (хотя бы в одной точке отличается на 1). Противоречие. Таким образом, справедлива

Теорема 4.16. *Не существует вычислимой тотальной функции двух аргументов, универсальной для класса всех вычислимых тотальных функций одного аргумента.*

Эта теорема является аналогом теоремы Гёделя о неполноте и, в общем-то, реализует ее центральную идею, только на функциях.

Среди универсальных функций особенную роль играют так называемые главные (или гёделевские) универсальные функции.

Главной (гёделевой) универсальной функцией (нумерацией) для класса унарных вычислимых функций называется такая вычислимая универсальная функция $U(n, x)$, что для любой вычислимой функции $V(m, x)$ существует тотальная вычислимая функция $s(m)$ такая, что

$$V(m, x) = U(s(m), x)$$

для всех m, x . При этом в случае, когда при каком-то аргументе значение слева не определено, оно не определено и справа, и наоборот (т.е. имеется ввиду совпадение графиков этих функций как множеств).

Теорема 4.17. *Существует главная универсальная функция.*

Прелесть главной универсальной функции (или как говорят, главной, или гёделевой, нумерации) состоит в следующей теореме.

Теорема 4.18. *Если U — главная универсальная функция для класса унарных вычислимых функций, то существует вычислимая тотальная функция $c(p, q)$ такая, что*

$$U(c(n, m), x) = U(n, U(m, x)).$$

Иначе говоря, имея номера n, m двух вычислимых функций, мы можем определить номер $c(n, m)$ функции, которая является их композицией $U_n(U_m)$.

Приведем некоторые свойства главных (гёделевых) нумераций.

Теорема 4.19. *Любые две главные нумерации изоморфны. То есть, если U_1 и U_2 — главные нумерации семейства вычислимых функций от одного аргумента, то существует такая вычислимая биекция $s : \mathbb{N} \leftrightarrow \mathbb{N}$, что $U_1(n, x) = U_2(s(n), x)$ и $U_2(n, x) = U_1(s^{-1}(n), x)$.*

Теорема 4.20 (Клини о неподвижной точке). *Для главной нумерации существует неподвижная точка. То есть, если U — главная нумерация семейства вычислимых функций от одного аргумента, то существует вычислимая тотальная функция $h(n)$ такая, что $U(h(n), x) = U(n, x)$ при любом натуральном x .*

Здесь имеется в виду, что мы рассматриваем функцию $h(n)$ как некий оператор над пространством вычислимых одноместных функций, и именно для такого оператора находим неподвижную точку. Действительно, взяв функцию U_n , применяем к ее номеру (одному из номеров) функцию h , получаем номер $h(n)$, берем функцию $U_{h(n)}$, т. е. получаем отображение $U_n \mapsto U_{h(n)}$. Теорема о неподвижной точке говорит о том, что функция h выдаст нам еще один номер той же самой функции, т. е. U_n перейдет в себя (как функция-множество) при смене номера.

Фазовое пространство?

Можно также показать, что таких неподвижных точек бесконечно много. Имеется следующее следствие теоремы о неподвижной точке.

Теорема 4.21. *Пусть U — главная нумерация семейства вычислимых функций от одного аргумента. Тогда существует такой номер p , что $U(p, x) = p$ для всех x .*

Объясним эту теорему следующим способом. Среди всех одноместных вычислимых функций, которые пронумерованы $U(n, x)$, несомненно, есть функции-константы вида $U_{n_k}(x) = p$ для всех x . У этих функций есть какие-то свои натуральные номера в общей нумерации: n_0, n_1, \dots . Теорема говорит

о том, что, перебирая эти функции, мы рано или поздно наткнемся на такую, номер которой совпадет с ее значением-константой: $U_{n_k}(x) = n_k$ для всех x .

Доказать это можно так. В теореме 4.20 в качестве функции h возьмем такую, которая по аргументу n находит такой номер n_k , что функция $U_{n_k}(x) = n$ для всех x . Тогда по теореме о неподвижной точке найдется такой номер $n = p$, что функции U_n и U_{n_k} совпадут, т.е. $U_p(x) = U_{n_k}(x) = p$, или $U(p, x) = p$ для всех x .

Заметим, что номер p функции можно интерпретировать довольно широко, если вспомнить о том, как в компьютере кодируется любая информация. Так, текст программы есть последовательность байтов, каждый из которых кодирует какой-либо символ этого текста. Стало быть, и весь текст можно считать каким-то (очень большим) числом, записанным по основанию 16 (или по основанию 2, если перейти к битам). Таким образом, и код программы, и ее выдачу в процессе работы (если она завершается за конечное число шагов) можно считать натуральными числами, а саму программу — вычислимой функцией одного аргумента.

Следовательно, к программам (на любом языке программирования) можно применять все теоремы о вычислимых функциях, в том числе теорему о неподвижной точке и ее следствие.

Таким образом, равенство $U(p, x) = p$ говорит нам о том, что в любом языке программирования существует программа, печатающая свой собственный код (на любом входе x).

Вот простейший пример такой программы для Python:

```
__='=%r;print(%%)';print(%%)
```

И вот еще один, правда, не совсем честный, т.к. использует распечатку файла с собственным кодом:

```
print(open(__file__).read())
```

Более ясный, но длинный код можно посмотреть в листинге С.4.

4.3.3 Про сигма-определимость

Скажем, что формула грамматики РА является Δ_0 -**формулой**, если все кванторы в ней имеют вид $(\exists x \leq t)$ и $(\forall x \leq t)$. Иначе говоря, область действия кванторов ограничена некоторым начальным отрезком натурального ряда (при этом ограничитель t является параметром формулы или термом с параметрами). Таким образом, все такие формулы есть, по сути, алгоритмы, выполнимые за конечное число шагов (кванторы можно рассматривать как циклы от 0 до t или соответствующие логические связки: для \exists — дизъюнкция, для \forall — конъюнкция). Усилением понятия Δ_0 -формул является Σ_1 -

формула (далее — *сигма-формула*), которая имеет вид

$$\exists \bar{x} \varphi_0(\bar{x}),$$

где φ_0 — это Δ_0 -формула, а \bar{x} обозначает набор числовых переменных (не вектор, а просто синтаксическую конструкцию из переменных, разделенных запятыми). Далее, Π_1 -формула получается навешиванием неограниченных кванторов всеобщности на сигма-формулу:

$$\forall \bar{x} \exists \bar{y} \varphi_0(\bar{x}, \bar{y}).$$

И далее процесс продолжается рекурсивно: Σ_2 навешивает \exists на Π_1 -формулы, а Π_2 — кванторы всеобщности на Σ_2 -формулы, и т.д. Таким образом, формула типа Σ_n имеет вид

$$\exists x_1 \forall x_2 \exists x_3 \dots \varphi_0(\bar{x}),$$

а формула типа Π_n имеет вид

$$\forall x_1 \exists x_2 \forall x_3 \dots \varphi_0(\bar{x}),$$

где количество кванторов равно n (вообще говоря, речь идет о группах одноименных подряд идущих кванторов вида $\exists x_i, x_j, x_k$ или $\forall x_i, x_j, x_k$).

Ограниченность кванторов, как мы уже отметили, позволяет считать Δ_0 -формулу алгоритмом, который, в частности, может определять график функции своей областью истинности, т.е. множество пар (\bar{x}, y) , для которых истинна $\varphi_0(\bar{x}, y)$, может оказаться графиком частичной функции $\bar{x} \mapsto y$. Такая функция, очевидно, является вычислимой, а пи-формула $\forall \bar{x} \exists y \varphi_0(\bar{x}, y)$ в таком случае будет утверждать, что она тотальна.

Комментарий 16.

Формула φ_0 может быть диофантовым уравнением с фиксированным числом параметров (коэффициентов) из \mathbb{N} , которые можно варьировать, получая разные уравнения. Множество всех значений этих параметров, при которых данное диофантово уравнение имеет натуральные решения, называется *диофантовым*. Например, уравнение $a - x^2 = 0$ определяет диофантово множество всех полных квадратов ($a = 1, 4, 9, 16, \dots$).

Так вот, известно, что диофантово множество перечислимо и, обратно, всякое перечислимое множество является диофантовым.²⁸

Отсюда следует известное негативное решение 10-й проблемы Гильберта [6]: не существует общего алгоритма, который узнавал бы по произвольному диофантову уравнению, имеет ли оно решения в целых числах или нет. Это утверждение

²⁸Это так называемая DPRM-теорема, совместный результат М. Дэвиса и Х. Патнема и Д. Робинсон.

выводится из приведенного нами ранее факта о существовании перечеислимого, но неразрешимого множества. Действительно, взяв в качестве M перечеислимое (т. е. диофантово) неразрешимое множество, мы не найдем алгоритма, распознающего для каждого натурального числа его принадлежность к M , т. е. строящего функцию-индикатор. А значит, нет и общего алгоритма.

Этот результат сильно напоминает теорему Галуа, не правда ли?

Если формула $\varphi(\bar{x})$ со свободными переменными $\bar{x} = x_1, \dots, x_k$ принадлежит классу Σ_1 , то область ее истинности, множество $P_\varphi \subseteq \mathbb{N}^k$, определяемое равенством

$$\bar{x} \in P_\varphi \leftrightarrow \varphi(\bar{x}),$$

называется

сигма-определимым. Главный результат о сигма-определимости состоит в следующем [9].

Теорема 4.22 (о сигма-определимости). *Множество $P \subseteq \mathbb{N}$ сигма-определимо тогда и только тогда, когда оно перечеислимо.*

В частности, пользуясь свойством Count2, получаем, что функция вычислима тогда и только тогда, когда ее график является сигма-определимым множеством.

Метатеорема 4.13. *Множество всех высказываний, истинных в стандартной модели \mathbb{N} , неперечеислимо.*

Доказательство. Пусть $K \subseteq \mathbb{N}$ перечеисливо и неразрешимо (теорема 4.15). По теореме о сигма-определимости найдётся формула $\varphi(x)$ такая, что²⁹

$$n \in K \Leftrightarrow \mathbb{N} \models \varphi(n).$$

Отсюда получаем

$$n \notin K \Leftrightarrow \mathbb{N} \models \neg\varphi(n).$$

Если множество всех истинных высказываний перечеисливо, то таково же и $\{n \in \mathbb{N} \mid \mathbb{N} \models \neg\varphi(n)\}$, так как по n эффективно восстанавливается формула $\neg\varphi(n)$ (подстановка символа натурального числа в фиксированную формулу является вычислимой операцией). Таким образом, будет перечеислимым также и дополнение множества K , что противоречит свойству Count5, в силу которого K получится разрешимым. \square

Как видим, данная теорема существенно использует понятие истинности формулы в стандартной модели \mathbb{N} , поэтому мы ее формулируем и доказываем как мататеорему, или как теорему в метатеории (коей является ZF).

²⁹Мы здесь пишем двойные стрелки эквиваленции \Leftrightarrow , поскольку работаем в метатеории.

На самом деле, если еще немного углубиться в формализацию понятий, то можно показать аналог этой метатеоремы внутри РА. При этом нам придется рассматривать арифметические коды высказываний.

Метатеорема 4.14 (Первая теорема Гёделя о неполноте). *Если теория T эффективно аксиоматизируема и $\mathbb{N} \models T$, то найдётся высказывание θ такое, что $T \not\vdash \theta$ и $T \not\vdash \neg\theta$.*

Доказательство. Поскольку $\mathbb{N} \models T$, множество T является частью множества всех высказываний, истинных в стандартной модели, значит, по предыдущей теореме найдётся истинное в этой модели высказывание θ такое, что $T \not\vdash \theta$. Так как $\mathbb{N} \models \neg\theta$, имеем $T \not\vdash \neg\theta$. \square

Данную теорему мы также оставили среди метатеорем, поскольку она основывается на предыдущей. Этот факт мы передокажем ниже, действуя непосредственно в формализме арифметики, т. е. получим как теорему теории РА.

4.3.4 Доказательство теорем Гёделя

Продолжая логику с кодом программы, который мы заменили двоичным числом, можно пойти еще дальше и рассмотреть кодирование любых арифметических формул. Каждая формула $\varphi(x)$ с одной свободной переменной x состоит из конечного числа символов алфавита грамматики. Занумеруем (хотя бы в порядке таблицы кодов `Unicode`) эти символы натуральными числами. Далее, имея текст формулы $\varphi(x)$, мы можем записать последовательность чисел k_0, k_1, k_2, \dots , соответствующих символам данного текста в порядке их следования. Далее, кодом формулы $\varphi(x)$ назовем число³⁰

$$\ulcorner \varphi(x) \urcorner \equiv p_0^{k_0} p_1^{k_1} p_2^{k_2} \dots,$$

где p_0, p_1, p_2 — простые числа в порядке возрастания.

Такое кодирование удобно тем, что оно однозначно восстанавливает текст формулы по числу (что следует из основной теоремы арифметики), а кроме того, алгоритмично, т. е. вычислимо. При этом, правда, стоит учитывать, что не всякое натуральное число соответствует корректной формуле грамматики, т. е. существуют «правильные» коды формул и «неправильные».

Далее, подставляя вместо x в формуле $\varphi(x)$ какое-либо конкретное число n , мы будем получать высказывания $\varphi(x||n)$ в языке арифметики, для которых можем точно так же вычислить код (он будет отличаться от кода $\varphi(x)$). Формально результат этих действий можно записать следующим образом:

$$S_{\ulcorner \varphi(x) \urcorner}(n) \equiv \ulcorner \varphi(x||n) \urcorner,$$

³⁰Мы не пользуемся здесь гёделевским β -кодированием для упрощения текста и концентрации внимания на главном.

при этом нужно понимать, что n в данном случае является аргументом функции $S_{\ulcorner \varphi(x) \urcorner}$, т. е. при вычислении кода каждый раз нужно сначала заменить его конкретным числом. А параметром этой функции (вторым аргументом) является код формулы $\varphi(x)$, в которой переменная x ничем не заменена. Тем еще интересней получается значение выражения $S_x(n)$, которое означает, что мы взяли формулу с номером x (это известный входной параметр), пусть это оказалась формула $\psi(x)$ (где x есть просто символ, который никак не связан с входным параметром x), а затем произвели замену $\psi(x||n_0)$ и вычислили код $\ulcorner \psi(x||n_0) \urcorner$. Формула с кодом x могла и не иметь свободной переменной x в своей записи или вообще не быть формулой, в этом случае замена $x||n_0$ оказалась бы тривиальной.

Например, пусть формула $x = 5$ имеет код 3 ($\ulcorner x = 5 \urcorner = 3$). Тогда

$$S_3(7) = \ulcorner (x = 5)(x||7) \urcorner = \ulcorner 7 = 5 \urcorner.$$

Или $S_3(3) = \ulcorner 3 = 5 \urcorner$. А если формула с номером 10 равна $y - 1 = z$, то результат будет $S_{10}(7) = S_{10}(10) = \ulcorner y - 1 = z \urcorner$, т. е. замена ничего не даст, исходные свободные переменные останутся на своих местах. Мы неспроста подчеркиваем здесь диагональный случай, когда параметр и аргумент функции S совпадают.

Пусть далее у нас фиксирована некоторая формула $\varphi(x)$. Посторим новую формулу

$$\psi(x) \equiv \varphi(x||z) \wedge (z = S_x(x)).$$

Это — формула с одной свободной переменной x , и у этой формулы тоже есть свой код. Наконец, определим высказывание без параметров:

$$\theta \equiv \psi(\ulcorner \psi(x) \urcorner).$$

Нетрудно видеть, что

$$\theta \leftrightarrow \psi(\ulcorner \varphi(x||z) \wedge (z = S_x(x)) \urcorner) = \varphi(\ulcorner \theta \urcorner),$$

поскольку

$$z = S_{\ulcorner \psi(x) \urcorner}(\ulcorner \psi(x) \urcorner) = \ulcorner \psi(\ulcorner \psi(x) \urcorner) \urcorner = \ulcorner \theta \urcorner.$$

Таким образом, мы получаем еще одну теорему о неподвижной точке, только для формул с параметром x .

Теорема 4.23. *Для произвольной формулы $\varphi(x)$ с единственной свободной переменной x существует высказывание θ такое, что $\theta \leftrightarrow \varphi(\ulcorner \theta \urcorner)$.*

Данная теорема в такой формулировке является метатеоремой для арифметики Пеано, поскольку оперирует таким объектом как формула. На самом деле, закодировав все формулы числами и рекурсивно определив понятие

«правильного» кода, ее можно записать чисто арифметически. Поэтому в данном случае можно сказать, что сама арифметика «знает» об этой теореме, т.е. приведенная теорема является теоремой РА (поэтому мы не стали называть ее метатеоремой).

Что же еще «знает» о себе арифметика?

Имея коды формул, очень просто построить коды доказательств (выводов). Действительно, пусть имеется конечная цепочка формул $\varphi_0, \varphi_1, \dots, \varphi_n$, каждая из которых либо является аксиомой РА, либо получается из предыдущих формул или тавтологий ИП по правилам вывода. В этом случае код такой цепочки определяется аналогично кодам формул:³¹

$$\ulcorner \varphi_0, \varphi_1, \dots, \varphi_n \urcorner \Rightarrow p_0^{\ulcorner \varphi_0 \urcorner} p_1^{\ulcorner \varphi_1 \urcorner} \dots p_n^{\ulcorner \varphi_n \urcorner}.$$

Поэтому на языке арифметики можно выразить следующий предикат:

$$Proof(x, y) \Rightarrow y \text{ есть код доказательства формулы с кодом } x,$$

и одноместный предикат:

$$Pr(x) \Rightarrow \exists y Proof(x, y),$$

выражающий тот факт, что формула с кодом x доказуема (выводима) в арифметике.

Теорема 4.24 (Первая теорема Гёделя о неполноте). *Если РА непротиворечива, то в языке РА существует формула, независимая от аксиом РА.*

Доказательство. Рассмотрим формулу $\neg Pr(x)$. По теореме о неподвижной точке существует высказывание θ такое, что

$$\theta \leftrightarrow \neg Pr(\ulcorner \theta \urcorner). \quad (4.6)$$

Но тогда высказывание θ не доказуемо и неопровержимо в РА. Действительно, если θ можно доказать, то $Pr(\ulcorner \theta \urcorner)$ истинно, откуда θ ложно. Если же можно доказать $\neg \theta$, то $Pr(\ulcorner \theta \urcorner)$ истинно, откуда существует доказательство θ . В любом случае мы получаем противоречие. А поскольку мы предположили непротиворечивость РА, то высказывание θ невозможно ни доказать, ни опровергнуть в РА. \square

Данная теорема хоть и выглядит как метатеорема, но и ее мы можем формализовать и доказать в РА. Именно, рассмотрим формулу $\text{Con}(РА) \Rightarrow$

³¹Заметим, что похожим образом мы определяли коды мультимножеств, но их арифметика была проще, т.к. не было такого разнообразия символов!

$\neg Pr(\ulcorner 0 = 1 \urcorner)$, которая в нашей кодировке означает невозможность доказательства равенства $0 = 1$. Это ровно то же самое, что утверждено о непротиворечивости РА. Тогда первая теорема Гёделя принимает арифметическую формулировку

$$\text{Con}(PA) \rightarrow \neg Pr(\ulcorner \theta \urcorner) \wedge \neg Pr(\ulcorner \neg \theta \urcorner),$$

где θ — гёделева неразрешимая формула (4.6).

Теорема 4.25 (Вторая теорема Гёделя о неполноте). *Если РА непротиворечива, то в ней невозможно доказать $\text{Con}(PA)$.*

Доказательство. Внутри самой арифметики утверждение теоремы равносильно высказыванию

$$\text{Con}(PA) \rightarrow \neg Pr(\ulcorner \text{Con}(PA) \urcorner).$$

Действительно, предполагая $\text{Con}(PA)$, по первой теореме Гёделя мы получаем, что истинна формула $\neg Pr(\ulcorner \theta \urcorner)$, которая эквивалентна самой θ . То есть, предполагая совместность РА, выразимую в самой РА, мы получили противоречие с первой теоремой Гёделя. Следовательно, либо РА несовместна, либо утверждение о ее совместности не выводимо средствами РА. \square

Подчеркнем, что невозможность доказать непротиворечивость РА средствами самой РА не означает, что она противоречива. Скорее всего, это не так. И в рамках ZF она действительно непротиворечива, т. к. имеет модель на ординале ω . При этом, правда, мы предполагаем непротиворечивость самой ZF. Но уверенность в этом нам доставляет во-первых, самоочевидная непротиворечивость теории «начальных» множеств (объекты которой — конечные строки из фигурных скобок), которая, к тому же, равносильна самой РА, а во-вторых, тот факт, что за более чем 100 лет ее «тестирования» противоречие не было найдено.

Обе теоремы Гёделя обобщаются на большинство «разумных» теорий, т. е. таких, которые способны к рефлексии — рассуждениям о самих себе. Требования к «разумным» теориям следующие:

Gödl1 В теорию можно погрузить вычислимые функции (т. е. в теории можно интерпретировать РА, вообще говоря, без аксиомы индукции)³²;

Gödl2 В теории можно формализовать алгоритм, распознающий аксиомы и верифицирующий доказательства (т. е. теория эффективно аксиоматизируема);

³²На самом деле тут речь идет о том, что можно вместо теории РА рассмотреть *арифметику Робинсона*, в которой рекурсивно задается порядок: $a \leq S(b)$, если $a \leq b \vee a = S(b)$, а аксиома индукции заменена на условие связности порядка.

Göд3 Теория должна быть корректной (все доказуемые высказывания должны быть истинными в любой модели).

Теории, удовлетворяющие этим условиям, принято называть **гёделевыми**. К таким теории относятся, например, ZF, ZFC и другие модификации как арифметики Пеано, так и теории множеств.

Отметим, что в современной теории доказательств вместо символа Pr принято использовать \Box . Так, формула $\Box\varphi$ означает, что формула φ доказуема в теории T (предполагается, что теория T задана в контексте). Кроме того, имеется сопряженный оператор $\Diamond\varphi$, означающий невозможность опровергнуть формулу φ . Таким образом,

$$\Diamond = \neg\Box\neg, \quad \Box = \neg\Diamond\neg$$

Наконец, несложно понять, что в ZF выполняются правила:

$$\text{Box1 } \Box(\varphi \wedge \psi) \leftrightarrow \Box\varphi \wedge \Box\psi;$$

$$\text{Box2 } \Diamond(\varphi \vee \psi) \leftrightarrow \Diamond\varphi \vee \Diamond\psi;$$

$$\text{Box3 } \Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi).$$

В тех же «разумных» теориях доказуема

Теорема 4.26 (Лёба). *Если доказуемо $\Box\varphi \rightarrow \varphi$, то φ .*

Действительно, пусть φ — некоторое высказывание. Тогда по теореме 4.23 о неподвижной точке имеем: существует высказывание θ такое, что

$$\theta \leftrightarrow (\Box\theta \rightarrow \varphi),$$

откуда по свойству Box3 получаем

$$\Box\theta \rightarrow (\Box\Box\theta \rightarrow \Box\varphi).$$

Далее, предполагая посылку $\Box\varphi \rightarrow \varphi$, получаем

$$\Box\theta \rightarrow (\Box\Box\theta \rightarrow \varphi),$$

откуда

$$\Box\theta \rightarrow \varphi,$$

что эквивалентно θ . То есть из посылки теоремы выводится θ :

$$(\Box\varphi \rightarrow \varphi) \vdash \theta,$$

т. е. θ доказуемо в предположении теоремы:

$$(\Box\varphi \rightarrow \varphi) \vdash \Box\theta,$$

откуда уже выводится φ , поскольку выше была выведена импликация $\Box\theta \rightarrow \varphi$. Таким образом, в предположении теоремы мы вывели φ , и теорема Лёба доказана.

Если говорить о произвольных «разумных» теориях в рамках теории доказательств, то там принято **Box3** считать аксиомой такого расширенного исчисления высказываний, и поэтому теорема Лёба в них также остается справедливой.

Вообще, если говорить о формализмах теории доказательств, то имеется список аксиом K4, состоящий из:

K4:1 всех аксиом исчисления высказываний;

K4:2 аксиомы $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$;

K4:3 аксиомы $\Box\varphi \rightarrow \Box\Box\varphi$;

K4:4 правил вывода: $\frac{\varphi}{\Box\varphi}$, modus ponens $\frac{\varphi, \varphi \rightarrow \psi}{\psi}$.

Отметим, что аксиома $\Box\varphi \rightarrow \Box\Box\varphi$ говорит нам о том, что если имеется доказательство φ , то мы можем проверить, что оно действительно является доказательством, и тем самым доказать утверждение $\Box\Box\varphi$. Данная аксиома в ZF является формализацией правила вывода $\frac{\varphi}{\Box\varphi}$, а без привязки к ZF постулируется как собственно аксиома.

Подробнее с языком и особенностями рассуждений в теории доказательств можно ознакомиться, например, в видеокурсе «Доказуемость» [S_7H9zCM1D0] д.ф.-м.н. Л. Беклемишева.

Мы наблюдаем здесь очередную попытку свести к исчислению не только теории, описывающие какие-то математические структуры, но и саму матлогику. Выясняется, что для всякой теории в таком расширенном ИВ можно построить некоторую алгебру доказуемости (на самом деле, булевскую), и рассматривать на ней действие оператора \Box так, как это принято в обычной Алгебре. Более того, утверждение о совместности $\text{Con}(T)$ формализуется как $\neg\Box\perp$ (символ \perp означает ложь, т.е. нельзя доказать ложь), и для совместной теории можно рассматривать самоприменение такого оператора: $T_0 = T$, $T_1 = T_0 + \text{Con}(T_0)$, $T_2 = T_1 + \text{Con}(T_1)$ и т.д. Так же, как в Алгебре, вводится понятие характеристики теории как того номера n , на котором данная цепочка оборвется из-за противоречивости теории T_n . Соответственно, «хорошие» теории вроде PA или ZF имеют характеристику ω (аналог поля характеристики 0). Исследования в этой области находятся на передовом крае науки и, как ни странно, имеют прямые применения своих формализмов в инженерии, где требуется построение некоторых формально-логических моделей поведения «умного» оборудования.

4.3.5 Несколько слов о рекурсиях

Выше мы видели, что три понятия — функции, множества и свойства (формулы) — тесно связаны в контексте понятий вычислимости, разрешимости и перечислимости. Как функция (через ее график), так и свойство (через его область истинности) сводятся к множеству, о котором можно сказать, вычислимо оно или нет, и обладает ли таковым свойством его дополнение (т. е. разрешимо ли само множество). Ну, а кодируя формулы и цепочки доказательств вычислимым кодом, можно вообще говорить о разрешимости теории (т. е. перечислимости всех ее доказательств и теорем).

Материал этого параграфа основан на **видеолекции «Доказуемо рекурсивные функции»** [Z3V0vtyXQvU] д.ф.-м.н. Л. Беклемишева.

Кроме того, вычислимость функции, оказывается, есть ровно то же самое, что сигма-определимость формулы, которая задает график этой функции. Такие функции также принято называть сигма-функциями. Оказывается, что класс таких функций настолько устойчив, что он определяется весьма разными, но, тем не менее, эквивалентными моделями:

M1 Вычислимые частичные функции;

M2 Сигма-определимые функции;

M3 Частично рекурсивные функции (К. Гёдель, С. Клини);

M4 Функции, заданные алгоритмами в λ -исчислении (А. Чёрч);

M5 Алгоритмы Маркова;

M6 Функции, вычисляемые машиной Тьюринга (А. Тьюринг, Э. Пост);

M7 Машины с неограниченными регистрами;

M8 C++, Python, Java, Lisp и т.д.

Эквивалентность всех этих моделей (обоснованная математически) говорит в пользу (нематематического) тезиса Чёрча–Тьюринга, который заключается в следующем суждении: *любая вычислимая в интуитивном смысле частичная функция вычислима на машине Тьюринга.*³³

Сужение этого класса функций возможно только путем совсем уж явного их определения.

³³ Существует также физическая интерпретация тезиса Чёрча–Тьюринга, которая говорит о том, что любая функция, вычисляемая на физическом устройстве, вычислима и на машине Тьюринга. Тем самым предполагается, что на физических устройствах (включая квантовые и аналоговые компьютеры), мы не получим ничего нового. Разница может быть лишь в огромной разности скоростей вычислений.

Далее нас будут интересовать т.н. доказуемо рекурсивные функции относительно некоторой «разумной» теории T , содержащей натуральные числа (например, PA или ZF) в том смысле, что в ней можно формализовать аксиомы Пеано (в варианте Робинсона), а также существует алгоритм распознавания аксиом и доказательств.

Функция $f : \mathbb{N} \rightarrow \mathbb{N}$ называется **доказуемо рекурсивной** (в T), если она, во-первых, вычислима, а во-вторых, ее график задан такой сигма-формулой $\varphi(x, y)$, что Π_2 -формула $\forall x \exists y \varphi(x, y)$ выводима (доказуема) в теории T ($T \vdash \forall x \exists y \varphi(x, y)$). Иначе говоря, доказуемо рекурсивная функция — это такая тотальная вычислимая функция, о которой теория T «знает», что она вычислима.

Изучение таких функций важно, прежде всего, для теории алгоритмов и Computer Science в целом. Действительно, имея код программы, нам всегда хочется знать ее корректность (доказуемость), в частности, то, что она завершит работу за конечное время при любых входных данных. Кроме того, изучая классы доказуемо рекурсивных функций двух разных теорий, мы можем сравнивать эти теории: если эти классы функций различны, то различны и сами теории.

Обозначим $\mathcal{F}(T)$ класс всех доказуемо рекурсивных функций теории T . Интересен следующий факт, показывающий, что класс $\mathcal{F}(T)$ строго меньше класса всех вычислимых тотальных функций.

Теорема 4.27. *Существует тотальная недоказуемая в теории T вычислимая функция.*

Действительно, пусть P — класс всех тотальных вычислимых функций, и предположим, что все функции из P доказуемы в T .³⁴ Тогда существует универсальная функция $U(n, x)$ для класса P , которая будет вычислимой.³⁵ Далее, положим $f(x) = U(x, x) + 1$, т.е. вновь прибегнем к диагональному методу (см. теорему 4.16). И увидим, что f не может находиться в классе P , хотя она тотальна и вычислима, а ее формула выводима в T . Имеем противоречие. Следовательно, существуют тотальные вычислимые недоказуемые в T функции.

Данная теорема показывает тот факт, что если теория T «разумна», то класс вычислимых функций, про которые она знает, что они тотальны, не может совпадать с классом всех тотальных вычислимых функций.

³⁴Мы здесь отождествляем доказуемость функции с доказуемостью определяющей ее Π_2 -формулы.

³⁵Точнее, при определении $U(n, x)$ мы пользуемся тем, что все доказательства теории T можно рекурсивно перенумеровать, и если n — номер вывода формулы $\forall x \exists y \varphi(x, y)$, задающей тотальную функцию, то в качестве ответа $U(n, x)$ возвращает такой y , что $\varphi(x, y)$. При этом нужно понимать, что может быть несколько выводов одной и той же формулы, т.е. U нумерует наши функции с избытком.

Это — слабый вариант первой теоремы Геделя о неполноте, поскольку мы пользуемся выводимостью формул с двумя кванторами (Π_2 -формулами). Тем не менее, он тоже демонстрирует нам тот существенный зазор, который возникает между истинностью и доказуемостью в более-менее сложных теориях. Ведь существование недоказуемой функции означает истинность ее формулы в некоторой модели, т. е. принципиальную непротиворечивость данной формулы. В то же время, отсутствие ее вывода в теории T означает независимость такой формулы от теории T .

Конечно, в нашем кратком доказательстве значительная часть логических построений и обоснований упрятана «под ковер». Например, обоснование вычислимости U или доказуемости f . Эти-то трудности и преодолевал в свое время Гёдель, не имея еще за плечами развитого аппарата теории доказательств, в частности, он сам стоял у истоков теории вычислимости, алгоритмов и т.п.

*Даже
машина
Тьюринга
появилась
позже!*

Насколько маленьким или большим может быть класс $\mathcal{F}(T)$ доказуемо рекурсивных функций в зависимости от теории T ?

Известно, что если в арифметике Пеано в аксиоме индукции заменить произвольную формулу на формулу типа Σ_1 , то в такой теории класс доказуемо рекурсивных функций совпадет с классом PR примитивно рекурсивных функций (Parsons, 1970). Теория PA , в которой таким способом ослабляется аксиома индукции, обозначается $I\Sigma_1$, так что теорема Парсонса гласит $\mathcal{F}(I\Sigma_1) = PR$.

Аналогично, если индукцию немного усилить и вместо сигма-формул разрешить использовать Π_2 -формулы без параметров, то мы получим теорию IP_2^- , для которой также установлено равенство $\mathcal{F}(IP_2^-) = PR$ (Беклемишев, 1996).

Самое время дать определение. Функция $f : \mathbb{N}^k \rightarrow \mathbb{N}$ называется **примитивно рекурсивной**, если она может быть получена из константы 0, функции следования $S(x) = x + 1$ и проектирующих функций $I_n^m(x_1, \dots, x_n) = x_n$ с помощью операций композиции (подстановки) и примитивной рекурсии. Функция f получена из исходных функций g и h примитивной рекурсией, если

$$\begin{cases} f(0, \vec{x}) = g(\vec{x}), \\ f(n + 1, \vec{x}) = h(f(n, \vec{x}), n, \vec{x}). \end{cases}$$

Рассмотрим конкретные примеры. Пусть

$$\begin{cases} a(0, x) = x \\ a(n + 1, x) = a(n, x) + 1 \end{cases}$$

Функция a определяет сложение натуральных чисел и является примитивно

рекурсивной. Аналогично

$$\begin{cases} m(1, x) = x \\ m(n+1, x) = a(m(n, x), x) \end{cases}$$

определяет умножение. Напомним, что ровно таким способом мы определяли сложение и умножение ординалов (плюс правило для предельного ординала).

Рассмотрим еще пример:

$$A(n, x, y) = \begin{cases} x^y, & n = 1 \\ 1, & y = 0 \\ A(n-1, x, A(n, x, y-1)), & \text{иначе.} \end{cases}$$

В стрелочной нотации Кнута эта функция задает следующие функции. При $n = 1$ возведение в степень x^y , при $n = 2$ привычную нам башню степеней $x \uparrow \uparrow y$, которая является y -кратным самоприменением операции возведения в степень, при $n = 3$ имеем $x \uparrow \uparrow \uparrow y$ — y -кратное самоприменение операции с двумя стрелками, и т.д. В общем виде можно записать так:

$$A(n, x, y) = x \uparrow^n y = \underbrace{x \uparrow^{n-1} (x \uparrow^{n-1} (\dots \uparrow^{n-1} x))}_{y \text{ копий } x}$$

Такая функция носит название **функции Аккермана**. Можно дать упрощенный вариант, немного модифицируя рекурсию:

$$A_0(x) = x + 1, \quad A_{n+1}(x) = \underbrace{A_n(A_n(\dots A_n(x)))}_{x+1 \text{ раз}}$$

Например, $A_1(x) = 2x + 1$, $A_2(x) > 2^x x$, $A_3(x) > 2^{2^{\dots}}$ (башня из x двоек), и т.д. То есть, она ведет себя примерно так же, как вышеопределенная функция при $x = 2$.

Функция Аккермана интересна тем, что она доставляет пример вычислимой, но не примитивно рекурсивной (в арифметике) функции (это знаменитая *теорема Аккермана*,³⁶) т.е. класс вычислимых (сигма-) функций заведомо шире класса примитивно рекурсивных функций. Если мы фиксируем n , то $A_n(x)$ все еще остается примитивно рекурсивной, но функция вида $A_x(x)$ уже не является примитивно рекурсивной (это следует из того, что она растет быстрее, чем может расти примитивная рекурсия согласно упомянутой теореме Аккермана).

³⁶ Аккерман — ученик и последователь Гильберта

Нужно заметить, что доказательство существования $A_x(x)$ использует индукцию с двумя кванторами и потому находится в сумме теорий $I\Sigma_1$ и IP_2^- . А так как известно, что функция $A_x(x)$ выходит за пределы класса PR , то вместе с упомянутыми результатами $\mathcal{F}(I\Sigma_1) = PR = \mathcal{F}(IP_2^-)$ мы получаем, что никакая из этих теорий не содержится в другой, т. е. это существенно разные теории.

Упражнение 4.13.
Постройте доказательство с двумя вложенными индукциями.

Примечательно, что о классе $\mathcal{F}(ZF)$ нам практически ничего не известно.

Далее. Обозначим $A_\omega(x) = A_x(x)$. Ничто не мешает нам в (ZF) определить примитивные рекурсии, отправляясь уже от этой функции (ранее мы стартовали с арифметической функции $S(x) = x + 1$):

$$A_{\omega+n+1}(x) = \underbrace{A_{\omega+n}(A_{\omega+n}(\dots A_{\omega+n}(x)))}_{x+1 \text{ раз}}.$$

И дальше, для предельного ординала α выберем какую-нибудь конфинальную ему последовательность α_n ($\sup \alpha_n = \alpha$) и положим

$$A_\alpha(x) = A_{\alpha_x}(x).$$

Выбрать эти последовательности можно конструктивно (без аксиомы выбора) для ординалов $\alpha < \varepsilon_0$, поскольку каждый такой α имеет конечную запись в виде разложения по супероснованию ω (нужно самые верхние «омеги» в самых правых (меньших) башнях степеней подменять натуральными числами).

Таким способом мы построим класс $\mathcal{F}_{\varepsilon_0}$ доказуемо рекурсивных функций всей арифметики Пеано, т. е. $\mathcal{F}(PA)$ (Аккерман, Швихтенберг, Вайнер [40]). При этом предельная функция Аккермана $A_{\varepsilon_0}(x) = A_{\alpha_x}(x)$ для последовательности α_n , конфинальной ординалу ε_0 , уже выйдет за пределы класса $\mathcal{F}_{\varepsilon_0}$, т. е. не будет доказуемой в арифметике PA . По этой причине ординал ε_0 так важен и называется *доказательным ординалом арифметики Пеано*.

В связи с этим мы вновь можем вспомнить последовательность Гудстейна (раздел 1.4). А именно, обозначим через $Goodstein(x)$ такой номер n , при котором последовательность Гудстейна, стартовав с числа x , завершится на n -ом шаге (в ZF это доказуемо, как мы видели ранее). Известно, что функция $Goodstein(x)$ имеет порядок роста такой же, как функция $A_{\varepsilon_0}(x)$. Этот результат получен в 1982 году (Jeff Paris & Laurie Kirby [37]). Отсюда, собственно говоря, и следует недоказуемость теоремы Гудстейна в арифметике.

Можно ли каким-то образом определить доказательный ординал ZF (возможно, с помощью аксиомы существования каких-то сверхбольших кардинальных чисел), в настоящее время неизвестно.

Финал главы 4

В этой главе мы рассмотрели только те исчисления, которые относятся к математической логике. И при этом все равно за бортом осталось много чего интересного. Например, мы совсем не осветили теории, связанные с машиной Тьюринга и λ -исчислением, а также известную проблему $P=NP$, важную как для Computer Science, так и для матлогики в целом.

Тем не менее, надеюсь, что нам удалось показать, насколько глубоко матлогика проникает в самую суть математики, и сколько общего возникает у математики и Computer Science на уровне ее оснований. А это о многом говорит!

Комментарий 17. Об одной NP-полной задаче

Порой возникает потребность в такой арифметической задачке как подбор слагаемых из заданного набора так, чтобы их сумма совпала с заданным числом. Вроде бы чистая теория, однако для успешной сдачи отчетности в наши славные органы госконтроля находится применение и этой задаче. Проблема в том, что решается она почти полным перебором и только в целых числах, что существенно замедляет алгоритм (перебор напрямую связан с величиной суммы).

Тем не менее, ее можно решать в несколько этапов: сначала найти (хоть вручную, хоть этим же алгоритмом) несколько крупных слагаемых и, тем самым, свести задачу к подбору значительно меньшей суммы. Алгоритм работает с некоторым допуском ошибки, поскольку не всегда из заданных слагаемых можно сложить целевую сумму. Кроме того, величина допуска влияет на скорость работы алгоритма.

Задача очень похожа на знаменитую Coin Problem о представлении произвольной суммы денег монетами заданного достоинства. Только в этой задаче одинаковых монет можно брать сколько угодно, в то время как в нашей задаче каждое из слагаемых данного набора может входить в сумму лишь однажды. Coin Problem, являясь более слабой задачей, также в общем случае имеет сложность NP.

Упоминание о Coin Problem мы еще встретим в разговоре о случайных графах (см. стр. 575). В конце книги можно увидеть реализацию алгоритма поиска слагаемых под заданную сумму из разрешенного списка возможных слагаемых. См. листинг C.5.

Рассмотренные здесь исчисления дают твердую основу под всей математикой в том смысле, что сводят процесс порождения и проверки математических утверждений к некоторой простой игре по формальным правилам, которой можно научить компьютер. И действительно, существуют компьютерные программы (например, *Coq*), которые способны переваривать фор-

мализм теории и в ряде случаев проверять, действительно ли утверждение φ выводимо из аксиом теории, а также перебором искать новые теоремы.

Конечно, как мы знаем хотя бы из гёделевских теорем, это не всегда возможно. Пространство теорем, порождаемое механически по правилам игры, строго меньше пространства всех истинных (относительно ZF) утверждений. Тем не менее, человеку оказываются доступны и такие «недоказуемые» вещи.

Здесь могут возразить, что невыводимые в арифметике теоремы все-таки выводимы в ZF, а значит, машина в состоянии проверить и их, если ей дать аксиомы ZF. Но дело в том, что человек при этом способен заглянуть и за пределы ZF и ZFC, и поставить вопросы об истинности таких утверждений, как континуум-гипотеза, существование измеримого кардинала и т.п. А их, как мы знаем, невозможно доказать и в рамках ZFC.

Иначе говоря, какова бы ни была алгоритмическая (машинная) система описаний мира, человек способен найти какое-то знание вне этой системы и поставить вопрос о возможности присоединения такого знания к уже описанному миру, после чего появится возможность его формализации и машинного анализа. Это — квинтэссенция архетипа *неограниченного расширения*. Это то, что отличает математика от компьютера.

Тем не менее, сама возможность рассматривать теории как алгебраические структуры и получать некоторые общие представления о теориях вообще и о конкретных, хорошо известных, теориях в частности — это колоссальный прорыв математики 20-го века. Как мы уже отмечали ранее, исчисления превращают открытия в системные знания. Сами по себе исчисления — это конвейер, который воплощает в себе уникальные знания инженеров и тиражирует их в виде новых сущностей по строго выверенным правилам. Именно строгость этих правил позволяет нам рассчитывать на то, что все порождаемые конвейером продукты математического знания ничем не уступают исходным аксиомам в смысле истинности и непротиворечивости и могут быть верифицированы любым достаточно грамотным специалистом, т. е. не являются субъективными. Во многом благодаря такому свойству математики ракеты доставляют спутники на заданную орбиту, GPS в вашем телефоне работает с высокой точностью, а роботы делают качественные автомобили уже на физическом конвейере.

Разобравшись с основаниями математики, мы переходим (или возвращаемся) в собственно математику, и в следующей главе рассмотрим некоторые исчисления, позволяющие работать с уже известными нам числовыми структурами. К той картине, о которой мы говорили в разделах 3.9 и 1.1.3, мы теперь добавим динамику: среди выстроенных ранее башен, полей и механизмов возникнут улицы, коммуникации, движение теней, картинка оживет и задышит.

Собственно, отчасти мы это уже пронаблюдали, работая с формальными теориями, когда изучали влияние аксиомы выбора на математические резуль-

таты, а также строили исчисление предикатов. Длинные цепочки выводов и порождающих их алгоритмов, ветвления в точках принятия или непринятия тех или иных аксиом — это и есть своего рода улицы и перекрестки в каменных джунглях математических теорий. Здесь уместно также упомянуть известные логикам модели Крипке, где основные объекты изучения прямо так и называются: *мирами*. А наследование свойств миров и их параллельность в логическом смысле, вместе с их абсолютной детерминированностью, являет собой метафизическую суть оснований математики.

Анализ. Исчисления II

В этой главе мы рассмотрим несколько задач анализа, которые никак не связаны с предыдущей главой, а целиком ложатся в вещественный или комплексный анализ, но играют значительную прикладную роль. Особенность подхода в изложении данных задач будет заключаться в том, что мы их рассмотрим одновременно, не взирая на большую разницу в их сложности. Как мы уже отмечали, цель книги не в том, чтобы дать материал в обучающей последовательности, скорее наоборот — подать его подготовленному читателю в виде сравнительных срезов.

Отметим также, что на этот раз теорема Гудстайна не появится в качестве назидательного и глубокого примера, как это было ранее.

Для более глубокого изучения свойств топологических пространств рекомендуем читателю обратиться к книге [105].

5.1 Пространства и отображения

Мы много раз употребляли термин «пространство», но до сих пор прозвучало только такое его определение в финале главы 1: *множество с системой подмножеств называется пространством, поскольку такая конструкция использует «пространные» объекты — подмножества.*

Было также дано определение векторного пространства, которое, вообще говоря, идет вразрез с предыдущим определением через систему подмножеств, хотя в дальнейшем мы увидим, что это как раз не проблема определения, а просто подход с другой стороны.

А еще было
простран-
ство
Бэра!

Чтобы как-то определиться, договоримся термин **пространство** всегда ассоциировать с топологическим пространством и всевозможными его надстройками и пристойками. Даже если топология не задана явно (как в случае с векторным пространством), но ее можно «естественным» образом задать, то будем считать, что мы имеем дело с пространством. Кроме того, функции, действующие из одного пространства в другое, чаще всего называют **отображениями**, а в более конкретных случаях они могут называться числовыми функциями, операторами, функционалами и т.п. Термин «отображение» тоже более топологический, чем теоретико-множественный, как и термин «пространство».

Почему именно топология? Дело тут, прежде всего, в том, что под про-

странством нам хочется понимать некоторый объем, внутри которого можно так или иначе выделять некоторые области или подпространства, имеющие, вообще говоря, нетривиальную структуру. Например, отрезок $[0; 1]$, квадрат или куб — пространства, а вот ординал $\omega + 1$ — вряд ли. Слишком уж он дискретен и специфичен своей теоретико-множественной интерпретацией. Хотя, конечно, топология — столь гибкое понятие, что ее можно «натянуть» на любое множество. Но это скорее дань формализму и простоте аксиоматизации данного понятия, чем самоцель.

5.1.1 Связность и непрерывность

Итак, в соответствии с методологией двухтактной модели построения математических систем (см. стр. 106), для произвольного множества X (оно может быть даже пустым!) возьмем его булеан $\mathcal{P}(X)$ и выделим в нем систему подмножеств $\tau \subseteq \mathcal{P}(X)$, удовлетворяющую следующим аксиомам:

τ — это не кардинал!

T1 $\emptyset \in \tau, X \in \tau$;

T2 $(A \in \tau) \wedge (B \in \tau) \rightarrow (A \cap B \in \tau)$;

T3 $\tau' \subseteq \tau \rightarrow \cup \tau' \in \tau$.

Строго говоря, это — не аксиомы топологии, а определение некоего понятия в рамках теории ZF. Поэтому, как и в случае групп, мы часто сталкиваемся с тем, что мы видим одну и ту же топологию в разных формальных объектах. Как группа V_4 — это целый класс изоморфных групп, так и топология окружности — это целый класс изоморфных топологий. К этому нужно просто привыкнуть, как к Алгебре.

Система подмножеств τ множества X , удовлетворяющая свойствам T1–T3, называется **топологией** на множестве X , пара (X, τ) — **топологическим пространством**, а элементы τ — **открытыми множествами**, их дополнения — **замкнутыми**. Договоримся также, что под словом «точка» в отношении пространств мы будем понимать их элементы. Если открытое множество O содержит точку x , то O также называют *окрестностью* x . Кроме того, в соответствии с архетипом базового множества, само множество X мы также будем называть топологическим пространством, неявно подразумевая заданную на нем топологию.

Очевидно, что в силу аксиомы T1 в топологии всегда есть как минимум два открыто–замкнутых множества: пустое множество и все пространство X . Соответственно, есть и два крайних случая топологии: *дискретная*, которая включает все подмножества, т. е. $\tau = \mathcal{P}(X)$, и *тривиальная*, которая включает только X и \emptyset .

Дискретная топология носит такое название, т. к. в ней каждая точка сама по себе уже является открытым множеством ($x \in X \rightarrow \{x\} \in \tau$). Такие точки называются *изолированными*, а если точка не составляет открытое множество, то она называется *предельной*. Смысл же открытых множеств как раз и состоит в том, чтобы показывать, насколько точки пространства связаны друг с другом, т. е. насколько это пространство непрерывно.

[В скобках отметим, что определение предельной и изолированной точки дается для произвольного подмножества пространства. Приведенное выше определение — частный случай, когда подмножество совпадает со всем пространством. Более точно, если $x \in M \subseteq X$ и существует окрестность точки x , не пересекающаяся с $M \setminus \{x\}$, то x называется изолированной точкой M множества (в данной топологии), а если $x \in X$ такова, что любая ее окрестность пересекается с $M \setminus \{x\}$, то x называется **предельной точкой множества M** (в данной топологии). Предельная точка не обязана быть точкой множества M . Кроме того, важным понятием является **граничная точка**. Если x такова, что любая ее окрестность пересекается как с M , так и с $X \setminus M$, то x называется граничной точкой M . Множество всех граничных точек M называется **границей M** и обозначается ∂M .]

Топология τ называется **связной** (пространство X — связным), если не существует разбиения пространства $X = O_1 \sqcup O_2$ на непустые открытые множества (напомним, что \sqcup означает объединение непересекающихся множеств). Как видим, дискретная топология несвязна (и даже *вполне несвязна*), в то время как тривиальная топология связна.

Ясно, что всякое пустое пространство и пространство с одной точкой связно. В случае двух точек $\{0, 1\}$ топология $\{\emptyset, \{0\}, \{0, 1\}\}$ связна и называется связным двоеточием.¹ Обе точки в этой топологии являются предельными.

Вообще, понятия *связность*, *многосвязность*, *раздельность*, *отделимость*, *непрерывность*, *сепарабельность* и т.п. заслуживают того, чтобы объединить их под общим названием **архетип связности-непрерывности**.

Понятие непрерывности тесно связано с этим архетипом. Так, если мы рассмотрим линейно упорядоченное множество $(L, <)$, то мы можем определить на нем топологию интервалов. Для этого определим семейства левых и правых лучей:

$$Rl \Rightarrow \{(-\infty; x) \mid x \in L\}, \quad Rr \Rightarrow \{(x; +\infty) \mid x \in L\},$$

где $(-\infty; x) = \{y \in L \mid y < x\}$, $(x; +\infty) = \{y \in L \mid x < y\}$ — лучи. И далее,

$$\tau_{<} \Rightarrow \bigcap \{\tau \supset Rl \cup Rr\},$$

*Непрерывно
связано со
связностью
:)*

¹Вместо чисел 0 и 1, разумеется, можно взять любые два различных множества, но именно в случае с 0 и 1 получается забавная ситуация, когда фон Неймановское число 3 является топологией связного двоеточия на двойке.

т. е. $\tau_{<}$ есть наименьшая (в смысле вложения множеств) топология, содержащая все лучи L .

Здесь можно пару слов добавить о том, что такое база и предбаза топологии. **Базой** топологии называется семейство множеств, из которых всеми возможными объединениями можно получить все открытые множества. Ясно, что пересечение конечного числа базовых множеств должно быть представимо как объединение других базовых множеств, иначе мы не получим топологию. Например, все интервалы с рациональными концами в \mathbb{R} являются базой обычной евклидовой топологии, а все круги (без границы) — базой топологии на плоскости. **Предбаза** — это семейство множеств, все конечные пересечения которых образуют базу топологии. Множество $Rl \cup Rr$, определенное выше, образует предбазу интервальной топологии. Кроме того, заметим, что всякое открытое множество $\tau_{<}$ является объединением интервалов и/или лучей.²

Если (X, τ_X) — топологическое пространство и $Y \subseteq X$, $\tau_Y = \{O \cap Y \mid O \in \tau\}$, то τ_Y называется **индуцированной** в множество Y топологией пространства X , а (Y, τ_Y) — **подпространством** пространства (X, τ_X) .

Помимо подпространства можно получать и надпространства стандартными теоретико-множественными способами. Например, если X и Y — топологические пространства с топологиями τ_X и τ_Y , то на прямом произведении $X \times Y$ можно задать базу топологии как

$$\{O_1 \times O_2 \mid O_1 \in \tau_X, O_2 \in \tau_Y\}.$$

Аналогично строится топология для более высоких степеней и произведений множеств.

Ранее мы давали определение *непрерывного линейного порядка*. Для этого он должен быть, во-первых, плотным, а во-вторых, непрерывным, т. е., согласно теореме 1.13, всякое ограниченное множество должно иметь точную грань со стороны ограничения.

Напомним, что множество B плотно в л.у.м. $(L, <)$, если любой интервал в L пересекается с B . Аналогичное определение из топологии: множество B **плотно** в т.п. (X, τ) , если любая окрестность Ox произвольной точки $x \in X$ пересекается с B . Почти очевидно, что плотное в л.у.м множество B **плотно** в интервальной топологии, но не наоборот. Например, \mathbb{Z} не плотно в себе в смысле линейного порядка, однако оно плотно в себе в смысле топологии.

Теорема 5.1. Пусть $(L, <)$ — л.у.м. с интервальной топологией. L непрерывно в смысле порядка тогда и только тогда, когда оно связано в

²Мы не можем ограничиться только лишь интервалами, поскольку в случае существования $\min L$ или $\max L$ эти точки не будут покрываться никаким интервалом, а значит, L не будет открытым множеством в противоречии с определением топологии.

интервальной топологии.

Доказательство. Необходимость. Пусть L непрерывно. Предположим, что оно несвязно, т. е. $L = O_1 \sqcup O_2$. Пусть $a \in O_1$, $b \in O_2$, причем $a < b$. Пусть $A = O_1 \cap (a; b)$. Очевидно, что A — непустое открытое множество. Ясно, что $A < b$, поэтому существует $c = \sup A$ (в силу непрерывности) и $a < c \leq b$. Покажем, что c не может лежать ни в O_1 , ни в O_2 .

Пусть $c \in O_1$. Ясно, что $c \neq b$, т. к. $c \in O_1$, $b \in O_2$, поэтому $c \in (a; b)$. Так как O_1 открытое, найдется интервал $(c'; c'') \ni c$, целиком лежащий в O_1 , а интервал $(\max(c', a); \min(c'', b)) \ni c$ целиком лежит в A . Но тогда любая точка $d \in (c, \min(c'', b))$ (существует в силу плотности порядка) будет точкой A и при этом $d > \sup A$, противоречие.

Аналогично исключается случай $c \in O_2$ (найдется $d: A < d < c$). Получаем противоречие с тем, что $L = O_1 \sqcup O_2$.

Достаточность. Пусть интервальная топология связна. Допустим, что L не плотно, т. е. существует пара точек $a < b$, между которыми нет точек L . Тогда, полагая $O_1 = (-\infty; b)$, $O_2 = (a; +\infty)$, мы получаем разбиение L на два открытых множества в противоречии со связностью. Так что L плотно.

Далее, пусть $A \subseteq L$ ограничено сверху. При этом мы сразу можем исключить случаи, когда A конечно или когда выше A в L лежит только конечное число точек. В этом случае нахождение $\sup A$ не представляет проблемы.

Положим $O_1 = \bigcup_{x \in A} (-\infty; x)$, $O_2 = \bigcup_{x \notin O_1} (x; +\infty)$. Нетрудно видеть, что $O_1 \cap O_2 = \emptyset$, т. к. $O_1 < O_2$. Оба эти множества непустые в силу указанных предположений.

Пусть $S = L \setminus (O_1 \cup O_2)$. Ясно, что $O_1 < S < O_2$, если S непусто. Множество S не может содержать 2 и более точек, т. к. если $O_1 < s' < s''$, то $s'' \in (s', +\infty) \subseteq O_2$, т. е. $s'' \in O_2$. Таким образом, $S = \{s\}$ или пусто. Но пустым оно не может быть в силу связности интервальной топологии.

Так что $S = \{s\}$, следовательно, $s = \sup A$. Отсюда по теореме 1.13 получаем, что L непрерывно. \square

Итак, на примере линейно упорядоченных множеств (прежде всего \mathbb{R}) мы видим некую взаимосвязь между непрерывностью линейного порядка и связностью топологии, которая порождена этим порядком. Что убеждает нас в корректном соотношении двух определений.

Если, например, мы рассмотрим \mathbb{Q} с обычным порядком и интервальной топологией, то мы увидим, что оно плотно, но не связно. Действительно, дедекиндово сечение для $\sqrt{2}$

$$\mathbb{Q} = \{r \in \mathbb{Q} \mid r^2 < 2 \vee r < 0\} \cup \{r \in \mathbb{Q} \mid r^2 > 2 \wedge r > 0\}$$

разбивает \mathbb{Q} на два непустых открытых множества (они открыты как объединения соответствующих лучей).

При этом \mathbb{Q} не дискретно, поскольку любое одноточечное множество $\{r\}$ не содержит никакой интервал или луч.

Следующий пример: \mathbb{Z} . Это множество не только не связно в интервальной топологии, но и дискретно.³

Наконец, если рассмотреть произвольный ординал $\alpha > \omega$, то он, конечно, не связан в интервальной топологии, поскольку для любого $\beta + 1 < \alpha$ имеем $\alpha = (-\infty; \beta + 1) \cup (\beta; \alpha)$. В то же время, в нем есть изолированные (в смысле топологии) точки: если $\beta = \gamma + 1$, то $\{\beta\} = (\gamma; \beta + 1)$ — открытое множество. И есть предельные (в смысле топологии) точки: таковыми являются предельные ординалы. Иначе говоря, свойства предельности/изолированности в смысле топологии и в смысле теории ординалов эквивалентны в случае интервальной топологии.

Связность в таком общем виде позволяет нам обобщить (в силу доказанной теоремы) понятие непрерывности на очень широкий класс математических структур. В самом деле, рассмотрим комплексную плоскость \mathbb{C} , в которой в качестве базы топологии возьмем круги произвольного радиуса без объемлющей окружности

$$\{z \mid |z - z_0| < R\}, \quad z_0 \in \mathbb{C}, \quad R \in (0; +\infty).$$

Упражнение 5.2. То, что это действительно база (а не предбаза или что-то еще) топологии, мы предлагаем доказать читателю самостоятельно.

Упражнение 5.3. Докажите связность \mathbb{C} . Связность \mathbb{C} с такой (евклидовой) топологией достаточно очевидна. При этом \mathbb{C} невозможно линейно упорядочить таким образом, чтобы интервальная топология этого порядка совпадала бы с евклидовой топологией. Иначе говоря, мы не можем на плоскости совместить порядок и топологию так, как это делается на вещественной прямой.

Тем не менее, со связностью плоскости нет никаких проблем. Более того, если мы рассмотрим произвольную прямую на \mathbb{C} , то индуцированная топология на этой прямой будет изоморфна интервальной топологии \mathbb{R} . На самом деле, существует еще более глубокая взаимосвязь топологий плоскости и прямой через непрерывные функции.

Забывая временно классическое определение непрерывности вещественной функции, скажем, что функция $f : X \rightarrow Y$, где X и Y — топологические пространства, **непрерывна**, если прообраз относительно данной функции открытого в Y множества открыт в X :

$$\forall B \in \tau_Y : f^{-1}B \in \tau_X.$$

³Заметим, что если \mathbb{Z} рассматривать как подмножество \mathbb{R} , то оно еще и не плотно в \mathbb{R} , хотя и остается плотным в себе просто потому, что всякое топологическое пространство плотно в себе.

Нетрудно показать, что классическое определение непрерывности эквивалентно приведенному выше в случае действительной прямой или комплексной плоскости.⁴

Непрерывная функция с точки зрения топологии обладает следующими свойствами:

1. сохраняет связные множества, т. е. если A связно, то и образ fA связан;
2. график непрерывной функции связан в топологии прямого произведения $X \times Y$, если само оно связно.

[Интересно, что связность графика функции не обеспечивает ее непрерывность. Достаточно рассмотреть график функции $y = \sin(1/x)$ if $x > 0$, else 0. Он связан на плоскости, но функция терпит разрыв в точке 0.]

Второе свойство подводит нас к еще одной разновидности связности пространств, а именно — линейной связности.

Пусть $\Gamma = \{(x, y) \mid y = f(x)\}$ — график непрерывной функции f на плоскости \mathbb{C} , причем f определена для всех вещественных x . Можно определить новую функцию $g(x) = (x, f(x))$, которая будет действовать из \mathbb{R} в \mathbb{C} , «вычерчивая» на плоскости \mathbb{C} график Γ . При этом g непрерывна, а ее область значений связна. Тем самым, мы видим, что не только любая прямая, но и любая *непрерывная кривая (линия)* на плоскости будет связной. От этого свойства легко перейти к линейной связности.

Пусть $g : [0; 1] \rightarrow X$ ($[0; 1] \subset \mathbb{R}$) — непрерывная функция со значениями в топологическом пространстве X . Пусть, кроме того, для точек $a, b \in X$ известно, что $g(0) = a$ и $g(1) = b$. В этом случае говорят, что точки a и b *линейно связаны*, или g есть *путь* из a в b . Если же такая функция g найдется для любой пары различных точек в X , то говорят, что пространство X **линейно связно**.

Таким образом, обобщая сюжет с произвольными прямыми на плоскости, мы переносим непрерывность \mathbb{R} (а точнее, отрезка $[0; 1]$) с помощью непрерывной функции в произвольное пространство X . И здесь мы снова можем увидеть теорему о том, что *линейная связность* (непрерывность) *влечет топологическую связность*. Действительно, если $X = O_1 \sqcup O_2$, то, выбирая $a \in O_1$, $b \in O_2$ и непрерывную функцию $g : [0; 1] \rightarrow X$ такую, что $g(0) = a$, $g(1) = b$, мы получим $[0; 1] = g^{-1}O_1 \sqcup g^{-1}O_2$, что невозможно в силу связности (и непрерывности) \mathbb{R} .

Обратное, к сожалению, неверно. Только в случае конечного множества X из его топологической связности следует линейная связность.

Тем не менее, мы видим, что все время ходим вокруг одного и того же **архетипа связности–непрерывности**, причем применяемого как к пространству, так и к функции. По большому счету, вся топология построена

⁴На самом деле, в случае любого метрического пространства.

именно вокруг этого архетипа. Сохранение структуры связности различных тел и фигур при непрерывных деформациях — основной инвариант топологии!

Рассмотрим множество $T = \mathbb{C}^X$ всех функций из X в \mathbb{C} . Определим ε -окрестность элемента $f \in T$ следующим образом:

$$O_{\varepsilon, x_1, \dots, x_n}(f) = \{g \in T \mid \max_{1 \leq k \leq n} |f(x_k) - g(x_k)| < \varepsilon\}.$$

Заметим, что в случае конечного X эти окрестности представляют собой кубики размерности $\|X\|$ с центром в точке f .

Окрестности вида $O_{\varepsilon, x_1, \dots, x_n}(f)$ определяют предбазу топологии на T . Эта топология называется **топологией поточечной сходимости**, т. к. в этой топологии замкнутое множество включает все поточечные пределы своих последовательностей. Действительно, если $f_n(x) \rightarrow f(x)$ при $n \rightarrow \infty$ в каждой

Упражнение 5.4. Докажите это! | точке $x \in X$, и все $f_n \in K$, где K — замкнутое в T множество, то и $f \in K$. Для этого нужно показать, что любая окрестность f пересекается с K .

Конструкцию \mathbb{C}^X можно обобщить до пространства Y^X , где Y — произвольное топологическое пространство. Для этого придется немного модифицировать определение окрестностей, определяющих слабую топологию на Y^X . Как это сделать — предлагаем разобраться читателю самостоятельно.

Упражнение 5.5. | Имея в основном дело с вещественными или комплексными числами, мы можем заметить, что хотя каждое открытое множество есть произвольное объединение базовых множеств, но в \mathbb{R} и в \mathbb{C} достаточно ограничиться лишь счетным объединением базовых множеств. Иначе говоря, топология в этих пространствах является *счетно порожденной*. Кроме того, саму базу топологии можно выбрать счетной, включая в нее интервалы и лучи только с рациональными концами. Понятно, что если база счетна, то любое открытое множество есть счетное объединение элементов базы.

Для «хороших» (метрических) топологических пространств наличие счетной базы равносильно тому, что это пространство содержит счетное всюду плотное множество. Соответственно, пространство X называется **сепарабельным**, если оно содержит счетное всюду плотное множество.

Так, \mathbb{R} сепарабельно, поскольку \mathbb{Q} плотно в нем.

Для топологии со счетной базой удобно определять согласованную с ней σ -алгебру (см. раздел 3.1.3). Действительно, минимальная σ -алгебра, содержащая счетную базу (или предбазу) топологии, будет полностью содержать всю топологию. То есть каждое открытое множество будет элементом данной σ -алгебры. Кроме того, и каждое замкнутое множество будет элементом такой σ -алгебры.

В общем случае, если σ -алгебра порождена всей топологией пространства, то такая σ -алгебра называется **борелевской**. Ясно, что гораздо удобнее в

качестве набора порождающих множеств выбирать что-то простое, как, например, интервалы в случае линейно упорядоченного множества. И если база топологии является счетной, то она же порождает и борелевскую σ -алгебру.

В частности, в Анализе большую роль играет σ -алгебра, порожденная интервалами действительной прямой.

Итак, мы теперь знаем о трех разных системах подмножеств, приводящих к нетривиальным математическим структурам. Ранее мы определили фильтры и ультрафильтры (с которыми даже успели поработать), а также алгебры и сигма-алгебры, а в данном разделе к ним добавилась топология. Проведем их небольшое сравнение в таблице 5.1

Таблица 5.1: Системы подмножеств

	Фильтр	Ультрафильтр (неглавный)	Топология	Алгебра множеств	Сигма-алгебра (борелевская)
включает всё множество X	✓	✓	✓	✓	✓
включает пустое множество	✗	✗	✓	✓	✓
включает объединение	любое	любое	любое	конечное	счетное
включает пересечение	конечное	конечное	конечное	конечное	счетное
дополнение к элементу	✗	✗	X и \emptyset , в несвязном еще варианты	✓	✓
монотонность	✓	✓	✗	✗	✗
особые приметы	критерий компактности	либо A либо $X \setminus A$	отвечает за непрерывность	является алгеброй	содержит в себе топологию, если ее база счетная

5.1.2 Непрерывность и группы

Линейная связность предоставляет замечательную возможность привлечь в топологию теорию групп, причем интуитивно понятным способом.

Речь идет о *фундаментальной группе* топологического пространства. «На пальцах» это можно объяснить так. Рассмотрим сначала комплексную плоскость и будем рисовать на ней произвольные замкнутые кривые, т. е. деформированные окружности. Любую такую кривую можно непрерывно стянуть в одну точку («скомкать»). И сколько бы мы ни наматывали витков по этой деформированной окружности, ситуация не изменится — всю петлю можно сжать до одной точки непрерывным образом. В этом смысле у нас все замкнутые кривые на плоскости будут эквивалентны точке. Или, иначе говоря, тривиальной группе $\{e\}$.

Теперь, если на плоскости поставить препятствие в виде закрашенного круга и при сжатии петель запретить протаскивать их по этому кругу, то какие-то петли все еще можно будет сжать в точку (это петли, не огибающие полностью данное препятствие), а какие-то уже нельзя. При этом разными становятся замкнутые пути, совершающие разное количество витков вокруг закрашенной области и/или в разном направлении обхода (по часовой стрелке или против).

Отсюда возникает естественное соответствие между замкнутыми путями на плоскости и целыми числами: 0 соответствует стягиваемым в точку путям, положительное n — путям, совершающим n оборотов против часовой стрелки вокруг закрашенной области, и отрицательное n — путям, совершающим $|n|$ оборотов по часовой стрелке вокруг закрашенной области.

Таким образом, плоскость с одной дыркой соответствует группе целых чисел. Если добавить вторую дырку, то мы получим группу $\mathbb{Z} \times \mathbb{Z}$ и т.д.

Итак, действуя более формально, скажем, что два пути $f : [0; 1] \rightarrow X$ и $g : [0; 1] \rightarrow X$ (термин **путь** означает непрерывность этих функций) **гомотопически эквивалентны (гомотопны)**, если существует непрерывное отображение

$$F(t, s) : [0; 1] \times [0; 1] \rightarrow X$$

такое, что $F(t, 0) \equiv f(t)$, $F(t, 1) \equiv g(t)$. При этом $F(t, s)$ называется **гомотопией**.

Определение можно обобщить следующим образом: два отображения $f, g : X \rightarrow Y$ гомотопны, если существует гомотопия $F(x, s) : X \times [0; 1] \rightarrow Y$ такая, что $F(x, 0) \equiv f(x)$ и $F(x, 1) \equiv g(x)$.

Неформально это значит, что можно одно отображение совместить с другим отображением непрерывным смещением в пределах обоих пространств. Здесь мы можем усмотреть аналогию с геометрическими преобразованиями, о которых шла речь в разделе 3.6.

Там же мы отмечали инвариант $\text{Inv}5$, что связность сохраняется при гомеоморфизмах. Стоит отметить, что **гомеоморфизм** пространств (или подмножеств) X и Y — это биекция $f : X \leftrightarrow Y$, непрерывная в обе стороны.

Важность того обстоятельства, что в определении гомеоморфизма обратная функция f^{-1} должна быть непрерывна, можно продемонстрировать на простом примере. Функция $f : [0; 1) \rightarrow S^1$, действующая по правилу

Упражнение 5.6. Почему f^{-1} разрывна? $f(t) = e^{2it\pi}$, непрерывно и взаимно однозначно отображает полуинтервал в окружность. Однако, обратная функция терпит разрыв в точке $1 \in \mathbb{C}$. Мы предлагаем читателю самостоятельно разобраться с этим примером, чтобы усилить свою топологическую интуицию.

Обобщением свойства гомеоморфности пространств является свойство **гомотопической эквивалентности** пространств X и Y . Оно состоит в том, что существуют две функции $f : X \rightarrow Y$ и $g : Y \rightarrow X$ такие, что композиция

$f \circ g$ гомотопна id_Y , а композиция $g \circ f$ гомотопна id_X . Ясно, что наличие гомеоморфизма обеспечивает гомотопическую эквивалентность.

Обратное утверждение, строго говоря, неверно, однако существует замечательное исключение: если n -мерное многообразие гомотопически эквивалентно n -мерной сфере, то оно ей гомеоморфно. Это утверждение носит название «Гипотеза Пуанкаре» и была доказана Григорием Перельманом в серии статей 2002–2003 годов.⁵



Жюль Анри
Пуанкаре

Для гомотопически эквивалентных пространств связность также является инвариантом.

Вернемся к нашим путям. Путь f называется **петлей**, если $f(0) = f(1)$. Рассмотрим т.н. *пунктированное* пространство (X, x_0) , т.е. пространство X с выделенной точкой $x_0 \in X$. В этом пространстве рассмотрим все петли, начинающиеся и заканчивающиеся в точке x_0 , т.е. $f(0) = x_0 = f(1)$. Петля **стягивается** в точку x_0 , если она гомотопна тривиальному отображению $[0; 1] \rightarrow \{x_0\}$.

Далее введем на этих петлях отношение эквивалентности: $f \sim g$, если эти петли гомотопны. Образует соответствующее фактор-множество

$$G = \{[f] \mid f : [0; 1] \rightarrow X, f \text{ — непрерывно, } f(0) = x_0 = f(1)\},$$

где $[f]$ — класс всех петель, гомотопных f .

Теперь введем на G операцию умножения по правилу $[f] \cdot [g] = [f \star g]$, где

$$(f \star g)(t) = \begin{cases} f(2t), & 0 \leq t \leq 1/2 \\ g(2t - 1), & 1/2 \leq t \leq 1 \end{cases}$$

Нетрудно проверить, что множество G с операцией \cdot является группой. | Упражнение 5.7.

Группа (G, \cdot) называется **фундаментальной группой** пунктированного пространства (X, x_0) и обозначается $\pi_1(X, x_0)$.

Можно показать, что если пространство *линейно связно*, то фундаментальная группа не зависит от выбора точки x_0 (с точностью до изоморфизма групп). В этом случае фундаментальная группа обозначается $\pi_1(X)$.

Примеры фундаментальных групп:

FG1 $\pi_1(\mathbb{R}^n) = \{e\};$

FG2 $\pi_1(S^1) = \mathbb{Z}$ (группа окружности);

FG3 $\pi_1(T^2) = \mathbb{Z} \times \mathbb{Z}$ (T^2 — поверхность тора $S^1 \times S^1$);

⁵Точнее, Перельман доказал ее для случая $n = 3$, а остальные случаи были проверены ранее.

$$\text{FG4 } \pi_1(S^n) = \{e\} \quad (n \geq 2);$$

$$\text{FG5 } \pi_1(\mathbb{RP}^2) = \mathbb{Z}_2;$$

Фундаментальная группа также является инвариантом гомотопической эквивалентности пространств (и гомеоморфизма пространств). Поэтому, в частности, если известно, что фундаментальные группы пространств не изоморфны, то такие пространства не могут быть гомеоморфны. Например, по этой причине тор не гомеоморфен сфере, т.е. тор невозможно превратить путем деформации в сферу, не разрезая его.

У понятия фундаментальная группа существует естественное обобщение: **гомотопическая группа**. Если фундаментальная группа строится на базе линейных путей (отображений отрезка $[0; 1]$ в X), то гомотопическая — на базе отображений n -мерного куба в X . При этом композиция отображений производится по первой координате:

$$f \star g = \begin{cases} f(2t_1, t_2, \dots, t_n) & 0 \leq t_1 \leq 1/2, \\ g(2t_1 - 1, t_2, \dots, t_n) & 1/2 \leq t_1 \leq 1, \end{cases}$$

а склеивание концов отрезка заменяется стягиванием в точку поверхности куба, т.е. $f\partial[0; 1]^n \equiv x_0$.

Комментарий 18.

Читая известное произведение Стивена Хокинга [Вселенная Стивена Хокинга. — М.: АСТ, 2019], я обратил внимание на такую, казалось бы, безобидную сентенцию: «Поскольку поверхность области схлопывается до нуля, то это верно и в отношении объема». Речь идет о теореме Пенроуза о коллапсирующей звезде, но потом идея обобщается на любой коллапсирующий (в будущем или в прошлом) объект, в том числе Вселенную, так что «любая модель расширяющейся Вселенной фридмановского типа должна была начаться с сингулярности», т.е. Большого взрыва. Однако же в топологии мы часто имеем дело с ситуацией, когда такой вывод несостоятелен. Например, стягивая границу круга (или квадрата) в точку, мы получаем сферу, а стягивая бесконечно удаленную границу плоскости, мы получаем сферу Римана. Быть может, и разбегание галактик обусловлено тем, что «кто-то» проткнул «дыру» в трехмерной сфере, в результате чего она стала расплзаться, превращаясь в трехмерный шар? А вся внутренность этого шара могла вообще не измениться в объеме.

Ясно, что фундаментальная группа есть частный случай гомотопической группы при $n = 1$.

Высшие гомотопические группы обозначаются $\pi_n(X, x_0)$. Определение фундаментальной группы было дано основателем топологии Анри Пуанкаре, а высшие гомотопические группы были введены Витольдом Гуревичем.

Заметим одну характерную особенность определения гомотопических групп. В случае $n = 1$ они строятся с помощью отрезка $[0; 1]$ со склееными концами (склейка происходит за счет использования петель). Но при этом граница отрезка в своем родном пространстве \mathbb{R} состоит из двух точек $\{0, 1\}$ и не является связной. В случае же $n > 1$ группа строится с помощью гиперкуба со стянутой в точку поверхностью. Поверхность n -мерного куба гомеоморфна сфере S^{n-1} , которая уже является связным т.п. Такое существенное с точки зрения топологии отличие должно каким-то образом сказываться на свойствах гомотопических групп. И действительно, все высшие гомотопические группы являются абелевыми, в то время как фундаментальная группа не обязана быть коммутативной. В большинстве сложных случаев именно так и происходит.

Тем не менее, несмотря на такое замечательное свойство высших гомотопических групп (абелевы группы в алгебре заметно проще и не так разнообразны, как группы вообще — достаточно вспомнить группы перестановок), вычисление конкретных групп (даже для многомерных сфер) часто является очень трудной задачей.

Гомотопические группы можно определять несколько иначе.⁶ А именно, вместо отображений n -мерного куба в т.п. X можно рассматривать отображение n -мерной сферы в X . Так, отрезок $[0; 1]$ заменяется на окружность S^1 (т.е. его концы склеиваются не только в X , но уже в области определения отображений), квадрат $[0; 1]^2$ — на сферу S^2 (граница квадрата стягивается в точку), и т.д. При этом точке x_0 пространства X будет соответствовать некоторая выделенная точка на этой сфере (поскольку мы уже заранее склеили границу в точку). Соответственно, если раньше мы рисовали петлю сразу в X , то теперь мы имеем заранее заготовленную петлю с помеченной точкой, и просто приклеиваем ее к X так, чтобы точка на петле совпала с точкой x_0 .

Такое определение становится удобным, когда мы изучаем гомотопические группы сфер, т.е. $\pi_k(S^n)$ (точку x_0 здесь можно пропустить в силу линейной связности сферы). При $k < n$ у нас имеется естественное вложение сферы меньшей размерности в сферу большей размерности (например, окружность на обычной сфере): $S^k \hookrightarrow S^n$, и в этом случае нетрудно показать, что все такие отображения гомотопны тривиальному, переводящему сферу S^k в точку, т.е. все сферы S^k на поверхности сферы S^n стягиваются. Это значит, что $\pi_k(S^n) = \{e\}$ при $k < n$.

При $k = n$ имеем $\pi_n(S^n) = \mathbb{Z}$. Чтобы представить себе, каким способом можно «наматывать» одну сферу S^2 на другую сферу S^2 , можно представить мяч, положенный в пластиковый пакет. Причем одна точка пакета приклеена к одной точке на мяче. Этот пакет можно несколько раз намотать на мяч в ту

⁶Именно такой подход используется в видеолекциях д.ф.-м.н. А. Савватеева «Геометрия и группы» [hqTB9PTKvXU] и в классической книге по гомотопической топологии [95].

или другую сторону относительно выделенной на поверхности мяча точки, откуда и возникает группа \mathbb{Z} . Если пакет запечатать, то он будет гомотопичен сфере S^2 .

При $k > n$ начинаются проблемы. Но для малых значений k, n эти группы вычислены, их можно найти, например, [здесь](https://en.wikipedia.org/wiki/Homotopy_groups_of_spheres) (https://en.wikipedia.org/wiki/Homotopy_groups_of_spheres).

Так, в случае $\pi_3(S^2)$ снова возникает \mathbb{Z} , порожденная расслоением Хопфа, которое нам встречалось при изучении проективной геометрии в разделе 3.6.5. Расслоение Хопфа проецирует сферу S^3 на S^2 так, что S^3 представляется как семейство параллельных окружностей, каждая из которых проецируется в точку на сфере S^2 .

Если мы снова вернемся к кубам со стянутой в точку поверхностью, то S^3 можно представить как куб $[0; 1]^3$, а S^2 — как квадрат $[0; 1]^2$. При этом вертикальное проецирование куба на его нижнюю грань будет представлять собой расслоение куба на семейство вертикальных отрезков $\{(x, y)\} \times [0; 1]$ (которые соответствуют окружности S^1). Отсюда же видно, что такие отображения сводятся к отображению квадрата на квадрат, а их уже легко представить, как сложенный в несколько раз коврик, лежащий на полу. Количество сгибов данного коврика будет нумероваться целыми числами, и мы вновь имеем дело в группой \mathbb{Z} .

Чтобы не сложилось впечатление, что в гомотопических группах всегда можно отделаться \mathbb{Z} или тривиальной группой, приведем несколько непростых примеров:

$$\begin{aligned}\pi_4(S^3) &= \mathbb{Z}_2, & \pi_{10}(S^2) &= \mathbb{Z}_{15}, & \pi_7(S^4) &= \mathbb{Z} \times \mathbb{Z}_{12}, \\ \pi_{14}(S^2) &= \mathbb{Z}_{84} \times \mathbb{Z}_2^2, & \pi_{14}(S^4) &= \mathbb{Z}_{120} \times \mathbb{Z}_{12} \times \mathbb{Z}_2\end{aligned}$$

Теперь мы покажем на примере нескольких утверждений, как изложенная теория работает в топологии.

Дадим следующее определение. Если X — топологическое пространство и непустое $A \subseteq X$, кроме того, функция $f : X \rightarrow A$: 1) непрерывна, 2) $f|_A = \text{id}_A$, то f называется **ретракцией** пространства X , а множество A — его **ретрактом**. Отметим, что ретракция является своей собственной неподвижной точкой: $f \circ f = f$.⁷ Это свойство также используется как определение ретракции, а ретрактом в этом случае называется $\text{ran } f$.

Ретракт (если его правильно выбрать!) позволяет свести **Архетип редукции!** изучение сложного топологического пространства к более простому без существенной потери свойств.

Лемма 5.1. Пусть A — ретракт т.п. X . Тогда для любого $a \in A$ и любого $n \geq 1$ группа $\pi_n(A, a)$ является подгруппой группы $\pi_n(X, a)$.

⁷Имеется ввиду, что на основе f задается оператор на группе функций X^X , переводящий g в $f \circ g$. Именно для этого оператора f и будет неподвижной точкой.

Доказательство. Пусть $f : X \rightarrow A$ — ретракция. Пусть заданы непрерывные функции

$$p_1 : [0; 1]^n \rightarrow A, \quad p_2 : [0; 1]^n \rightarrow A, \quad p_1 \partial[0; 1]^n = p_2 \partial[0; 1]^n = a.$$

Пусть $[p_1]_A$ и $[p_2]_A$ — классы гомотопных относительно A функций, $[p_1]_X$ и $[p_2]_X$ — классы гомотопных относительно X функций.

Ясно, что $[p_1]_A \subseteq [p_1]_X$ и $[p_2]_A \subseteq [p_2]_X$, поскольку $A \subseteq X$, а кроме того, если p_1 гомотопна p_2 в X , то это же верно и в A , т. к. гомотопия, заданная в X , переводится в гомотопию в A с помощью ретракции f :

$$f(F(x, t)) : A \rightarrow A, \quad f(F(x, 0)) \equiv p_1(x), \quad f(F(x, 1)) \equiv p_2(x),$$

т. к. $f|_A = \text{id}_A$. Если же p_1 не гомотопна p_2 в A , то они также не гомотопны в X .

Это значит, что существует инъекция из $\pi_n(A, a)$ в $\pi_n(X, a)$, которая классу $[\]_A$ ставит в соответствие класс $[\]_X$.

Кроме того, очевидно, что данная инъекция сохраняет операцию умножения классов, т. е. мы имеем изоморфное вложение групп:

$$\pi_n(A, a) \hookrightarrow \pi_n(X, a),$$

что завершает доказательство. □

Отметим, что изоморфизм указанных групп здесь далеко не всегда возможен. Например, если X — это круг с двум дырками, то его фундаментальная группа равна \mathbb{Z}^2 . В то же время его ретрактом является ограничивающая его окружность с фундаментальной группой \mathbb{Z} .

Кроме того, во всяком т.п. любое одноточечное множество является его ретрактом. Но фундаментальной группой точки является группа $\{e\}$, в то время как у самого т.п. группы бывают разные.

«Неправильно
выбранный»
ретракт.

Основной смысл леммы состоит в том, что ретракт *не увеличивает* гомотопические группы.

Лемма 5.2. *Сфера S^n не является ретрактом шара B^{n+1} .*

Это следует непосредственно из того факта, что группа π_n шара тривиальна, а сферы — нет ($\pi_n(S^n) = \mathbb{Z}$).

Теорема 5.2 (Брауэра). *В конечномерном евклидовом пространстве любое непрерывное отображение замкнутого шара в себя имеет неподвижную точку.*

Доказательство. Предположим, что существует непрерывное отображение $f : B^{n+1} \rightarrow B^{n+1}$, не имеющее неподвижной точки.

Далее определим отображение $g : B^{n+1} \rightarrow S^n$ следующим способом. Пусть $x \in B^{n+1}$. В силу отсутствия неподвижной точки $f(x) \neq x$. Это значит, что существует единственная прямая l_x , проходящая через точки x и $f(x)$. Прямая l_x пересекает сферу S^n в двух точках. Выберем в качестве $g(x)$ ту точку пересечения, которая находится на l_x со стороны точки x , так что на прямой имеем расположение точек $g(x) \leq x < f(x)$.

Легко видеть, что $g(x)$ непрерывна, кроме того, $g|_{S^n} = \text{id}$. т. е. g является ретракцией шара на сферу.

Но это противоречит предыдущей лемме. Следовательно, неподвижная точка у $f(x)$ существует! \square

Теорема Брауэра — одна из «жемчужин» топологии и математики вообще. У нее существует прямое и весьма красивое доказательство, основанное на раскраске точек n -мерного симплекса в n цветов, сводящее теорему Брауэра к лемме Шпернера из комбинаторной топологии. Доказательство можно найти во многих книгах, например, в [98].

Напомним, что с теоремой о неподвижной точке мы уже встречались в разделе 4.3.2, это теорема 4.20 Клини о неподвижной точке для главной нумерации. Кроме того, можно вспомнить о неподвижной точке дифференциального оператора — экспоненте, и о многих других случаях. Наконец, сама ретракция является своей собственной неподвижной точкой, как уже отмечалось выше. Таким образом, мы имеем полное право выделить **архетип неподвижной точки**.

Теорема 5.3. Пусть отображения $f, g : S^n \rightarrow S^n$ таковы, что

$$f = \text{id}, \quad g(x) \equiv x_0 \in S^n.$$

Тогда f и g не гомотопны на сфере (сфера не стягивается в свою точку).

Доказательство. Предположим, что существует гомотопия $F : S^n \times [0; 1] \rightarrow S^n$ такая, что $F(x, 0) \equiv x$, $F(x, 1) \equiv x_0$, $x \in S^n$. Множество $S^n \times [0; 1]$ можно представить как замунутый шар B^{n+1} , где параметр $t \in [0; 1]$ означает глубину сферического слоя $S^n \times \{t\}$ в этом шаре. Слой при $t = 1$ соответствует поверхности шара, а слой при $t = 0$ соответствует центру шара, т. к. отображается целиком в точку x_0 .

Пусть S_t^n — сфера радиуса t . Определим на шаре B^n непрерывное отображение

$$T(x) = -F(x/t, 1 - t) \quad (x \in S_t^n), \quad T(0) = -x_0.$$

То, что оно непрерывно, следует из непрерывности F . Кроме того, оно переводит внутренность шара на его поверхность, и каждую точку поверхности в ей

противоположную относительно центра сферы. Это значит, что отображение T не имеет неподвижных точек.

Но такая ситуация невозможна в силу теоремы Брауэра! □

Нужно подчеркнуть, что в \mathbb{R}^n , безусловно, сфера S^n стягивается в любую точку. В доказанной леме опровергается возможность стянуть сферу в точку, не выходя за ее границы (оставаясь на поверхности мяча).

5.1.3 Фильтры и пределы

В общей топологии, ввиду отсутствия возможности строить те или иные «естественные упорядочения» на топологическом пространстве, возникает необходимость заменить привычным в Анализе последовательностям и пределам. В связи с этим дается следующее определение.⁸

Точка $x \in X$ называется **предельной точкой фильтра** (базы фильтра) \mathcal{F} , заданного в X , если она является предельной точкой всех элементов этого фильтра (базы фильтра). Точка $x \in X$ называется **пределом фильтра** (базы фильтра) \mathcal{F} , если каждая ее окрестность $Ox \in \mathcal{F}$ (соответственно, принадлежит фильтру, порожденному базой). Множество всех пределов фильтра \mathcal{F} обозначается $\lim \mathcal{F}$. Если \mathcal{B} — база фильтра, то множество ее пределов обозначается $\lim \mathcal{B}$. Если множество пределов состоит из единственной точки $a \in X$, то также пишут, что $a = \lim \mathcal{F}$.

Например, проверьте, что для главного ультрафильтра \mathcal{U} , порожденного точкой x_0 , имеем $x_0 = \lim \mathcal{U}$.

Нетрудно видеть, что *каждая предельная точка ультрафильтра есть предел этого ультрафильтра*. Действительно, пусть \mathcal{U} — ультрафильтр и a — его предельная точка. Покажем, что $a \in \lim \mathcal{U}$. Пусть Oa — некоторая окрестность точки a . Тогда для любого элемента $A \in \mathcal{U}$ имеем $Oa \cap A \neq \emptyset$. Предположим, что $Oa \notin \mathcal{U}$, тогда по определению ультрафильтра $A = X \setminus Oa \in \mathcal{U}$. А это противоречит тому, что $Oa \cap A \neq \emptyset$. Следовательно, $Oa \in \mathcal{U}$.

Таким образом, наличие пределов и предельных точек фильтров — это топологические свойства фильтров. С другой стороны, множество всех окрестностей какой-либо точки топологического пространства является фильтром. И это уже «фильтровое» свойство топологии.

А фильтрованное пиво — это топологическое свойство? :)

Понятие предела фильтра позволяет обобщить предел функции на случай произвольного топологического пространства. В «обычных» пространствах Анализа мы привыкли считать, что функция $f(x)$ непрерывна, если в любой точке ее области определения существует предел $\lim_{x \rightarrow a} f(x)$. Но в произвольном т.п., как мы видели выше, непрерывность определяется исключительно

⁸Определение фильтра дано на стр. 158

в связи с топологией как сохранение открытости множеств (в обратную сторону). Фильтры, а точнее их базы, позволяют говорить о пределе функции, аналогичном «обычному».

Связь между фильтрами и понятием непрерывности выявляет следующая

Теорема 5.4. *Отображение $f : X \rightarrow Y$ непрерывно тогда и только тогда, когда для любой базы \mathcal{B} фильтра и ее образа $f\mathcal{B}$ осуществляется вложение пределов:*

$$f \lim \mathcal{B} \subseteq \lim f\mathcal{B}.$$

Оставляя без доказательства эту теорему, рассмотрим пример. Предположим, что $f : X \rightarrow Y$ непрерывна, и в качестве базы \mathcal{B} возьмем какую-нибудь базу фильтра, состоящего из всех окрестностей точки $x_0 \in X$ (при этом мы можем держать в уме систему интервалов, содержащих точку x_0 , и работать в \mathbb{R}). Очевидно, $x_0 \in \lim \mathcal{B}$.

Далее, образом базы фильтра в X при действии функции f будет база $\mathcal{B}_Y = f\mathcal{B}$ фильтра в Y . Пусть $y_0 = f(x_0)$. В силу приведенной теоремы $y_0 \in \lim \mathcal{B}_Y$. Это значит, что любая окрестность Oy_0 принадлежит фильтру, порожденному базой \mathcal{B}_Y . А это, в свою очередь означает (в силу непрерывности f в топологическом смысле), что множество $f^{-1}Oy_0$ содержит в себе элемент базы \mathcal{B} , т. е. некоторую окрестность Ox_0 . Иначе говоря, прообраз произвольной окрестности y_0 содержит некоторую окрестность x_0 . [Сравните: прообраз ε -окрестности y_0 содержит δ -окрестность x_0 .]

В таком случае мы можем сказать, что $y_0 = \lim_{x \rightarrow x_0} f(x)$.

Вообще, говорят, что y_0 является **пределом f по базе** (пределом вдоль фильтра) \mathcal{B} (обозначение: $y_0 = \lim_{\mathcal{B}} f(x)$), если для любой окрестности Oy_0 существует множество $B \in \mathcal{B}$ такое, что $fB \subseteq Oy_0$.

В более простых пространствах (как, например, \mathbb{R}^n) это определение упрощается до обычного определения из Анализа на языке ε - δ .

5.1.4 Компактность

Одним из центральных понятий в топологии является *компакт*. Топологическое пространство (или множество в нем) называется **компактным**, если из любого его открытого покрытия можно выделить конечное подпокрытие. То есть если $X \subseteq \bigcup_{\sigma} O_{\sigma}$, то существует такой конечный набор индексов $\sigma_1, \dots, \sigma_k$, что $X \subseteq \bigcup_i O_{\sigma_i}$.

Само по себе определение компакта простое и непонятное, однако чисто механически его себе можно представить. Пусть у нас есть некая фигура на плоской поверхности, которую мы пытаемся закрасить при помощи стрельбы из маркера для игры в пейнтбол. Каждый выстрел соответствует одному

из множеств O_σ , которое представляет собой пятно на поверхности с фигурой. Размер пятна может варьироваться в зависимости от угла и дальности стрельбы, тем не менее за конечное число выстрелов можно закрасить всю фигуру, если она ограничена.

Именно возможность остановки через конечное число шагов алгоритма «покрывания» множества открытыми подмножествами и является тем уникальным свойством, которым обладают компакты. Недаром это слово просочилось даже в матлогику и дало название теоремам о компактности ИВ и ИП.

По этим и многим другим причинам свойство **компактности** заслуживает того, чтобы его отнести к **архетипам**.

Примеры и свойства компактных пространств:

- К1 Замкнутое ограниченное множество в \mathbb{R}^n (в стандартной топологии) — компакт;
- К2 Конечные подмножества топологических пространств компактны;
- К3 Пространство Стоуна булевых алгебр компактно;
- К4 Пространство X компактно тогда и только тогда, когда каждый фильтр в X имеет предельную точку;
- К5 Образ компакта при непрерывном отображении — компакт;
- К6 Любая непрерывная на компакте вещественно-значная функция достигает своих \max и \min ;
- К7 Замкнутое подмножество компакта компактно;
- К8 (теорема Тихонова) Прямое произведение любого количества (не обязательно конечного) компактов компактно (в топологии произведения);
- К9 Метрическое т.п. компактно тогда и только тогда, когда любая последовательность в нем содержит сходящуюся подпоследовательность.

Последнее свойство обязывает нас перейти, наконец, к метрическим пространствам, спуститься, так сказать, с облаков на землю.

5.1.5 Метрика

Здесь мы впервые начинаем использовать тот класс функций, который принято называть **функционалами**. А именно, функции, действующие из пространства в числовое множество (чаще всего имеется в виду \mathbb{R}).

Пусть X — произвольное непустое множество, и функционал $d : X \times X \rightarrow \mathbb{R}$ удовлетворяет следующим условиям:

M1 $d(a, b) = 0$ тогда и только тогда, когда $a = b$ (тождество);

M2 $d(a, b) = d(b, a)$ (симметричность);

M3 $d(a, b) \leq d(a, c) + d(c, b)$ (неравенство треугольника).

В этом случае d называется **метрикой**.

Заметим, что метрика неотрицательна, поскольку $0 = d(a, a) \leq d(a, b) + d(b, a) = 2d(a, b)$. Более того, если неравенство треугольника заменить аксиомой M3' $d(a, b) \leq d(a, c) + d(b, c)$, то M1+M3' выводит M2.

Метрика порождает топологию. Действительно, множество открытых шаров

$$\mathcal{B} = \{B(a, r) \mid (x \in B(a, r) \leftrightarrow d(a, x) < r) \wedge (r > 0)\}$$

Упражнение 5.8. | представляет собой базу топологии. Проверьте, что любое конечное пересечение шаров можно представить как объединение меньших шаров.

Полученная таким образом топология называется *порожденной данной метрикой*.

Топологическое пространство с топологией, порожденной метрикой, называется **метризуемым пространством**. Обычно разделяют понятие метрического пространства (X, d) и метризуемого (X, τ) , где τ порождена метрикой. Но поскольку выше мы условились под пространством понимать все, что может быть представлено как топологическое пространство, метрическим пространством мы будем называть тройку (X, τ, d) , где топология τ порождена метрикой d .

Ясно, что одну и ту же топологию могут порождать разные метрики (они называются эквивалентными), поэтому выбросить из обозначения метрического пространства метрику нельзя.

Существует целый ряд теорем (метризационные теоремы), которые дают условия метризации топологии по некоторым ее внутренним свойствам.

Поскольку метрика добавляет в топологическое пространство вещественные числа, это, с одной стороны, упрощает работу с пространством и обогащает его свойствами, с другой стороны, как мы знаем, вещественные числа «не видят» глубже уровня $\omega + 1$ в сюрреальных числах, и потому скрывают многие тонкие вещи в топологии, невыполнимые в метрических пространствах.

Упражнение 5.9. | Возможно, если расширить понятие метрики на сюрреальные числа произвольного ранга, то удалось бы получить или переоткрыть целый пласт новых свойств топологических пространств.
Подумайте об этом на досуге :)

Тем не менее, даже с вещественной метрикой роль метрических пространств в математике велика. Как правило, мы имеем дело с метризу-

емыми пространствами, в которых топология порождается некоторой естественной метрикой, получаемой из структуры самого пространства.

Метрика $d(x, y) = 0$ if $x = y$ else 1, называется **дискретной** по вполне понятным причинам — она метризует дискретную топологию. Данная метрика использует из всех чисел только 0 и 1, поэтому на языке сюрреальных чисел можно сказать, что она имеет глубину детализации 1 (см. стр. 171).

С помощью метрики, порождающей топологию, легко переформулировать определение непрерывности отображения. Пусть $f : X \rightarrow Y$, где X и Y — метрические пространства с метриками, соответственно, d_X и d_Y . Тогда f непрерывна в точке x_0 , если

$$\forall \varepsilon > 0 \exists \delta > 0 \forall x \in X (d_X(x, x_0) < \delta \rightarrow d_Y(f(x), f(x_0)) < \varepsilon).$$

И f непрерывна (на X), если она непрерывна в каждой точке X .⁹

В метрических пространствах можно обобщить и многие другие понятия из Анализа. Например, x_0 есть предел последовательности $\{x_n\}$, если последовательность $\{d(x_n, x_0)\}$ сходится к нулю. Это же позволяет дать эквивалентное определение непрерывности функции в точке: если $x_n \rightarrow x_0$, то $f(x_n) \rightarrow f(x_0)$ для любой последовательности $\{x_n\}$.

Метрика позволяет определить *ограниченные подмножества* ($\sup d(x, y) < \infty$). Более того,

Лемма 5.3. *Для каждого метрического пространства (X, τ, d) существует метрика d' , ограниченная числом 1 и эквивалентная метрике d*

Доказательство. Достаточно взять

$$d'(x, y) = \min\{1, d(x, y)\}$$

□

Часто можно встретить метрику вида

$$d_p(x, y) = \left(\sum_{k=0}^{\infty} |x_k - y_k|^p \right)^{1/p},$$

где $x = \{x_k\}$, $y = \{y_k\}$, $p > 1$. В том, что это метрика, мы предлагаем удостовериться читателю самостоятельно. Подсказкой здесь служит *неравенство Гёльдера*:

$$\left| \sum_{k=0}^{\infty} x_k y_k \right| \leq \left(\sum_{k=0}^{\infty} |x_k|^p \right)^{1/p} \left(\sum_{k=0}^{\infty} |y_k|^q \right)^{1/q},$$

⁹Мы не говорим о подмножествах X , поскольку все определения для подмножеств сводятся к определениям для подпространств с индуцированной топологией.

где p и q — сопряженные показатели, т. е. $p^{-1} + q^{-1} = 1$.

Метрика d_p применима к последовательностям, суммируемым со степенью p . Следовательно, если взять в качестве X множество всех вещественных (или комплексных) последовательностей, суммируемых с показателем p , то на этом множестве d_p будет метрикой. Такое пространство обозначается ℓ^p и представляет собой пример *гильбертова пространства*.

Другой пример гильбертова пространства получается, если вместо последовательностей взять числовые функции, а вместо суммы — интеграл Лебега:¹⁰

$$D_p(f, g) = \left(\int_a^b |f(x) - g(x)|^p \right)^{1/p},$$

где функции f и g интегрируемы с показателем $p \geq 1$ на отрезке $[a; b]$ и принимают значения из \mathbb{R} или \mathbb{C} . Пространство таких функций с данной метрикой обозначается $L^p[a; b]$.

Если рассмотреть предельный случай ($p = \infty$), то получатся метрики

$$d_\infty(x, y) = \sup_k |x_k - y_k|, \quad D_\infty(f, g) = \sup_{[a; b]} |f(x) - g(x)|,$$

Упражнение 5.11. где предполагается ограниченность последовательностей $\{x_k\}$, $\{y_k\}$ и непрерывность функций f и g на отрезке $[a; b]$. Соответствующие пространства обозначаются ℓ^∞ и $C[a; b]$ (обозначение $L^\infty[a; b]$ используется для пространства измеримых и почти всюду ограниченных на отрезке $[a; b]$ функций с метрикой $\text{ess sup } |f(x) - g(x)|$.)

Наконец, если взять обычное векторное пространство \mathbb{R}^n или \mathbb{C}^n , то для его векторов можно определить метрику

$$d_p^n(x, y) = \left(\sum_{k=0}^{n-1} |x_k - y_k|^p \right)^{1/p},$$

которая при $p = 2$ превращается в стандартную **евклидову метрику**. И, таким образом, пространства \mathbb{R}^n и \mathbb{C}^n можно рассматривать как метрические топологические пространства с естественной метрикой и порожденной ею топологией.

При евклидовой метрике шары $B(x, r)$ представляют собой обычные геометрические шары.

Если же снова перейти к пределу при $p \rightarrow \infty$, то получим **метрику Манхэттена**:

$$d_\infty^n(x, y) = \max_{0 \leq k < n} |x_k - y_k|,$$

¹⁰Здесь неявно предполагается, что мы не различаем функции, отличающиеся на подмножестве меры ноль. Об этом — чуть позже.

в такой метрике шары станут геометрическими n -мерными кубами.

Комментарий 19.

Довелось мне работать в одной системе управления контентом, где мы помимо текстов занимались обчетом товаров. Чем больше у товара технических характеристик, выражаемых числом (с единицей измерения), тем больше о нем можно «сказать» арифметическими формулами! Беда была в том, что конструктор формул не позволял¹¹ составлять какие угодно выражения, включающие элементарные функции, можно было только складывать, умножать и возводить в степень, т. е. минимальный набор.

Вопрос: как посчитать максимум двух величин a и b при помощи только таких операций?

Ответ: вспоминаем предельную формулу $(|a|^p + |b|^p)^{1/p} \rightarrow \max\{|a|, |b|\}$ при $p \rightarrow +\infty$. Стало быть, если взять какое-то большое число p , то результат будет близкий к максимуму. При этом, чтобы избежать слишком больших чисел, можно нормировать исходные a и b на некоторое усредненное число (обычно из природы a и b известен их порядок), что позволит взять достаточно большие степени (мне удалось использовать $p = 100$ для числового типа `double word`).

Полученный таким способом ответ не является абсолютно точным, но весь фокус в том, что как правило точность выше 2 знаков после запятой и не используется в оценочных параметрах, а это значит, что приближенные методы вполне себе могут выдавать ответ с нужной нам точностью.

Позже, правда, пришла в голову более экономичная в вычислительном плане формула (и более точная):

$$\max\{a, b\} = \frac{a + b + ((a - b)^2)^{0.5}}{2}, \quad \min\{a, b\} = \frac{a + b - ((a - b)^2)^{0.5}}{2},$$

где $((a - b)^2)^{0.5}$ — это не что иное, как $|a - b|$.

К сожалению, оптимальные идеи не всегда приходят в голову первыми :)

Добавим, что т.п. X называется **хаусдорфовым**, если для любых двух различных точек $x_1, x_2 \in X$ существуют окрестности, *отделяющие* эти точки: $Ox_1 \cap Ox_2 = \emptyset$. Всякое метрическое пространство является хаусдорфовым, поскольку если $d(x_1, x_2) = r > 0$, то шары $B(x_1, r/2)$ и $B(x_2, r/2)$ отделяют точки x_1 и x_2 .

Метрическое пространство с метрикой d называется **полным**, если любая фундаментальная последовательность в нем имеет предел в этом пространстве. При этом последовательность $\{x_k\}$ **фундаментальна** (относительно

¹¹Позже эта проблема решилась тем, что мне дали возможность использовать язык программирования для обвеса продуктов.

данной метрики), если

$$\forall \varepsilon > 0 \exists N \forall n, m > N d(x_n, x_m) < \varepsilon,$$

т.е. с ростом номера N все члены хвоста последовательности становятся сколь угодно близкими.

Всякое метрическое пространство можно пополнить до полного. Таким способом, например, пополняется \mathbb{Q} до \mathbb{R} . Для этого в исходном пространстве берутся классы эквивалентных фундаментальных последовательностей, причем под эквивалентностью $\{x_k\} \sim \{y_k\}$ понимается условие $d(x_k, y_k) \rightarrow 0$. Эти-то классы и будут точками нового пространства, которое окажется полным.

Полнота метрического пространства — это еще одна ипостась архетипа **связности-непрерывности**.

Примерами полных пространств являются: \mathbb{R} , \mathbb{C} и вообще все конечномерные линейные пространства над этими полями. Пространства $L^\infty[a; b]$, $C[a; b]$, $L^p[a; b]$, ℓ^∞ и ℓ^p также являются полными.

Уже здесь, при определении метрических пространств мы начинаем подмечать такую особенность математических конструкций, когда мы, действуя по аналогии, обобщаем простые понятия (длину вектора) и вместе с тем поступаемся некоторой незначительной общностью. Например, при переходе к ω -векторам (последовательностям) мы требуем сходимости ряда или ограниченности, а при переходе к функциям — непрерывности или интегрируемости по Лебегу. Причем, в последнем случае приходится еще и ослаблять понятие равенства (отличающиеся на множестве меры ноль функции мы считаем равными).

Иначе говоря, такое «генеалогическое» обобщение конструкций требует определенных ограничений и дополнительных построений, правда, не слишком существенных (в физике, например, принято не обращать внимания на подобные ограничения). Тем не менее, всегда нужно иметь ввиду, что формальная система не прощает «забывчивости в мелочах» и может привести к различным антиномиям вроде парадокса об удвоении шара, неконструктивности и/или неполноте теории. Умение держать в голове подобные мелочи, аккуратно обходя формально-логические ловушки при получении результатов есть часть математической интуиции.

5.1.6 Норма и скалярное произведение

Следующий шаг, снижающий энтропию определения топологического пространства — введение *нормы*. Переход от метрики к норме существенно ограничивает нас тем, что теперь мы в качестве пространства рассматриваем только модуль над некоторым числовым полем, т.е. имеем дело с линейным пространством.

Пусть X — модуль над \mathbb{L} , где $\mathbb{L} = \mathbb{R}$ или $\mathbb{L} = \mathbb{C}$.

Нормой на X называется всякая числовая функция $N : X \rightarrow \mathbb{R}$, которая удовлетворяет условиям (аксиомам):

$$N1 \quad (N(x) = 0) \rightarrow (x = 0);$$

$$N2 \quad N(\lambda x) = |\lambda|N(x) \text{ (однородность);}$$

$$N3 \quad N(x + y) \leq N(x) + N(y) \text{ (неравенство треугольника),}$$

где $x, y \in X$. Линейное пространство с нормой называется **линейным нормированным пространством**.

Нужно сразу же отметить, что такое определение нормы не согласуется с евклидовой нормой, определенной нами для гауссовых и эйзенштейновских чисел. Евклидова норма, удовлетворяющая некоторым условиям, связанным с алгоритмом деления с остатком, была введена как квадрат нормы, обычно применяемой в качестве меры длины в евклидовом пространстве. В теории делимости это делается специально для того, чтобы значение нормы оставалось целым числом и вписывалось в теорию делимости натуральных чисел.

На самом деле, не так уж важно, считать ли нормой N или N^2 . В большинстве теорем одно на другое заменяется без потери строгости, в остальных же нужно лишь корректно переписать выкладки, связанные со свойством однородности. Кроме того, если выполнено $N^2(x + y) \leq N^2(x) + N^2(y)$, то, очевидно, выполнено и $N(x + y) \leq N(x) + N(y)$, т. е. неравенство треугольника для N^2 влечет неравенство треугольника для N . Но традиция есть традиция!

К трем аксиомам нормы можно добавить еще одну

$$N4 \quad N(xy) \leq N(x)N(y),$$

если X есть алгебра, т. е. допускает умножение векторов. В этом случае норма называется **субмультипликативной**. Такую норму можно встретить, например, при изучении матриц и операторов. Требование субмультипликативности обеспечивает непрерывность произведения векторов относительно нормы.

В соответствии с традицией будем обозначать норму $\|x\|$. Да, это совпадает с нашим обозначением мощности множества, впрочем, как и модуль числа обозначается аналогично порядковому типу. Но такое разночтение не приводит к коллизиям, поскольку из контекста обычно понятно, о чем идет речь.

Отметим, что $\|x\|$ как мощность *мультимножества* x вполне можно считать его нормой, если под суммой мультимножеств понимать их объединение, а под умножением на число — умножением кратностей элементов множества. В этом случае мощность будет удовлетворять аксиомам нормы, с той только разницей, что коэффициентами могут быть лишь натуральные числа (см.

раздел 1.1.8). В этом смысле можно сказать, что обозначение нормы и мощности согласованы.

Чего не
скажешь о
|·|

Некоторые простые свойства нормы:

$$\text{Norm1 } (\|x\| = 0) \leftrightarrow (x = 0);$$

$$\text{Norm2 } \|x\| \geq 0;$$

$$\text{Norm3 } \|x - y\| \text{ является метрикой на } X;$$

$$\text{Norm4 } |||x| - |y|| \leq \|x \pm y\| \leq \|x\| + \|y\|;$$

$$\text{Norm5 } \frac{\|x\|^2 + \|y\|^2 - \|x - y\|^2}{2\|x\|\|y\|} \in [-1; 1] \text{ (косинус угла между } x \text{ и } y).$$

Поскольку норма задает метрику, с ее помощью легко задать топологию на нормированном пространстве, рассматривая в качестве базы топологии открытые шары положительного радиуса r :

$$B(a, r) = \{x \in X \mid \|x - a\| < r\}.$$

С нормой тесно связано понятие **скалярного произведения**, аксиомы которого мы уже приводили на стр. 252.

Скалярное произведение $x \cdot y$ естественным образом задает норму:

$$\|x\| = \sqrt{x \cdot x},$$

а значит, задает метрику и топологию.

Упражнение 5.12. Нетрудно видеть, что аксиомы нормы в этом случае выводятся непосредственно из аксиом скалярного произведения.

Не всякая норма может быть задана скалярным произведением, но существует точный критерий такой возможности. Норма в произвольном нормированном пространстве порождается некоторым скалярным произведением тогда и только тогда, когда выполнено *тождество параллелограмма*:

$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2).$$

В случае вещественного пространства скалярное произведение можно определить через норму по формуле:

Это не
фамилия,
поэтому
пишем с
маленькой
буквы :)

$$x \cdot y = \left\| \frac{x + y}{2} \right\|^2 - \left\| \frac{x - y}{2} \right\|^2,$$

а в случае комплексного:

$$x \cdot y = \left\| \frac{x + y}{2} \right\|^2 - \left\| \frac{x - y}{2} \right\|^2 + i \left\| \frac{x + iy}{2} \right\|^2 - i \left\| \frac{x - iy}{2} \right\|^2$$

Этой формулой мы уже пользовались на странице 254.

Чаще всего норма и скалярное произведение задаются некоторым естественным образом одновременно, а именно, путем обобщения скалярного произведения в евклидовом пространстве. Например, в пространстве ℓ^p это

$$x \cdot y \Rightarrow \left(\sum_{k=0}^{\infty} (x_k \bar{y}_k)^p \right)^{1/p}, \quad \|x\| \Rightarrow \left(\sum_{k=0}^{\infty} |x_k|^p \right)^{1/p},$$

а в пространстве $L^p[a; b]$:

$$f \cdot g \Rightarrow \left(\int_a^b (f(x) \bar{g}(x))^p dx \right)^{1/p}, \quad \|f\| \Rightarrow \left(\int_a^b |f(x)|^p dx \right)^{1/p},$$

Нетрудно видеть, что эти нормы (и скалярные произведения) задают ранее определенные метрики d_p и D_p . Аналогично — для случая $p = \infty$.

Линейное нормированное пространство называется **банаховым**, если оно является полным метрическим пространством по метрике, порожденной нормой. Банаховы пространства — основной объект изучения функционального анализа. Кроме того, в них очень часто можно задать произведение векторов, удовлетворяющее требованию субмультипликативности нормы, тем самым получив **банахову алгебру**.

Настолько же, насколько недалеко от нормы отстоит скалярное произведение, гильбертово пространство отстоит от банахова: пространство называется **гильбертовым**, если это линейное пространство со скалярным произведением, полное в метрике, порожденной данным скалярным произведением.

Эти пространства хороши тем, что редко вводят в заблуждение нашу евклидову интуицию.

Банаховыми (и гильбертовыми) пространствами являются ℓ^p , L^p (включая случай $p = \infty$) и $C[a; b]$ с определенными выше нормами (скалярными произведениями).

Банаховыми алгебрами являются векторные пространства конечной размерности над полями \mathbb{R} и \mathbb{C} с евклидовой нормой и поточечным произведением $xy = (x_1 y_1, \dots, x_n y_n)$, которое не следует путать со скалярным и векторным произведением.

В банаховых алгебрах можно доказать многие соотношения, обобщающие таковые из Анализа. Например, мы можем определить экспоненту линейного оператора A

$$\exp(A) = E + A + \frac{1}{2}A^2 + \frac{1}{3!}A^3 + \dots$$

и доказать для нее предельное соотношение

$$\exp(A) = \lim_{n \rightarrow \infty} \left(E + \frac{1}{n}A \right)^n,$$

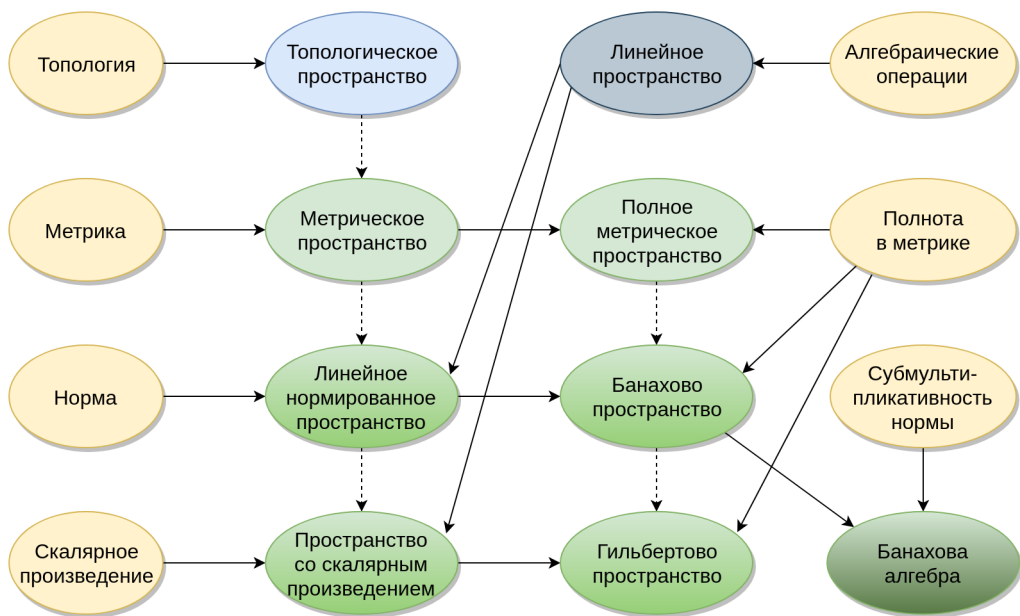


Рис. 5.1: Пространства (сплошные стрелки задают определение, пунктирные — переход от общего к частному).

а также определить тригонометрические функции

$$\cos(A) = \frac{\exp(iA) + \exp(-iA)}{2}, \quad \sin(A) = \frac{\exp(iA) - \exp(-iA)}{2i}$$

и вообще любые целые функции.

Следующая конструкция позволяет строить новые пространства (в том числе банаховы) из уже имеющих.

5.1.7 Ограниченные операторы

Ранее мы уже знакомились с линейными операторами на пространствах \mathbb{R}^n и \mathbb{C}^n . В общем случае, если у нас имеется два топологических пространства X и Y , то можно говорить о **непрерывных операторах** $X \rightarrow Y$, а если это линейные пространства, то — о **линейных**. Обычно предполагается, что оператор сохраняет структуру пространства, поэтому их также называют морфизмами пространств.

Функциональный анализ изучает, в основном, линейные непрерывные операторы. К ним относятся и операторы, заданные матрицей, и многие преобразования функций вроде преобразований Фурье и Лапласа.

Если на пространстве имеется норма или метрика, то можно потребовать от оператора, чтобы он переводил ограниченное множество в ограниченное. Такие операторы называются **ограниченными**. В случае линейного оператора определение ограниченности упрощается до требования

$$\|Ax\| \leq C\|x\|, \quad x \in X,$$

где C — некоторое положительное число, не зависящее от x .

Наименьшее такое число C называется нормой оператора A и обозначается $\|A\|$.

Нетрудно видеть, что линейный оператор между двумя нормированными пространствами непрерывен тогда и только тогда, когда он ограничен.

Множество всех ограниченных линейных операторов из пространства X в пространство Y обозначается $L(X, Y)$ (в случае $X = Y$ пишут $L(X)$). Если теперь определить сумму операторов и умножение на число

$$(A + B)(x) = A(x) + B(x), \quad (\alpha A)(x) = \alpha A(x),$$

то вместе с определенной выше нормой оператора пространство $L(X, Y)$ становится линейным нормированным пространством.

Если X — банахово пространство, то $L(X)$ — также банахово пространство.

Наконец, если для операторов из $L(X)$ банахова пространства X определить произведение как функциональную композицию

$$(AB)(x) = A(B(x)),$$

то $L(X)$, будучи банаховым пространством, превратится в банахову алгебру!

Поэтому, например, линейные операторы над \mathbb{R}^n образуют банахову алгебру (и, как следствие, это же относится к квадратным матрицам).

Вообще, стоит заметить, что линейные функции (в частности, операторы) играют очень важную роль везде, где требуются числовые модели — от математики до биологии и экономики. Дело в том, что линейная функция — это простейшая зависимость двух величин. С ней легко работать: анализировать, рассчитывать, проверять. Поэтому в тех случаях, когда сложную задачу удастся свести (хотя бы в первом приближении) к линейной зависимости, мы получаем в руки простой и хорошо разработанный математический аппарат, который легко погружается в компьютер.

В Анализе линейность проявляется, прежде всего, при разложении функции в ряд до первой степени. На линейности основано понятие дифференциала. Поэтому найти способ линеаризовать функцию или оператор в окрестности заданной точки — очень важная математическая задача. И она успешно решается в линейных нормированных пространствах.

Упражнение
5.13.

Проверьте,
что это действительно
норма.

Упражнение
5.14.

При изучении вариаций мы столкнемся с этой задачей в достаточно общем случае.

5.1.8 Псевдометрика

Аналогично
ослаблению
аксиом
геометрии.

У метрики существуют вариации и обобщения, ослабляющие в той или иной степени аксиомы метрики, приведенные выше.

Так, метрика называется **псевдометрикой**, если вместо аксиомы M1 требуется лишь

$$M1' \quad d(x, x) = 0,$$

т.е. нулевое расстояние допускается для неравных точек пространства.

Чтобы привести пример, как это может получиться, вспомним, что расстояние между двумя комплексными числами определяется скалярным произведением

$$d(z, z') = \sqrt{(z - z')\overline{(z - z')}} = \sqrt{(x - x')^2 + (y - y')^2},$$

где $z - z' = (x - x') + i(y - y')$.

Ранее мы определяли двойные числа вида $x + jy$, где $j^2 = -1$. Для них естественным определением скалярного произведения будет

$$(x + jy)(x' - jy'),$$

а для нормы

$$\|x + jy\| = \sqrt{|(x + jy)(x - jy)|} = \sqrt{|x^2 - y^2|}$$

и метрики

$$d(z, z') = \sqrt{|(x - x')^2 - (y - y')^2|}.$$

Отметим, что если перейти в новый базис преобразованием координат $\xi = x + y, \eta = x - y$ (матрица перехода $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$), то в нем метрика примет очень простой вид: $d(z, z') = \sqrt{|(\xi - \xi')(\eta - \eta')|}$.

Полученная метрика будет псевдометрикой, а соответствующая норма называется **псевдонормой**. Нетрудно заметить, что такая норма обращается в ноль на делителях нуля алгебры двойных чисел.

Тем не менее, псевдометрика (и псевдонорма) могут, с некоторыми ограничениями, использоваться для анализа вещественных или комплексных пространств.

Так, вместо формулы $x + iy = r(\cos(t) + i \sin(t))$ (где $r = \sqrt{x^2 + y^2}$, $t = \text{atan}(y/x)$) появляется формула

$$x + jy = Ir(\cosh(t) + j \sinh(t)),$$

где $r = \sqrt{|x^2 - y^2|}$, $t = \operatorname{atanh}(y/t)$, а коэффициент $I = \pm 1$ или $I = \pm j$ в зависимости от квадранта, в который попадает точка (x, y) . Угол $t = \pm \infty$ соответствует диагоналям $x = y$ и $x = -y$ на координатной плоскости. Эти диагонали делят плоскость на 4 квадранта, в каждом из которых существует как бы независимый геометрический мир.

Соответственно, движения в таком мире происходят своеобразно: в каждом квадранте имеется своя группа движений, причем, вращение осуществляется скольжением по гиперболическим окружностям, т. е. гиперболам. Мы предлагаем читателю в качестве развлечения поизучать движения плоскости в этом «странном» гиперболическом мире и построить таблицу композиций движений или их классификацию подобно тому, как мы это делали в евклидовой плоскости (см. раздел. 3.6.1). После чего нужно вспомнить о том, какие преобразования переводят окружности в гиперболы и построить алгебраический мостик между евклидовыми движениями и гиперболическими.

Упражнение 5.15.

Как и в случае кольца вычетов по простому модулю (где также есть делители нуля), в алгебре двойных чисел многочлен степени n может иметь корней больше, чем n (вплоть до n^2). Более подробно см. [85].

Мы не будем здесь углубляться в многочисленные «вариации на тему», т. к. это требует отдельной большой книги. Скажем только, что нестандартные метрики все чаще используются в различных задачах, причем свою историю они ведут уже как минимум столетие. Достаточно вспомнить геометрию Минковского, которая обобщает рассмотренную здесь геометрию двойных чисел на случай 4-х измерений,¹² и вообще любую псевдоевклидову метрику, которая в строгом смысле является не метрикой, а метрическим тензором.

Звучит как нестандартный анализ...

Другим обобщением геометрии двойных чисел является пространство Бервальда–Моора, где в подходящем базисе метрика выглядит довольно просто

$$d(x, y) = \sqrt[n]{(x_1 - y_1) \dots (x_n - y_n)},$$

однако она не является ни евклидовой, ни псевдоевклидовой метрикой. Подробнее об исследованиях на эту тему можно посмотреть в [78, 103].

Комментарий 20. О «метриках»

Отметим, что при анализе подобия текстов и других объектов, которые можно свести к числовым (чаще — натуральночисленным) векторам, в машинном обучении часто используется косинусная мера сходства, которая представляет собой вычисление косинуса угла между векторами x, y по формуле $\cos(\theta) = x \cdot y / \|x\| \|y\|$ (иногда добавляются весовые или корреляционные коэффициенты на пары координат у этих векторов). $\cos(\theta)$ нельзя считать метрикой,

¹²Скалярное произведение задается формулой $x_1 y_1 + x_2 y_2 + x_3 y_3 - x_4 y_4$.

т.к. он тем ближе к нулю, чем «ортогональнее» векторы, но поправка в виде $1 - \cos(\theta)$ для векторов с неотрицательными координатами уже будет метрикой (проверьте, что это так!).

Еще один вариант «метрики» или меры сходства можно найти в статистике. Пусть у нас имеются наблюдения $\bar{x} = (x_1, \dots, x_n)$ некоторой неизвестной случайной величины X , для которой мы проверяем гипотезу H_0 о том, что X имеет распределение с функцией $F(t) = F(t, \bar{\theta})$, где $\bar{\theta}$ — вектор параметров. Стандартный метод проверки — это определить сначала некоторую обобщающую статистику $T(\bar{x})$, а затем сравнить ее конкретное числовое значение t_0 на данной выборке с ее же значениями на произвольных выборках, полученных в рамках гипотезы H_0 . Для этого рассмотрим случайный вектор (ξ_1, \dots, ξ_n) , где все компоненты независимы и одинаково распределены по закону $F(t)$. В этом случае мы можем построить случайную величину $\tau = T(\xi_1, \dots, \xi_n)$, т.е. рассмотреть случайную статистику.

Как же сравнить выборочное значение t_0 с распределением величины τ ?

Предлагается следующий подход. Во-первых, помимо гипотезы H_0 рассматривается альтернативная гипотеза H_1 , которая тоже что-то утверждает про распределение X , например, что выборочная статистика t_0 будет сильно отклоняться вверх от основной части распределения τ . H_1 не обязана быть строгим отрицанием H_0 .

Разобьем область значений τ на две области $A_0 \sqcup A_1$, считая, что τ попадает в A_1 с незначительной вероятностью. Эта вероятность, обозначаемая α , называется уровнем значимости.¹³ Форма множества A_1 выбирается в зависимости от гипотезы H_1 . Так, если H_1 утверждает, что t_0 значимо уклоняется от τ в любую сторону, то $A_1 = (-\infty, -c) \cup (c, +\infty)$ (двусторонняя гипотеза), если же H_1 утверждает, что t_0 значимо отклонится вверх, то $B = (c, +\infty)$ (односторонняя гипотеза), и т.д. Соответственно, попадание выборочной статистики t_0 в множество A_1 отвергает гипотезу H_0 в пользу H_1 .

Попадание t_0 в то или иное множество можно расценивать по-разному в зависимости от того, как глубоко оно там оказалось. Вот на этом и строится метрика близости выборки \bar{x} и распределения $F(t)$. Рассмотрим вероятность $p = P\{|\tau| > |t_0|\}$ (в одностороннем случае: $p = P\{\tau > t_0\}$). Такая вероятность называется достигаемым уровнем значимости.

Ясно, что $|t_0| > c$ (для односторонней: $t_0 > c$) равносильно $p < \alpha$, так что при $p < \alpha$ гипотеза H_0 отвергается в пользу H_1 . Это значит, что близость p к нулю можно рассматривать как показатель расхождения выборки \bar{x} и распределения $F(t)$. Примерно так же ведет себя косинусная мера сходства. При этом величину $1 - p$ можно считать чем-то вроде расстояния между выборкой и теоретическим распределением. Это расстояние определяется гипотезой H_0 (распределением

¹³Обычно выбирается $\alpha = 0.05$.

$F(t)$), статистикой T и гипотезой H_1 (видом множества A_1).

В рассмотренных выше путях обобщения метрики (расстояния, схожести) можно усмотреть действие архетипа **неограниченного расширения**. Но сами метрики и их обобщения можно объединить общим архетипом **подобия**, который является родственным архетипам равенства и изоморфизма, но не входит ни в тот, ни в другой.

5.1.9 Мера и интеграл

Продолжим конструировать математические системы с помощью подмножеств. Выше мы ввели понятие σ -алгебры, в том числе борелевской, т. е. построенной на топологии.

Предположим теперь, что на множестве X имеется некоторая алгебра множеств \mathcal{A} , на которой задана функция $\mu : \mathcal{A} \rightarrow \mathbb{R} \cup \{\infty\}$, удовлетворяющая условиям (аксиомам):

Est modus in rebus... (мера должна быть во всем — Горацій)

MES1 $\mu(\emptyset) = 0$;

MES2 если $A = A_1 \sqcup A_2$, то $\mu(A) = \mu(A_1) + \mu(A_2)$ (аддитивность);

MES3 $\mu \geq 0$,

тогда функция μ называется **мерой** на X , а пара (X, μ) — пространством с мерой, элементы алгебры \mathcal{A} — **измеримыми множествами**.

Мера является обобщением геометрических понятий «длина», «площадь», «объем».

Заметим, что μ может принимать бесконечные значения, при этом арифметика с числом ∞ индуцируется из вычисления пределов, т. е. $\infty \pm x = \infty$ для любого конечного $x \in \mathbb{R}$. В данном случае мы не прибегаем ни к помощи арифметики ординалов, ни к нестандартному анализу или сюрреальным числам. Оперирование с ∞ здесь чисто мат.аналитическое, в том же смысле, в каком мы понимаем сходимость функции или интеграла к бесконечности.

Предполагая, что на множестве X задано отношение эквивалентности \sim , мы можем индуцировать меру μ на фактор-множество X/\sim , например, одним из следующих способов:

1. Если множество X конечно и факторизация согласована с мерой так, что внутри каждого класса $[a] \in X/\sim$ мера всех точек одинакова, то полагаем $\mu_\sim([a]) = \mu(\{a\})$.

2. В более общем случае рассмотрим множество \mathcal{D} разбиений, более крупных, чем X/\sim , и при этом измеримых относительно алгебры \mathcal{A} , затем построим разбиение $Z = \inf \mathcal{D}$ по формуле (1.10). Если разбиение Z окажется измеримым (что не

*Упражнение 5.16.
Когда Z измеримо?
Подумайте!*

всегда верно!), то с его помощью нетрудно индуцировать алгебру и меру в X/\sim . Для этого определим разбиение Z' на факторе X/\sim следующим образом:

$$Z' = \{ \{ [x] \in X/\sim \mid [x] \subseteq z \} \mid z \in Z \},$$

после чего зададим меру для всех его элементов $z' \in Z'$

$$\mu'(z') = \mu(\cup z'), \quad (5.1)$$

а затем продолжим ее на алгебру, порожденную разбиением Z' (элементами такой алгебры являются всевозможные конечные, в том числе пустые, объединения элементов Z' , а также их дополнения). Смысл формулы (5.1) в том, что для разбиения X отыскивается включающее его измеримое разбиение, которое диктует, какие классы фактора X/\sim нужно объединять в один новый класс (т. к. не обязательно сам фактор является измеримым разбиением), и далее этому новому классу приписывается его исходная мера.

Например, рассмотрим \mathbb{R} с мерой Лебега и скажем, что $x \sim y$, если их целые части равны. Тогда элементами фактор-множества будут интервалы вида $[n, n+1)$, где $n \in \mathbb{Z}$. Поскольку сами эти интервалы измеримы по Лебегу, множество $Z = \{ [n, n+1) \mid n \in \mathbb{Z} \}$ совпадает с \mathbb{R}/\sim , а Z' состоит из одноточечных множеств $\{ [n, n+1) \}$. Их мера по формуле (5.1) равна 1. Так как \mathbb{R}/\sim можно отождествить с \mathbb{Z} , то мы получаем индуцированную дискретную меру на \mathbb{Z} , которая каждой точке приписывает меру 1.

Если не принимать требование MES3, то функция μ называется **зарядом** (в этом случае потребуется различать $+\infty$ и $-\infty$, и оставить неопределенными операции типа $\infty - \infty$). Термин «заряд» был впервые введен А. Д. Александровым [100–102].

Если, наоборот, оставить MES3, а MES2 усилить: предположить, что \mathcal{A} является σ -алгеброй и выполняется условие

MES4 если $A = \bigsqcup_k A_k$, то $\mu(A) = \sum_k \mu(A_k)$ (счетная аддитивность),

то такая мера называется **счетно-аддитивной**.

Обычно, если не сказано обратное, считается, что мера счетно-аддитивна. Поэтому далее мы также будем предполагать, что рассматриваемые меры счетно-аддитивны, а кроме того, σ -алгебра является борелевской, т. е. порождена топологией.

Мера μ называется **конечной**, если $\mu(X) < \infty$. В частности, если $\mu(X) = 1$, то μ называется **вероятностной мерой** и обычно обозначается буквой P .

Мера называется **σ -конечной**, если все пространство можно представить как объединение не более чем счетного набора измеримых множеств конечной меры:

$$X = \bigcup_{k=0}^{\infty} A_k, \quad \forall k \mu(A_k) < \infty.$$

Меру можно задать на алгебре множеств, порожденной базой топологии, и корректно продолжить до борелевской σ -алгебры. Точнее, справедлива

Теорема 5.5 (Каратеодори). *Если на алгебре \mathcal{A} задана мера μ , то ее можно продолжить до счетно-аддитивной меры μ' на сигма-алгебре $\sigma(\mathcal{A})$.*

При этом, если мера μ является σ -конечной, то ее продолжение μ' единственное и также σ -конечное.

Мы не будем доказывать эту теорему, отсылая читателя к книге [77]. Приведем только способ определения меры μ' :

$$\mu'(A) = \inf \left\{ \sum_{k=0}^{\infty} \mu(E_k) \mid E_k \in \mathcal{A}, A \subseteq \bigcup_{k=0}^{\infty} E_k \right\}.$$

Эта т.н. *внешняя мера* произвольного подмножества $A \subseteq X$ является счетно-аддитивной мерой не на всем $\mathcal{P}(X)$, а на некоторой специальной сигма-алгебре. Чтобы ее описать, введем следующую функцию «расстояния» на $\mathcal{P}(X)$:

$$d(A, B) = \mu'(A \setminus B \cup B \setminus A).$$

Это — не метрика в смысле определения, однако свойства данной функции очень похожи на метрику:

1. $d(A, B) \geq 0$, $d(A, A) = 0$;
2. $d(A, B) = d(B, A)$, $d(A, B) = d(X \setminus A, X \setminus B)$;
3. $d(A, B) \leq d(A, C) + d(C, A)$;
4. $d(A \cap B, C \cap D) \leq d(A, C) + d(B, D)$;
5. $d(\bigcup_k A_k, \bigcap_k B_k) \leq \sum_k d(A_k, B_k)$;
6. $|\mu'(A) - \mu'(B)| \leq d(A, B)$.

Последнее свойство означает, что μ' *равномерно непрерывна* относительно функции расстояния d .

Далее, скажем, что множество $A \in \mathcal{P}(X)$ является **аппроксимируемым**, если оно является пределом последовательности множеств $\{A_k\}$ из алгебры \mathcal{A} в смысле расстояния d :

$$d(A_k, A) \rightarrow 0 \text{ при } k \rightarrow \infty.$$

Можно показать, что совокупность всех аппроксимируемых множеств образует сигма-алгебру $\sigma_{\mathcal{A}}$, очевидно, содержащую \mathcal{A} , причем на данной сигма-алгебре функция μ' является счетно-аддитивной мерой. Тем самым, она является счетно-аддитивной мерой и на $\sigma(\mathcal{A})$, поскольку $\sigma(\mathcal{A})$ содержится в любой сигма-алгебре, содержащей в себе алгебру \mathcal{A} .

В случае интервальной топологии на \mathbb{R} в качестве алгебры множеств нужно взять все возможные конечные суммы интервалов (не обязательно открытых)

$$I_1 \sqcup I_2 \sqcup \dots \sqcup I_n,$$

где I_k — это $(a; b)$, или $[a; b)$, или $(a; b]$, или $[a; b]$, причем $a \geq -\infty$, $b \leq +\infty$ и $a, b \in \mathbb{Q}$.

Легко видеть, что объединение и разность таких конструкций имеет такой же вид, следовательно, они образуют алгебру. Естественной мерой для интервалов является их длина, так что, полагая $\mu((a; b)) = \mu([a; b)) = \mu((a; b]) = \mu([a; b]) = b - a$, мы определяем меру, которая единственным способом продлевается на сигма-алгебру, порожденную интервальной топологией на \mathbb{R} . Такая мера называется **мерой Лебега**. Отметим, что счетная аддитивность меры Лебега выводится в предположении аксиомы счетного выбора AC_ω (теорема 4.6). А в предположении обычной аксиомы выбора AC доказывается существование неизмеримого по Лебегу множества (пример Витали в теореме 4.12). Наконец, известный парадокс Банаха–Тарского об удвоении шара путем разрезания его на несколько частей также опирается на неизмеримые множества: шар нужно разрезать на неизмеримые по Лебегу части.



Анри Леон
Лебег

Отметим также, что мера Лебега является единственной (с точностью до постоянного коэффициента) мерой на \mathbb{R} , инвариантной относительно сдвига (точнее, она инвариантна относительно группы движений). Данное утверждение является содержанием *теоремы Лебега*.

Если вместо длины интервала в качестве его меры использовать количество целых точек в нем, то мы получим дискретную меру, которая каждому синглету $\{n\}$ ($n \in \mathbb{Z}$) сопоставляет меру, равную 1. В случае такой меры измеримым может быть любое подмножество \mathbb{R} , его мера определяется количеством целых точек в нем. В этом смысле дискретная мера шире, чем мера Лебега.

Есть, с другой стороны, мера более узкая и более простая, чем мера Лебега.

Для любого подмножества $A \subseteq \mathbb{R}$ определим внешнюю меру как \inf всех мер интервальных покрытий

$$A \subseteq I_1 \sqcup \dots \sqcup I_n, \quad \mu^{ext}(A) = \sum_{k=1}^n \mu(I_k),$$

где $\mu(I_k)$ — длина интервала. Иначе говоря, мы покрываем A интервалами так, чтобы у них не было перекрытий и считаем меру этого покрытия, затем берем точную нижнюю грань таких «полумер». В итоге получаем оценку сверху для меры множества A .

Затем проделываем такую же процедуру, только интервалы вписываем внутрь множества A

$$A \supseteq I_1 \sqcup \dots \sqcup I_n, \quad \mu^{int}(A) = \sum_{k=1}^n \mu(I_k),$$

и получаем внутреннюю меру A .

Далее, если $\mu^{ext}(A) = \mu^{int}(A)$, то множество A называется **измеримым по Жордану**, а общее значение внешней и внутренней мер — **мерой Жордана** множества A .

Ясно, что всякое измеримое по Жордану множество также измеримо и по Лебегу, причем на базовых множествах (интервалах) эти меры совпадают. Однако, измеримые по Жордану множества не образуют сигма-алгебру.¹⁴ Например, множество $\mathbb{Q} \cap (0; 1)$ имеет внешнюю меру 1, а внутреннюю — 0. Вообще, любое счетное плотное в интервале множество неизмеримо по Жордану, но измеримо по Лебегу (в силу счетной аддитивности меры Лебега).

Тем не менее, мера Жордана интуитивно более понятна, поскольку реализует известный со времен Архимеда *метод исчерпывания*, и потому используется в Анализе для определения обычного интеграла Римана (об этом ниже). К тому же исторически она была придумана раньше (Пеано и Жорданом), чем мера Лебега.

Комментарий 21. Мера и метрика в Machine Learning.

В машинном обучении довольно часто встречается задача сравнения или кластеризации текстов. Под текстом можно понимать книги, статьи, а также, например, прайсовые названия товаров или публикации в интернете. Возможность обучить машину выделять кластеры подобных текстов или предсказывать их класс (тему) на основе обучающей выборки дает возможность решать ряд полезных задач, таких, как: автоматическое распределение новостей по темам, подбор похожих товаров, определение автора текста, выявление спама, составление ключевых слов для тем, и т.д.

Одним из методов, помогающих решить описанные задачи, является векторизация текстов в некотором многомерном пространстве. Делается это следующим способом. Для текстов производится так называемый стемминг или лемматизация слов. При этом текст превращается в последовательный набор (вектор) слов из имеющегося корпуса слов (словаря). Обычно слова в этом корпусе лишены окончаний, обезличены и приведены к единственному числу именительного падежа или начальной форме глагола. Это позволяет отождествить различные

¹⁴На самом деле они образуют *кольцо множеств*, т.е. систему подмножеств, замкнутую по операциям пересечения («умножение») и симметрической разности («сложение»). Для того, чтобы быть алгеброй множеств, не хватает только единицы — всего \mathbb{R} . Так что, если \mathbb{R} считать измеримым по Жордану, то данное семейство будет алгеброй.

формы одного и того же слова. После стемминга и, может быть, дополнительных действий по очистке текста, мы можем оцифровать текст.

Делается это частотными методами, одним из которых является мера TF-IDF. Итак, пусть у нас текст t представлен в виде набора слов (w_1, w_2, \dots, w_n) . Лучше всего в данном случае использовать представление не векторное, а виде мультимножества, т. е. как набор слов с кратностями: $t = \{k_1 \bullet w_1, \dots, k_m \bullet w_m\}$ (в МО это называется «мешком слов»). Далее положим

$$tf(w_i, t) \Rightarrow \frac{k_i}{k_1 + \dots + k_m},$$

т. е. $tf(w, t)$ — это относительная частота, с которой слово w встречается в данном тексте t .

Теперь мы можем взять используемый корпус слов и для каждого слова указать его частоту, вычисленную для данного текста t . Получится числовой вектор частот, соответствующий данному тексту. Ясно, что если у нас корпус слов содержит 10000 единиц, а текст только 100 слов, то 99% значений в этом векторе будут нулевые. Тем не менее, каждому тексту мы сопоставим вектор в 10000-мерном пространстве. А с этим уже можно работать методами линейной алгебры. Особенно, если использовать методы понижения размерности и выявления главных компонент.

Помимо того, что это числовой вектор, он же представляет собой еще и вероятностное распределение в пространстве слов, поскольку сумма весов $tf(w, t)$ для данного конкретного текста t равна 1. Таким образом, на тексты можно смотреть как на различные вероятностные меры на пространстве слов. Сказанное позволяет вводить различные метрики на текстах — от обычной евклидовой до расстояния между распределениями (Кульбака—Лейблера и т. п.)

Заметим, что в силу особенностей языка или специфики предметной области некоторые слова априори могут встречаться чаще других в любом из изучаемых текстов (например, слова «память» и «объем» в наборе текстов, описывающих компьютерные товары). Их можно выбрасывать еще на этапе стемминга (так поступают, скажем, с артиклями, союзами, междометиями и им подобными короткими словами), а можно нивелировать их влияние, если производить нормировку меры $tf(w, t)$ на частоту слов в исследуемом наборе текстов.

Итак, пусть у нас есть набор текстов $T = \{t_1, \dots, t_N\}$ (обычное множество). Для каждого слова w можно определить количество текстов, в которых оно встречается: $\#\{t \in T \mid w \in t\}$. Здесь мы уже не учитываем кратность вхождения слова в каждый текст. Положим теперь

$$idf(w, T) \Rightarrow \log \frac{N}{\#\{t \in T \mid w \in t\}}.$$

Так, если количество текстов равно $N = 10000000$, а слово w встречается в 1000 из них, то десятичный логарифм даст результат $idf = 4$, а если слово

встречается практически всюду, то idf будет близко к 1. Таким образом, idf можно считать мерой уникальности слов, которая также задана на пространстве всех слов из заданного корпуса слов.

Наконец, модифицируем исходную частотную меру tf текста следующим образом:

$$tfidf(w, t, T) \Rightarrow tf(w, t)idf(w, T).$$

Мы вновь для каждого текста имеем некоторую дискретную меру в пространстве слов, однако она уже не является вероятностной, причем если в каком-то тексте встречаются редкие слова, они получают более значительный вес, чем стандартные слова, и поэтому индивидуальность каждого текста усиливается в векторном пространстве распределений. Это значит, что если расстояние между текстами, определенное через $tfidf$ -частоты слов, мало, то они действительно очень похожи.

Следующий логический шаг в теории меры состоит в том, чтобы научиться перемножать пространства с мерами. Аналогично тому, как мы умеем умножать топологические пространства. Вообще, если есть два (или более) множеств, наделенных своими системами подмножеств, то их произведение наделяется системной подмножеств, получаемых как все попарные произведения исходных подмножеств:

$$(X, \tau) \times (Y, \eta) = (X \times Y, \tau\eta),$$

где $\tau\eta = \{a \times b \mid a \in \tau, b \in \eta\}$, а замыкание над $\tau\eta$ означает, что мы «допиливаем» это произведение до некоторого замкнутого состояния, получая структуру того же рода, что исходные τ и η .

Так, в случае топологических пространств мы берем прямые произведения всех открытых множеств, а затем достраиваем это семейство до топологии (т. е. считаем $\tau\eta$ за базу новой топологии). Аналогично, при произведении пространств с мерой мы составляем семейство всех прямых произведений измеримых множеств (исходных пространств), а затем замыкаем их до алгебры или сигма-алгебры. При этом мера каждого такого произведения $a \times b$ приравнивается к произведению мер сомножителей:

$$\mu_{X \times Y}(a \times b) \Rightarrow \mu_X(a)\mu_Y(b). \quad (5.2)$$

Как в случае топологий, так и в случае измеримых множеств достаточно определять произведение, отправляясь от базовой системы множеств: в случае топологии — от базы топологии, в случае меры — от кольца или алгебры множеств. В этом случае получается, что произведение баз есть снова база, а значит, замыкание до топологии/алгебры становится однозначным и конструктивно таким же, как в исходных пространствах.

Докажите, что результат произведения топологий и сигма-алгебр множеств не зависит от того, использовали ли мы сами топологии/сигма-алгебры при построении, или же только базу топологии/алгебру. То есть

Упражнение
5.18.

$$\overline{\tau(\beta_1)\tau(\beta_2)} = \tau(\beta_1\beta_2), \quad \overline{\sigma(\alpha_1)\sigma(\alpha_2)} = \sigma(\alpha_1\alpha_2),$$

где β_i — база топологии $\tau(\beta_i)$, α_i — алгебра, порождающая сигма-алгебру $\sigma(\alpha_i)$.

Мера, которая получается как продолжение меры, заданной способом (5.2), на минимальную алгебру (или сигма-алгебру) множеств, называется **произведением мер**. В случае, если исходные меры были счетно-аддитивными, то и их произведение будет счетно-аддитивным.

Таким способом можно задать меру Жордана и меру Лебега на пространстве \mathbb{R}^n . Это вполне согласуется с тем, например, что площадь прямоугольника $(a; b) \times (c; d)$ равна $(b - a)(d - c)$. Меры Жордана и Лебега на \mathbb{R}^n инвариантны относительно группы движений в этих пространствах.

При таком мультиплицированном порождении новых конструкций (однородных с исходными) мы одновременно видим и архетип **порождающего элемента**, и **рекурсию** (если требуется перепрыгивать через предельные ординалы при трансфинитном произведении пространств), и **неограниченное расширение**.

Как видим, несмотря на кажущуюся сложность рассматриваемых конструкций, все они укладываются в двухтактовую модель построения математических структур в рамках теории множеств, о которой мы говорили выше, хотя количество шагов в этом построении может быть огромным. Отчасти это напоминает построение физических макрообъектов из элементарных частиц. Порожденные структуры зачастую не только сильно сложнее порождающих, но и полностью меняют состав основных свойств, определяющих их «жизнедеятельность».

Почти...

Мера позволяет относиться к множествам не так щепетильно, как это принято в логике и теории множеств. Дело в том, что могут существовать непустые множества меры ноль.

Например, в случае меры Лебега точка, любой счетный набор точек и некоторые континуальные множества (например, «канторова гребенка») имеют нулевую меру. Присоединение и/или изъятие множеств меры ноль никак не меняет меру исходного множества.

Таким образом, мы можем предъявить еще один вариант равенства множеств (правда, ограниченный некоторым пространством с мерой): множества

A и B **почти равны**, если их симметрическая разность имеет меру ноль:

$$\mu(A \Delta B) = \mu(A \setminus B \cup B \setminus A) = 0.$$

Упражнение 5.19. Нетрудно доказать, что почти равенство является отношением эквивалентности.

Если мы озаботимся тем, чтобы значения меры рассматривать в \mathbb{R} , то у нас появится возможность говорить о бесконечно большой и бесконечно малой мере. В этом случае определение почти равенства станет еще интереснее. Скажем, что множества A и B почти равны, если мера их симметрической разности бесконечно мала по отношению к мере их суммы:

$$\mu(A \Delta B) = o(\mu(A \cup B)).$$

Здесь мы используем o -нотацию (которая ранее встречалась только в примерах), в стандартном анализе означающую, что величина слева бесконечно мала относительно величины справа при некотором предельном переходе. Однако по смыслу этого понятия ничто не мешает нам использовать его и в случае гипердействительных чисел.

Больше того, это же обозначение пригодится при сравнении последовательностей множеств: $\mu(A_n) = o(\mu(B_n))$, если $\mu(A_n)/\mu(B_n) \rightarrow 0$ при $n \rightarrow \infty$.

Обобщение понятие «почти» равенства позволяет, например, ввести понятие «почти всюду» для множеств любой меры, в том числе и бесконечной. Скажем, что формула $\varphi(x)$ выполняется **почти всюду** (относительно меры μ) на множестве X , если область ее истинности почти равна X .

Например, на множестве ω задаем равномерную меру, т. е. $\mu(n) = 1$. Тогда, равенство двух функций натурального числа $f, g : \omega \rightarrow \mathbb{R}$ почти всюду означает, что они равны всюду, за исключением множества конечной мощности (оно может быть и пустым), поскольку только любое конечное число бесконечно мало в сравнении с бесконечностью.

Кроме того, множества, почти равные ω , образуют фильтр, но не образуют ультрафильтр.

Еще пример: когда говорят, что функция почти всюду на отрезке равна нулю, это значит, что она равна нулю всюду, кроме точек множества меры ноль (обычно предполагается мера Лебега).

Термин «почти» является еще одной реализацией **архетипа подобия**.

Мера, как и фильтры, выводят нас на некоторое модифицированное представление о сходимости. Пусть $f, f_n : X \rightarrow Y$, где X — топологическое пространство с борелевской мерой μ , а Y — пространство с метрикой d .

Тогда $f_n \rightarrow f$ при $n \rightarrow \infty$ **μ -почти всюду** на X , если множество точек $x \in X$, где имеет место обычная поточечная сходимость $f_n(x) \rightarrow f(x)$ (в смысле

топологии), почти равно X . Такая сходимость еще называется сильной μ -сходимостью.

$f_n \rightarrow f$ на X **по мере** μ , если¹⁵

$$\forall \varepsilon > 0 \lim_{n \rightarrow \infty} \mu \{x \in X \mid d(f_n(x), f(x)) > \varepsilon\} = 0,$$

т. е. множество точек существенных отклонений от f сходится по мере к пустому множеству. Такая сходимость еще называется слабой. Обозначение сходимости по мере:

$$f_n \xrightarrow{\mu} f.$$

Легко видеть, что из сходимости почти всюду (сильной сходимости) следует сходимость по мере (слабая сходимость). | Упражнение 5.20.

Таким образом, нам известны следующие сходимости функций:

Conv1 Равномерная сходимость по метрике: $f_n \Rightarrow f$;

Conv2 Поточечная сходимость по топологии (метрике): $f_n(x) \rightarrow f(x)$;

Conv3 Сходимость почти всюду $f_n(x) \rightarrow f(x)$ (кроме множества меры 0);

Conv4 Сходимость по мере: $f_n \xrightarrow{\mu} f$;

Conv5 Сходимость вдоль фильтра (обобщение топологического предела).

При этом, очевидно, Conv1 \rightarrow Conv2 \rightarrow Conv3 \rightarrow Conv4, т. е. первые 4 вида предела расположены в порядке ослабления, а пятый можно отнести к любому из первых четырех, поскольку предел вдоль фильтра на ω — это просто предел при $n \rightarrow \infty$.

Об интегралах

В простейших прикладных задачах часто встречается необходимость просуммировать некоторые числовые показатели с весами. Например, показатели c_1, \dots, c_n с весами w_1, \dots, w_n суммируются так:

$$C = c_1 w_1 + \dots + c_n w_n,$$

в результате получается некий числовой функционал, определенный для измеримых событий, представленных вектором (c_1, \dots, c_n) , значимость компонент которого представлена весовыми коэффициентами w_1, \dots, w_n . При этом

¹⁵Конечно же, мы предполагаем здесь, что функция $d(f_n, f)(x)$ измерима относительно меры μ и меры Лебега в \mathbb{R} , иначе мы просто не сможем гарантировать существования мер у этого множества. Поэтому лучше всего сразу предполагать, что и на Y задана борелевская сигма-алгебра, так что функции f_n и f измеримы относительно борелевских сигма-алгебр в X и в Y .

обычно предполагается, что $w_k \geq 0$ и $\sum w_k = 1$. Тогда функционал C есть *средневзвешенное значение* выборки $\{c_k\}$.

Ситуации, где это применяется, могут быть самые разные: от голосований в совете директоров или на собрании жильцов многоквартирного дома (в первом случае весами являются доли в капитале, во втором — доли в общей площади дома) до теоретико-игровых ситуаций и вычисления состояний кварков.

С точки зрения теории меры мы имеем следующее. На множестве $X = \{1, \dots, n\}$ задана дискретная мера μ (каждой точке приписан ее вес), а также задана функция $f(k) = c_k$, определенная на X и принимающая значения в \mathbb{R} или \mathbb{C} . При этом на \mathbb{R} (\mathbb{C}) тоже имеется мера (Лебега), и прообраз любого измеримого подмножества будет измерим в X (поскольку там все 2^n подмножеств измеримы). В этом случае говорят, что функция f измерима.

Далее мы строим функционал, считающий среднее значение функции f относительно меры μ на X :

$$C(f) = \sum_{k=1}^n f(k) \mu(\{k\}).$$

Не правда ли, сильно напоминает определение интеграла Римана через интегральные суммы? Сравните:

$$\int_a^b f(x) dx = \lim_{n \rightarrow \infty} \sum_{k=1}^n f(x_{nk}) \Delta_{nk},$$

где Δ_{nk} — длина интервала I_{nk} , $[a; b] = I_{n1} \sqcup \dots \sqcup I_{nn}$ и $x_{nk} \in I_{nk}$. При этом $\max \Delta_{nk} \rightarrow 0$ при $n \rightarrow \infty$.

В случае дискретной конечной меры никакого предельного перехода не требуется, и функционал $C(f)$ является *интегралом f по мере μ* .

Дадим теперь общее определение. Во-первых, функция $f : X \rightarrow Y$ называется измеримой, если прообраз измеримого подмножества Y измерим в X (относительно заданных на этих пространствах алгебр и мер). Во-вторых, конечным измеримым разбиением множества $A \subseteq X$ называется разбиение $\alpha = \{A_1, \dots, A_n\}$ ($A = A_1 \sqcup \dots \sqcup A_n$), в котором все A_k измеримы (как следствие, A также измеримо). В-третьих, диаметром разбиения называется максимум мер его компонентов, т. е.

$$\text{diam}(\alpha) = \max_k \mu(A_k),$$

где μ — мера на X .

Теперь можно определить интеграл. Пусть $f : X \rightarrow \mathbb{R}$ (\mathbb{C}) измерима (относительно меры μ на X и меры Лебега на \mathbb{R}). Тогда число I называется

Сравните с определением непрерывности!

интегралом функции f по множеству A относительно меры μ , если

$$\forall \varepsilon > 0 \exists \delta > 0 \forall \alpha \left(\alpha \text{ — измеримое разбиение } A \right) \wedge (\text{diam}(\alpha) < \delta) \rightarrow \left| \sum_{x \in B \in \alpha} f(x) \mu(B) - I \right| < \varepsilon.$$

Иначе говоря, если при $\text{diam}(\alpha) \rightarrow 0$ средневзвешенное значение f относительно мер разбиения α стремится к числу I .

Интеграл I обозначается

$$\int_A f(x) d\mu.$$

В зависимости от меры и самого множества X такой интеграл может иметь много разных ипостасей. Например, как мы уже видели, в случае конечной дискретной меры он превращается в обычную сумму, в случае, когда мера задана на \mathbb{N} , он становится рядом (причем если эта мера равномерна, т. е. вес каждой точки $=1$, то это просто сумма или ряд значений функции f). Наконец, если $X = \mathbb{R}$, а мера μ является мерой Лебега, то перед нами **интеграл Лебега**, а в случае жордановой меры — **интеграл Римана** (он же — определенный интеграл).

Таким образом, теория меры объединяет в одно понятие суммы, ряды, взвешенные суммы, интегралы, а также характеристики случайных величин, поскольку случайная величина определяется как измеримая функция из пространства событий в \mathbb{R} . Естественным образом сюда же включаются интегралы от функций многих переменных, которые в некоторых случаях сводятся к повторным интегралам (**теорема Тонелли–Фубини**) и прекрасно работают в физике. А также интегралы по кривым в многомерных пространствах (т.н. контурные интегралы, хорошо известные нам в комплексном анализе и дифференциальной геометрии).

Заметим, кроме того, что если положить $f(x) \equiv 1$, то интеграл превращается в меру множества A , что позволяет вычислять площади и объемы различных фигур.

5.2 Преобразования пространств

Скажем пару слов о пользе различных интегральных преобразований. Предположим, мы изучаем пространство последовательностей с некоторыми «хорошими» ограничениями. Например, пространство ℓ^1 вещественных последовательностей, суммируемых с модулем. В этом случае каждой последовательности $x = (x_0, x_1, \dots)$ мы можем поставить в соответствие комплекснозначную функцию

$$F(x, z) = \sum_{k=0}^{\infty} z^k x_k.$$

Функция $F(x, z)$ называется **производящей функцией** последовательности x . К ее замечательным свойствам можно отнести, например, регулярность в круге $|z| < 1$. Заметим, что соответствие между этими функциями и последовательностями взаимно однозначное (с учетом регулярности в круге и сходимости в точке $z = 1$), т. к. по функции $F(x, z)$ можно найти ее коэффициенты по формуле $x_n = (d^n/dz^n)F(z)|_{z=0}/n!$.

Часто выражение $z^k x_k$ снабжается знаменателем $k!$. Это позволяет работать с быстро растущими последовательностями x_k за счет факториальной составляющей. Такие производящие функции называются *экспоненциальными* (от разложения в ряд экспоненты). Можно использовать и другие «ядра» при построении интегрирующего функционала — все определяется классом последовательностей x , которые мы изучаем.

Смысл такого интегрального преобразования заключается в том, что мы от практически произвольной последовательности переходим к очень удобной регулярной комплексной функции, тем самым привлекая мощный аппарат комплексного анализа. Таким образом, мы переходим от исходного пространства ℓ^1 (или ℓ^p и их аналогов) к пространству комплексных регулярных функций (определенных на всем \mathbb{C} или же какой-то его подобласти), получаем в нем некоторый результат, и затем возвращаемся обратно в исходное пространство с уже готовым результатом. Здесь работает **архетип редукции**, о котором мы уже говорили ранее.

Интеграл
здесь — это
сумма

Для того, чтобы переход от исходного пространства к вспомогательному и обратно был максимально формализован, полезно заметить некоторые свойства преобразования $x \mapsto F(x, z)$, т. е. построить миниатюрное исчисление:

1. $(\alpha x + \beta y) \mapsto \alpha F(x, z) + \beta F(y, z)$ (линейность);
2. $1 \mapsto \frac{1}{1-z}$;
3. $(q^{-k} \text{ for } k = 0, 1, \dots) \mapsto \frac{q}{q-z}$;
4. $(0, \dots, 0, x_0, x_1, \dots) \mapsto z^n F(x, z)$ (сдвиг на n позиций вправо);
5. $[z^n]F(x, z)F(y, z) = \sum_k x_k y_{n-k}$ (свертка).

В последней формуле нотация¹⁶ $[z^n]$ перед степенным рядом означает коэффициент при z^n . Естественно, для других ядер свойства 3–5 будут отличаться. Предлагаем читателю вывести аналогичные свойства для экспоненциальных производящих функций.

Упражнение
5.21.

В качестве примера рассмотрим последовательность Фибоначчи, которая описывается рекурсией:

$$x_1 = 1, \quad x_2 = 1, \quad x_n = x_{n-1} + x_{n-2} \quad (n \geq 3).$$

¹⁶Не путаем с нотацией Айверсона!

Что мы можем сделать с помощью производящих функций? Для начала заметим, что последовательность x можно продолжить влево с помощью $x_0 = x_{-1} = x_{-2} = 0$. Тогда рекурсия запишется в виде¹⁷

$$x_n = x_{n-1} + x_{n-2} + [n = 1] \quad (n \geq 0).$$

Теперь можем просуммировать от 0:

$$F(x, z) = \sum_{n=0}^{\infty} x_n z^n = \sum_{n=0}^{\infty} x_{n-1} z^n + \sum_{n=0}^{\infty} x_{n-2} z^n + z = zF(x, z) + z^2 F(x, z) + z,$$

откуда

$$F(x, z) = \frac{z}{1 - z - z^2} = -\frac{\phi_1/\sqrt{5}}{z - \phi_1} + \frac{\phi_2/\sqrt{5}}{z - \phi_2},$$

где

$$\phi_1 = \frac{-1 + \sqrt{5}}{2}, \quad \phi_2 = \frac{-1 - \sqrt{5}}{2}$$

— корни уравнения $z^2 + z - 1 = 0$. Отсюда, пользуясь свойством 3 и тем, что $\phi_1 \phi_2 = -1$, получаем, что

$$x_n = \frac{1}{\phi_1^n \sqrt{5}} - \frac{1}{\phi_2^n \sqrt{5}} = \frac{(-\phi_2)^n - (-\phi_1)^n}{\sqrt{5}}.$$

Наконец, заметим, что оба корня выражаются через золотое сечение $\phi = (1 + \sqrt{5})/2$, откуда окончательно имеем:

$$x_n = \frac{\phi^n - (-\phi)^{-n}}{2\phi - 1}.$$

Это последнее выражение называется *формулой Бине*. Отсюда следует, например, что асимптотика чисел Фибоначчи имеет вид

$$x_n = \frac{\phi^n}{\sqrt{5}}(1 + o(1))$$

при $n \rightarrow \infty$.

Преобразование последовательностей с помощью степенного ряда — не единственное удобное преобразование пространств. В общем случае интегральное преобразование T задается скалярным произведением ядра K этого преобразования и исходной функции f относительно меры μ на множестве Ω :

$$(Tf)(z) = \int_{\Omega} K(t, z) f(t) d\mu,$$

¹⁷ $[n = 1]$ — выражение, равное 1, когда $n = 1$, и равное 0 в противном случае.

где μ может быть мерой Лебега (и тогда $d\mu = dt$), или вероятностью P (и тогда $d\mu = dP$), или какой-либо другой мерой. Важно, чтобы указанный интеграл при этом существовал в каждой точке z из некоторого интервала в \mathbb{R} или области в \mathbb{C} (которые могут быть и бесконечными).

Так, рассмотренный выше пример преобразования последовательности x относится как раз к случаю дискретной меры μ_1 , равной единице в точках $n = 0, 1, 2, \dots$ и нулю вне этих точек:

$$Tx(z) = \int_{\mathbb{N}} z^t x_t d\mu_1 = \sum_{t=0}^{\infty} z^t x_t.$$

Здесь $K = z^t$. В случае $K = z^t/t!$ (правильнее было бы написать $K = z^t/\Gamma(t+1)$, т. к. t чаще обозначает непрерывную величину) получаем экспоненциальную производящую функцию.

Γ — это гамма-функция Эйлера.

Интегральное преобразование T является линейным оператором (см. выше свойство 1), действующим из пространства **оригиналов**, т. е. исходных функций от t , в пространство **изображений** (т. е. получаемых функций от аргумента z). Большинство интегральных преобразований обратимы, т. е. по изображению можно однозначно восстановить оригинал по формуле

$$f(t) = \int_{\Omega'} K^{-1}(z, t)(Tf)(z) d\mu',$$

где K^{-1} не означает обратную функцию или арифметическое обращение, в общем случае это ядро обратного интегрального преобразования. Для рассмотренного выше случая обратное преобразование будет таким:

$$x_n = \frac{1}{2\pi i} \oint_{\Gamma} z^{-n-1} F(x, z) dz,$$

где Γ — замкнутый контур, содержащий внутри своей области точку $z = 0$, что легко выводится из формулы Коши (см. раздел 2.4.7) для производной регулярной функции. Например, можно интегрировать по контуру $|z| = 1$, введя замену $z = e^{it}$, $t \in (-\pi; \pi)$, и тогда будем иметь

$$x_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(x, e^{it}) e^{-itn} dt.$$

Популярными интегральными преобразованиями являются следующие:¹⁸

- Преобразование Фурье ($K = e^{itu}$, $K^{-1} = e^{-itu}$), а также его действительная и мнимая части (синус- и косинус-преобразования);

¹⁸Мы не приводим здесь нормирующие множители для упрощения. На самом деле, преобразования, как правило, снабжены нормировками вроде $1/\sqrt{2\pi}$, связанными с «эстетикой» и физическими соображениями.

- Преобразование Лапласа ($K = e^{-tu}$, $K^{-1} = e^{ut}$);
- Вейвлет–преобразование (в качестве ядра выступает сопряженная вейвлет-функция с необходимой нормировкой).

Область интегрирования обычно является естественной для такого рода интегралов — либо прямая \mathbb{R} , либо ее положительный луч, либо некоторый интервал.

Наиболее полный список можно найти [по ссылке https://en.wikipedia.org/wiki/Integral_transform](https://en.wikipedia.org/wiki/Integral_transform). Рекомендуем также изучить источник [110].

5.3 Конечные разности и вариации

Пусть \mathbb{L} обозначет поле \mathbb{R} или \mathbb{C} . Пусть также A и K — некоторые линейные нормированные пространства над \mathbb{L} , причем A может быть снабжено мерой, а K — это как правило либо само поле \mathbb{L} , либо пространство линейных операторов в \mathbb{L}^n .

Обозначим через \mathcal{F} некоторое множество частичных функций из A в K . \mathcal{F} может не совпадать с K^A , кроме того, в нем могут быть функции с областью определения меньшей, чем A (частичные функции).

Условно мы договоримся считать (это не строгое определение!) функцию $f \in \mathcal{F}$:

1. **числовой последовательностью**, если $A = \mathbb{N}$ и $\text{dom}(f) = \mathbb{N}$, и $K = \mathbb{L}$;
2. **числовой функцией**, если $K = \mathbb{L}$;
3. **функционалом**, если A есть линейное пространство числовых функций вида $X \rightarrow \mathbb{L}$ (X — некоторое топологическое пространство) и $K = \mathbb{L}$;
4. **оператором**, если $f \in L(A, K)$.

Формально, функционал является также и числовой функцией в нашем определении, однако чаще всего предполагается, что у числовой функции аргументы пробегают \mathbb{L}^n , т. е. это функция от одного или нескольких числовых аргументов из поля \mathbb{L} .

5.3.1 Смещения и разности

Для каждого из четырех типов функции f мы можем единообразно определять ее **смещение** на заданное число (функцию/вектор) δ :

$$E^\delta[f](x) \doteq f(x + \delta),$$

где $E^\delta[f]$ определена, если x и $x + \delta$ не выходят за пределы $\text{dom}(f)$. В случае последовательности смещение δ может принимать только целые значения (чаще всего это просто единица).

На основе смещения вводится понятие **конечной разности**:

$$\Delta_\delta[f] \rightleftharpoons \Delta_\delta^1[f] = E^\delta[f] - f,$$

где единица указывает на то, что это определение может быть рекурсивно продолжено, а именно:

$$\Delta_\delta^2[f] \rightleftharpoons (\Delta_\delta^1 \circ \Delta_\delta^1)[f] = E^\delta[\Delta_\delta^1[f]] - \Delta_\delta^1[f] = f(x + 2\delta) - 2f(x + \delta) + f(x),$$

$$\Delta_\delta^{m+1}[f] \rightleftharpoons (\Delta_\delta^1 \circ \Delta_\delta^m)[f] = E^\delta[\Delta_\delta^m[f]] - \Delta_\delta^m[f].$$

В случае, когда $\delta = 1$ обозначения E^δ и Δ_δ^m упрощаются до E и Δ^m .

Заметим, что оператор Δ «съедает» константы-слагаемые, т.е. $\Delta_\delta[f(x) + a] = \Delta_\delta[f(x)]$, поэтому при его вычислении можно сразу же выбрасывать в записи f слагаемые, не зависящие от x . Кроме того, E^δ и Δ_δ являются линейными операторами, поскольку

$$E^\delta[\alpha f(x) + \beta g(x)] = \alpha E^\delta[f(x)] + \beta E^\delta[g(x)],$$

$$\Delta_\delta[\alpha f(x) + \beta g(x)] = \alpha \Delta_\delta f(x) + \beta \Delta_\delta g(x),$$

где $\alpha, \beta \in \mathbb{L}$.

5.3.2 Пространство смещений

Рассмотрим пространство

$$\text{Shift}(A, K) \rightleftharpoons \{\alpha_1 E^{\delta_1} + \dots + \alpha_k E^{\delta_k} \mid \alpha_i \in K, \delta_i \in A, k \geq 1\}$$

всех конечных линейных комбинаций всех возможных смещений для функций $f \in \mathcal{F}$. Элементы $\text{Shift}(A, K)$ производят над функциями f какие-то смещения (определяемые числами δ_i), а затем составляют их линейную комбинацию с коэффициентами из кольца K .

Для того, чтобы $\text{Shift}(A, K)$ действительно стало пространством, необходимо ввести на нем операции и топологию. В качестве сложения возьмем сложение операторов (оно очевидно из определения пространства), а в качестве умножения возьмем композицию операторов, обозначаемую \circ .

Нетрудно проверить, что $\text{Shift}(A, K)$ является алгеброй над кольцом K , поскольку для любых операторов $S_1, S_2 \in \text{Shift}(A, K)$ и любых $\alpha_1, \alpha_2 \in K$:

$$E^0 \circ S_1 = S_1,$$

$$(\alpha_1 + \alpha_2)S_1 = \alpha_1 S_1 + \alpha_2 S_1,$$

$$(\alpha_1 \alpha_2)S_1 = \alpha_1(\alpha_2 S_1),$$

$$\alpha_1(S_1 + S_2) = \alpha_1 S_1 + \alpha_1 S_2,$$

$$\alpha_1(S_1 \circ S_2) = (\alpha_1 S_1) \circ S_2,$$

и, учитывая, что сложение внутри A коммутативно, получаем бонусом коммутативность умножения в $\text{Shift}(A, K)$:

$$(E^\delta \circ E^\gamma)[f](x) = f(x + \gamma + \delta) = f(x + \delta + \gamma) = (E^\gamma \circ E^\delta)[f](x).$$

Наконец, в пространстве $\text{Shift}(A, K)$ выполняется экспоненциальный закон $E^\delta \circ E^\gamma = E^{\delta+\gamma}$, что позволяет степени композиций свести к алгебре в A .¹⁹

$$(E^\delta)^{\circ k} = E^{k\delta}.$$

В дальнейшем мы будем опускать символ \circ для упрощения записи.

Отметим также, что все операторы из $\text{Shift}(A, K)$ являются линейными, это свойство они наследуют от образующих операторов смещения E^δ .

Для операторов из $\text{Shift}(A, K)$ справедлива формула, называемая *биномом Ньютона*:

$$(S_1 + S_2)^n = \sum_{k=0}^n \binom{n}{k} S_1^k S_2^{n-k}, \quad (5.3)$$

Упражнение 5.22. | которую легко доказать, пользуясь коммутативностью операции \circ .

Заметим, что $\text{Shift}(A, K)$ является векторным пространством, если K — поле. При этом, все сдвиги E^δ образуют базис данного пространства,²⁰ а значит, элементы пространства можно записывать в виде последовательностей или функций $\alpha_k : A \rightarrow K$ с конечным носителем (!). Если мы дополнительно предположим, что поле K вкладывается в \mathbb{C} , то мы можем стандартным способом определить скалярное произведение на $\text{Shift}(A, K)$, тем самым определив на нем топологию. В этом смысле мы получаем право использовать термин *пространство* для структуры $\text{Shift}(A, K)$.

Наконец, для функций $f : A \rightarrow K$ можно ввести *sup-норму*²¹

$$\|f\| = \sup_{x \in A} |f(x)|,$$

если поле K вкладывается в \mathbb{C} , а все функции на A ограничены (это верно, например, в случае конечного A , либо если на функции f накладываются дополнительные ограничения вроде суммируемости или непрерывности).

¹⁹Здесь умножение на натуральное k — это просто другая запись суммы $\delta + \dots + \delta$ с k слагаемыми, т.е. умножение в A в данном случае не требуется.

²⁰По крайней мере, если в кольце K для любого конечного набора ненулевых чисел $\alpha_1, \dots, \alpha_n$ ($n \leq \|A\|$) найдутся числа f_1, \dots, f_n такие, что $\sum \alpha_i f_i \neq 0$.

²¹Если на A задана мера, то берется существенный супремум.

Давайте
также
опускать E
и δ , будет
еще проще!

Тогда все операторы из $\text{Shift}(A, K)$ окажутся непрерывными, поскольку смещение аргумента на δ у функций $f_n(x)$ не повлияет на сходимость к пределу $g(x)$, у которого аргумент также будет смещен:

$$\|f_n - g\| \rightarrow 0 \Rightarrow \|E^\delta[f_n - g]\| \rightarrow 0.$$

А так как линейный непрерывный оператор ограничен, появляется возможность определить и норму оператора $S \in \text{Shift}(A, K)$:

$$\|S\| \Leftarrow \sup_{\|f\|=1} \|Sf\|.$$

Следовательно, при некоторых не слишком сильных ограничениях мы можем считать, что пространство $\text{Shift}(A, K)$ есть пространство линейных непрерывных операторов.

К сожалению, метрику, порожденную нормой $\|S\|$, нельзя в общем случае считать полной, поскольку в $\text{Shift}(A, K)$ включены только конечные линейные комбинации сдвигов. Однако, компактность A (в той топологии, которую можно там задать естественным путем) может обеспечить существование пределов фундаментальных последовательностей операторов. И в этом случае $\text{Shift}(A, K)$ будет банаховой алгеброй.

Для наглядности наших построений рассмотрим простой пример. Пусть $A = \mathbb{R}$, $K = \mathbb{C}$, $f_r(x) = re^{ix}$. Будем рассматривать только функции f_r , $r \geq 0$. Ясно, что семейство $\mathcal{F} = \{f_r\}$ представляет собой множество концентрических окружностей на комплексной плоскости, радиус которых является их параметром.

В этом случае оператор E^δ представляет собой поворот на угол δ :

$$E^\delta f_r = re^{ix} e^{i\delta} = e^{i\delta} f_r.$$

Легко видеть также, что

$$(\alpha_1 E^{\delta_1} + \dots + \alpha_k E^{\delta_k}) f_r = (\alpha_1 e^{i\delta_1} + \dots + \alpha_k e^{i\delta_k}) f_r = \alpha E^\delta f_r,$$

где $\alpha e^{i\delta} = \alpha_1 e^{i\delta_1} + \dots + \alpha_k e^{i\delta_k}$, причем вещественное $\alpha \geq 0$.

Иначе говоря, все операторы пространства $\text{Shift}(A, K)$ в этом примере имеют очень простой вид: αE^δ . Это операторы поворота и растяжения.

Легко также найти соответствующие нормы:

$$\|f_r\| = r, \quad \|E^\delta\| = 1, \quad \|\alpha E^\delta\| = |\alpha|.$$

Эти нормы являются полными в силу полноты \mathbb{R} , а пространство операторов является банаховой алгеброй (оно изоморфно комплексной плоскости).

Теперь вернемся к оператору Δ . Легко видеть, что $\Delta_\delta = E^\delta - E^0$, кроме того, полагая $\Delta_\delta^0 = E^0$, имеем также формулу $E^\delta = \Delta_\delta + \Delta_\delta^0$. Отсюда по формуле (5.3) легко получить следующие выражения:

$$\Delta_\delta^n = \sum_{k=0}^n \binom{n}{k} (-1)^{n-k} E^{k\delta}$$

и

$$E^{n\delta} = \sum_{k=0}^n \binom{n}{k} \Delta_\delta^k.$$

Первая формула дает нам представление оператора конечных разностей произвольной степени через смещения, а вторая, обратная к первой, дает представление смещения через конечные разности:

$$f(x + n\delta) = \sum_{k=0}^n \binom{n}{k} \Delta_\delta^k [f](x). \quad (5.4)$$

5.3.3 Целочисленные разности

Существует одно удобное понятие, позволяющее упростить оперирование конечными разностями в том случае, когда δ не является бесконечно малым числом. А именно, величина

$$x^{\underline{n}} \equiv (x)(x - \delta)(x - 2\delta) \dots (x - (n - 1)\delta)$$

называется **убывающей (факториальной) степенью**, а величина

$$x^{\overline{n}} \equiv (x)(x + \delta)(x + 2\delta) \dots (x + (n - 1)\delta)$$

— **возрастающей (факториальной) степенью**, вместе именуемые **разностными (факториальными) степенями**.

Докажите следующие равенства:

Упражнение
5.23.

$$x^{\underline{n}} = (x - (n - 1)\delta)^{\overline{n}}, \quad x^{\overline{n}} = (x + (n - 1)\delta)^{\underline{n}},$$

$$\Delta_\delta[x^{\underline{n}}] = nx^{\underline{n-1}}\delta, \quad \Delta_\delta[x^{\overline{n}}] = nx^{\overline{n-1}}\delta.$$

Примечательно, что и тут при замене $\delta = dx$ мы снова получаем формулу для производной степени, поскольку разностные степени будут отличаться от обычной степени на бесконечно малое число.

Для отрицательной степени мы также можем определить разностные степени, пользуясь следующей схемой. Выпишем все смещения в ряд (предполагая, что мы действуем в области допустимых значений)

$$\dots \underbrace{(x - m\delta) \dots (x - 2\delta)(x - \delta)}_{(x - m\delta)^{\overline{m}}} \underbrace{(x)(x + \delta)(x + 2\delta) \dots (x + (n - 1)\delta)}_{x^{\overline{n}}} \dots$$

Мы видим, что выполняются равенства

$$(x - m\delta)^{\overline{m}} \cdot x^{\overline{n}} = (x - m\delta)^{\overline{m+n}}, \quad x^{\overline{m}} \cdot (x + n\delta)^{\underline{n}} = (x + n\delta)^{\overline{m+n}}, \quad (5.5)$$

которые остаются верными, как бы мы ни выбрали сам x , если только все участвующие в формуле окружающие его смещения корректно определены. Данное мультипликативное свойство степени позволяет корректно продолжить степень в отрицательную область, полагая $m = -n$ и выражая один из сомножителей в виде дроби:

$$x^{\overline{-n}} \Rightarrow \frac{1}{(x - n\delta)^{\overline{n}}} = \frac{1}{(x - \delta)^{\underline{n}}}$$

$$x^{\underline{-n}} \Rightarrow \frac{1}{(x + n\delta)^{\underline{n}}} = \frac{1}{(x + \delta)^{\overline{n}}}.$$

Здесь мы уже предполагаем, что A есть подмножество достаточно хорошего поля, чтобы разрешать деление. «Хорошесть» поля будет видна из дальнейшего.

Единство формул, как это обычно бывает в математике, диктует некоторое расхождение в визуальной форме определения — в случае отрицательной степени у нас появился сдвиг исходной переменной. Поэтому отрицательные степени можно также переписать в виде:

$$x^{\overline{-n}} = \frac{1}{(E^{-\delta}x)^{\underline{n}}}, \quad x^{\underline{-n}} = \frac{1}{(E^{\delta}x)^{\overline{n}}}.$$

Для отрицательных степеней сохраняется правило дифференцирования:

Упражнение 5.24.

$$\Delta_{\delta} x^{\overline{-n}} = -n x^{\overline{-n-1}} \delta, \quad \Delta_{\delta} x^{\underline{-n}} = -n x^{\underline{-n-1}} \delta$$

Отметим, что биномиальные коэффициенты также имеют прямое отношение к разностным степеням:

Упражнение 5.25.
А какие у них свойства при $\delta \neq 1$?

$$\binom{n}{k} = \frac{n^{\underline{k}}}{k!},$$

где сдвиг $\delta = 1$.

Интересно также, что если просуммировать все $n^{\underline{k}}/k!$ по $k \geq 0$, то мы получим величину 2^n , которая, как мы увидим в дальнейшем, является дискретным аналогом e^x . Сравните:

$$\sum_k \frac{n^{\underline{k}}}{k!} = 2^n, \quad \sum_k \frac{x^k}{k!} = e^x.$$

Оператор Δ вместе с разностными степенями позволяет в более компактной форме записывать многие арифметические тождества, указывающие одновременно на соответствующие им тождества из интегрального и дифференциального исчисления. Например, если функции f и g связаны равенством $g = \Delta_\delta f / \delta$, то для них (в области допустимых значений) выполняется тождество:

$$\sum_{k=0}^{n-1} g(x + k\delta) \delta = f(x + n\delta) - f(x), \quad (5.6)$$

так что если мы положим $a = x$, $b = x + n\delta$, то вышестоящее тождество можно записать в виде

$$\sum_a^b g(x) \delta = f(a) - f(b),$$

где знак суммы \sum_a^b означает суммирование с шагом δ на полуинтервале $[a, b)$. Стоит отметить, что в случае, когда множество A , область определения функций f и g , неупорядочено (например, это область в комплексной плоскости или гауссовых числах), мы не можем суммировать по произвольному интервалу $[a; b)$, т. к. его определение будет некорректным. Поэтому договоримся считать, что суммирование с шагом δ всегда означает, что $b - a$ кратно δ , т. е. $b = a + k\delta$ при некотором $k \in \mathbb{N}$. По сути, это означает суммирование вдоль дискретной кривой (по аналогии с интегрированием вдоль пути).

Как видим, мы получаем полный аналог основной формулы интегрального исчисления (формулы Ньютона–Лейбница)²²

$$\int_a^b g(x) dx = f(a) - f(b), \quad \text{где } f'(x) = g(x).$$

На самом деле, никакого секрета тут нет: формула Ньютона–Лейбница получается из (5.6) предельным переходом $\delta \rightarrow 0$ (или подстановкой вместо него бесконечно малого dx), если, конечно, g интегрируема на $[a; b)$.

Аналогично интегральному исчислению для конечных разностей можно определить понятие **неопределенной суммы**:

$$\sum g(x) \delta$$

²²Заметим, кстати, что и символ суммирования, и символ интегрирования являются термами-кванторами, т. к. они связывают переменную суммирования или интегрирования. Более того, они в некотором роде похожи на дизъюнкцию и, соответственно, на квантор существования.

есть класс всех функций $f(x)$ таких, что $\Delta_\delta f = g$. Нетрудно проверить, что при достаточно общих ограничениях имеем

$$\sum g(x)\delta = f(x) + C,$$

Упражнение
5.26.
Попробуйте
описать эти
ограничения.

т. е. все эти функции отличаются лишь на величину, не зависящую от x . Зная «таблицу производных» для разностных степеней, можно построить обратную таблицу неопределенных сумм:

$$\sum x^{\bar{n}}\delta = \frac{x^{\bar{n}+1}}{n+1} + C, \quad \sum x^n\delta = \frac{x^{n+1}}{n+1} + C,$$

где исключение, как обычно, составляет случай $n = -1$. В обычном интегральном исчислении мы знаем формулу $\int (1/x)dx = \ln x + C$. Для разностных степеней аналогом логарифма будет функция

$$\mathcal{H}_\delta(x) \rightleftharpoons \frac{\delta}{x} + \frac{\delta}{x-\delta} + \frac{\delta}{x-2\delta} + \cdots + \frac{\delta}{x-n_x\delta} = \frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{m-n_x},$$

где $x = m\delta$, $m \in \mathbb{N}^+$ (т. е. x кратно δ), а параметр n_x определяется как²³

$$n_x = \frac{x}{\delta} - [1/\delta], \quad \text{где } [y] \rightleftharpoons \min\{k \in \mathbb{N} \mid k \geq y\}.$$

Ясно, что при $\delta \geq 1$ имеем $n_x = (x/\delta) - 1 = m - 1$, а при $\delta = 1$ и $x = m$ мы получаем знаменитое *гармоническое число*

$$\mathcal{H}_m \rightleftharpoons \frac{1}{m} + \frac{1}{m-1} + \cdots + \frac{1}{2} + \frac{1}{1}, \quad \mathcal{H}_0 = 0.$$

Нетрудно видеть, что

Упражнение
5.27.

$$\Delta_\delta \mathcal{H}_\delta(x) = x^{-1}\delta, \quad \sum x^{-1}\delta = \mathcal{H}_\delta(x) + C,$$

поскольку $n_{x+\delta} = n_x + 1$ (или $\Delta_\delta [n_x] = 1$).

Таким образом, мы получаем таблицу неопределенных сумм разностных степеней для всех целых степеней.

Кроме того, имеем следующее равенство:

$$\mathcal{H}_\delta(x) = \mathcal{H}_m - \mathcal{H}_{m-n_x-1} = \mathcal{H}_m - \mathcal{H}_{[1/\delta]-1}.$$

²³Для определения функции «потолок» $[y]$ нам нужно, чтобы в A было задано упорядочение, что накладывает существенные ограничения на A . Однако, если δ нацело делит 1, такое ограничение уже не нужно, и мы просто полагаем $n_x = (x-1)/\delta$. Тем не менее, для простоты можно считать, что $A \subseteq \mathbb{R}$.

А так как мы знаем асимптотику при $m \rightarrow \infty$

$$\mathcal{H}_m = \ln(m) + \gamma + \varepsilon,$$

где $\varepsilon \rightarrow 0$, мы можем указать асимптотику для $\mathcal{H}_\delta(x)$ при фиксированном δ и $x \rightarrow \infty$:

$$\mathcal{H}_\delta(x) = \ln(m) + \gamma - \mathcal{H}_{\lceil 1/\delta \rceil - 1} + \varepsilon.$$

Вместе с тем, при фиксированном x и $\delta \rightarrow 0$ мы имеем $x^{-1} \rightarrow 1/x$, а определенная сумма

$$\sum_a^b x^{-1} \delta = \mathcal{H}_\delta(b) - \mathcal{H}_\delta(a) = \sum_{k=0}^{n-1} \frac{\delta}{a + k\delta} \rightarrow \int_{a^*}^{b^*} \frac{dx}{x} = \ln(b^*) - \ln(a^*),$$

где $a = m\delta \rightarrow a^*$ и $b = (n + m)\delta \rightarrow b^*$. Иначе говоря, мы выбираем a и b кратными δ , но так, чтобы они сходились к некоторым заданным константам a^* и b^* при $\delta \rightarrow 0$. И в этом случае наша определенная сумма будет интегральной суммой для функции $1/x$ на интервале (a^*, b^*) , а значит, будет сходиться к определенному интегралу, значение которого равно $\ln(b^*) - \ln(a^*)$. В частности, полагая $a^* = 1$, мы получаем асимптотику

$$\mathcal{H}_\delta(x) \rightarrow \ln(x), \quad \text{при } \delta \rightarrow 0.$$

Коль скоро речь зашла о логарифме, нужно вспомнить, что функция e^x является неподвижной точкой оператора дифференцирования в \mathbb{R} . Для разностного оператора Δ_δ неподвижной точкой будет функция $(1 + \delta)^{x/\delta}$, которая при замене $\delta = dx$ превращается в e^x , а при $\delta = 1$ — в 2^x , поэтому 2^n считается дискретным аналогом экспоненты.

Вообще, наибольшую ценность в дискретных формулах имеет значение $\delta = 1$, т. к. обеспечивает целый шаг. Например, при таком значении δ имеем

$$\Delta[c^x] = c^{x+1} - c^x = (c - 1)c^x,$$

откуда следует, что

$$\sum c^x = \frac{c^x}{c - 1}, \quad \sum_a^b c^x = \frac{c^b - c^a}{c - 1},$$

и мы получаем формулу суммы геометрической прогрессии.

Наконец, пользуясь формулой (5.4), мы можем выписать **ряд Ньютона**:

$$f(x + n\delta) = f(x) \binom{n}{0} + \Delta_\delta f(x) \binom{n}{1} + \Delta_\delta^2 f(x) \binom{n}{2} + \dots = \sum_{k=0}^{\infty} \frac{\Delta_\delta^k f(x)}{k!} n^k$$

где в степени n^k смещение $\delta = 1$ (наследие числа сочетаний), а в операторе Δ_δ — произвольное. Как видим, это — полный разностный аналог ряда

Здесь γ — не ординал, а константа Эйлера $\gamma = 0.5772\dots$

Тейлора в Анализе, где смещение задается дискретным образом с шагом δ . При этом бесконечная сумма здесь на самом деле является конечной за счет обращения в ноль степеней $n^{\underline{k}}$ при $k > n$. Нетрудно видеть, что если мы полагаем $n\delta = y$ — фиксированное, а $\delta \rightarrow 0$, то $\Delta_\delta^k f(x) n^{\underline{k}}$ превращается в $f^{(k)}(x) \delta^k n^{\underline{k}} = f^{(k)}(x) y^{\underline{k}}$ (если, конечно, f бесконечно дифференцируема). И поскольку экспоненциальный ряд позволяет переходить к пределу под знаком суммы, мы в итоге получим **формулу Тейлора**:

$$f(x + y) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x)}{k!} y^{\underline{k}}.$$

Еще несколько примеров (здесь $\delta = 1$).

$$(x + y)^{\underline{n}} = \sum_{k=0}^n \binom{n}{k} x^{\underline{k}} y^{\underline{n-k}}, \quad (x + y)^{\overline{n}} = \sum_{k=0}^n \binom{n}{k} x^{\overline{k}} y^{\overline{n-k}}$$

Эти формулы, строго говоря, не являются следствием формулы (5.3), т. к. разностная степень не сводится к композиции операторов сдвига, но совершенно очевидно, что у всех биномиальных формул в основе лежат одни и те же комбинаторные принципы, связывающие сумму и произведение.

$$\sum_{k=1}^n k = \sum_{k=1}^n k^{\underline{1}} = \frac{k^{\underline{2}}}{2} \Big|_1^{n+1} = \frac{(n+1)n}{2}.$$

Далее, поскольку $k^2 = k^{\underline{2}} + k^{\underline{1}}$, имеем

$$\sum_{k=1}^n k^2 = \frac{k^{\underline{3}}}{3} \Big|_1^{n+1} + \frac{k^{\underline{2}}}{2} \Big|_1^{n+1} = \frac{(n+1)(n+1/2)n}{3}.$$

Аналогично можно выразить $k^3 = k^{\underline{3}} + 3k^{\underline{2}} + k^{\underline{1}}$ и найти сумму кубов, что мы и предлагаем сделать читателю самостоятельно. | Упражнение 5.28. Вообще, любую степень k^n можно выразить как линейную комбинацию разностных степеней, а коэффициенты при разностных степенях определяются рекурсивно и называются **числами Стирлинга**.

Рассмотрим конечное множество мощности n . Сколько существует способов разбить его на k непустых непересекающихся подмножеств? Например, множество $\{1, 2, 3, 4\}$ можно разбить семью способами:

$$\begin{aligned} &\{1\} \cup \{2, 3, 4\}, \quad \{2\} \cup \{1, 3, 4\}, \quad \{3\} \cup \{1, 2, 4\}, \quad \{4\} \cup \{1, 2, 3\}, \\ &\{1, 2\} \cup \{3, 4\}, \quad \{1, 3\} \cup \{2, 4\}, \quad \{1, 4\} \cup \{2, 3\} \end{aligned}$$

Число способов разбить множество мощности n на k непустых непересекающихся подмножеств называется **числом Стирлинга для множеств** и обозначается

$$\left\{ \begin{matrix} n \\ k \end{matrix} \right\}$$

Из примера выше мы видим, что $\left\{ \begin{matrix} 4 \\ 2 \end{matrix} \right\} = 7$. Для трехточечного множества все еще проще:

$$\left\{ \begin{matrix} 3 \\ 1 \end{matrix} \right\} = 1, \quad \left\{ \begin{matrix} 3 \\ 2 \end{matrix} \right\} = 3, \quad \left\{ \begin{matrix} 3 \\ 3 \end{matrix} \right\} = 1$$

Можно заметить, что рекурсивное тождество, связывающее эти числа, таково:

$$\left\{ \begin{matrix} n \\ k \end{matrix} \right\} = k \left\{ \begin{matrix} n-1 \\ k \end{matrix} \right\} + \left\{ \begin{matrix} n-1 \\ k-1 \end{matrix} \right\}, \quad n > 0 \quad (5.7)$$

Определяя стартовые значения так, чтобы это тождество сохранялось,²⁴

Опять
артемин?

$$\left\{ \begin{matrix} n \\ 0 \end{matrix} \right\} = [n = 0],$$

мы можем вычислять рекурсивно любое число Стирлинга для множеств.

Еще один вид чисел Стирлинга — числа, отвечающие за разложение перестановки в циклы. **Число Стирлинга для циклов** равно количеству перестановок n элементов, содержащих ровно k циклов (при этом тривиальные циклы также считаются), и обозначается

$$\left[\begin{matrix} n \\ k \end{matrix} \right]$$

Например, в группе S_4 имеется 11 двухцикловых перестановок:

$$(123)(4), \quad (132)(4), \quad (124)(3), \quad (142)(3), \quad (134)(2), \quad (143)(2), \\ (234)(1), \quad (243)(1), \quad (12)(34), \quad (13)(24), \quad (14)(23),$$

т. е. $\left[\begin{matrix} 4 \\ 2 \end{matrix} \right] = 11$. Очевидно, что всегда выполняется неравенство

$$\left[\begin{matrix} n \\ k \end{matrix} \right] \geq \left\{ \begin{matrix} n \\ k \end{matrix} \right\}$$

Для чисел Стирлинга для циклов можно вывести рекурсивное тождество

$$\left[\begin{matrix} n \\ k \end{matrix} \right] = (n-1) \left[\begin{matrix} n-1 \\ k \end{matrix} \right] + \left[\begin{matrix} n-1 \\ k-1 \end{matrix} \right], \quad n > 0, \quad (5.8)$$

²⁴Напоминаем, что неформальный символ $[n = 0]$, именуемый нотацией Айверсона, принимает значение 1, если высказывание в скобках истинно, и 0, если ложно.

при этом стартовые значения такие же, как у предыдущих чисел Стирлинга:

$$\begin{bmatrix} n \\ 0 \end{bmatrix} = [n = 0].$$

С помощью чисел Стирлинга можно выразить обычные степени через разностные и наоборот:

$$k^n = \sum_{j=0}^n \left\{ \begin{matrix} n \\ j \end{matrix} \right\} k^{\underline{j}} = \sum_{j=0}^n \left\{ \begin{matrix} n \\ j \end{matrix} \right\} (-1)^{n-j} k^{\overline{j}}$$

$$k^{\overline{n}} = \sum_{j=0}^n \begin{bmatrix} n \\ j \end{bmatrix} k^{\underline{j}}, \quad k^{\underline{n}} = \sum_{j=0}^n \begin{bmatrix} n \\ j \end{bmatrix} (-1)^{n-j} k^j$$

Из последнего, в частности, следует (при $k = 1$), что

$$\sum_{j=0}^n \begin{bmatrix} n \\ j \end{bmatrix} = n!,$$

и это согласуется с тем, что группа S_n имеет порядок $n!$.

Приведенные выше равенства показывают некую двойственность чисел Стирлинга для множеств и для циклов.

Подобно тому, как мы продлили определение разностных степеней в отрицательную область с помощью рекурсивного тождества (5.5), мы можем сделать аналогичную процедуру с числами Стирлинга с помощью рекурсий (5.7) и (5.8). Это позволяет записывать суммы с ними без указания пределов суммирования, например,

$$\sum_k \begin{bmatrix} n \\ k \end{bmatrix} \left\{ \begin{matrix} k \\ m \end{matrix} \right\} (-1)^{n-k} = [n = m] = \sum_k \left\{ \begin{matrix} n \\ k \end{matrix} \right\} \begin{bmatrix} k \\ m \end{bmatrix} (-1)^{n-k}.$$

It's a kind of Magic!

Эти формулы обращения вновь подчеркивают двойственность и симметрию между двумя видами чисел Стирлинга.

Выше мы уже упоминали биномиальные коэффициенты $\binom{n}{k}$, которые не нуждаются в представлении. Отметим только, что и для них продолжение в отрицательную область осуществляется при помощи рекурсивного тождества

Сравните с (5.7) и (5.8)!!

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

Числа Стирлинга, как и биномиальные коэффициенты, возникают во многих комбинаторных задачах. Вообще, комбинаторное исчисление (в сравнении с интегральным и дифференциальным) выглядит более богатым различными тождествами с суммами, двойными и тройными суммами, многочисленными специальными коэффициентами (числовыми функциями от 2-3 и многих параметров) перед степенями переменной и т.п. Дело тут, по-видимому,

в том, что дискретные объекты обладают и дискретными свойствами, т.е. свойствами, не укладывающимися в единую символьную модель. Та же теория делимости, например, не имеет никакого смысла в полях, но очень богата свойствами в дискретных кольцах (вспомним числа Гаусса и Эйзенштейна). С одной стороны, это усложняет исследования (вплоть до неразрешимых задач), с другой — делает картину целых чисел и их аналогов весьма разнообразной и интересной.

Так же, как в интегральном исчислении, для разностей существуют правила дифференцирования, вот одно из них:

$$\Delta_\delta[f g] = f \Delta_\delta[g] + E^\delta g \Delta_\delta[f].$$

Упражнение 5.29. Несимметричность выражения справа нивелируется тем, что «лишний» оператор E^δ может стоять как перед g , так и перед f , формула при этом не изменится. Отсюда мы получаем правило суммирования «по частям»:

$$\sum f \Delta_\delta[g] = f g - \sum E^\delta g \Delta_\delta[f].$$

Например, вспоминая, что при $\delta = 1$ имеем $\Delta[2^k] = 2^k$, получаем

$$\begin{aligned} \sum_{k=0}^n k 2^k &= \sum_{k=0}^n k \Delta[2^k] = k 2^k \Big|_0^{n+1} - \sum_{k=0}^n 2^{k+1} = \\ &= (n+1)2^{n+1} - 2(2^{n+1} - 1) = (n-1)2^{n+1} + 2. \end{aligned}$$

Таким образом, мы можем собрать в таблицу 5.2 сравнение разностных и непрерывных формул.

За многочисленными практическими примерами мы отсылаем читателя к книге [2].

5.3.4 Вариации и дифференциалы

В приложениях важным является такой случай, когда конечную разность можно представить в виде

$$\Delta_\delta[f(x)] = L(x, \delta) + \varepsilon(x, \delta), \quad (5.9)$$

где функционал L линеен по второму аргументу, а *невязка* $\varepsilon(x, \delta)$ оценивается по норме (пространства K) величиной порядка $o(\|\delta\|)$ при $\|\delta\| \rightarrow 0$, вообще говоря, неравномерно по x .

Если такое представление возможно, то его линейная часть $L(x, \delta)$ называется **вариацией** функции f и обозначается δf . В частном случае, когда

Таблица 5.2: Дискретность и непрерывность

конечные разности	интеграл/производная
$\Delta_\delta[(1+\delta)^{x/\delta}] = (1+\delta)^{x/\delta}\delta$	$de^x = e^x dx$
$\sum_k \frac{n^k}{k!} = 2^n \ (\delta = 1)$	$\sum_k \frac{x^k}{k!} = e^x$
$\Delta_\delta x^{\overline{n}} = nx^{\overline{n-1}}\delta, \Delta_\delta x^n = nx^{n-1}\delta$	$dx^n = nx^{n-1}dx$
$\sum x^{\overline{n}}\delta = \frac{x^{\overline{n+1}}}{n+1} + C, \sum x^n\delta = \frac{x^{n+1}}{n+1} + C$	$\int x^n dx = \frac{x^{n+1}}{n+1} + C$
$\sum x^{-1}\delta = \mathcal{H}_\delta(x) + C$	$\int \frac{dx}{x} = \ln(x) + C$
$\Delta_\delta[f g] = f \Delta_\delta[g] + E^\delta g \Delta_\delta[f]$	$d(uv) = u dv + v du$
$\sum_a^b \Delta_\delta f(x)\delta = f(b) - f(a)$	$\int_a^b f'(x)dx = f(b) - f(a)$

$f : \mathbb{L}^n \rightarrow \mathbb{L}$, вариация δf называется **дифференциалом** f и обозначается df .

Пример 1: числовая функция $f(x) = x^2 - x + 1$. Тогда

$$\Delta_\delta[f(x)] = (x+\delta)^2 - (x+\delta) + 1 - (x^2 - x + 1) = (2x-1)\delta + \delta^2,$$

и здесь мы видим, что $L = (2x-1)\delta$, $\varepsilon(x, \delta) = \delta^2$, так что условие линеаризации (5.9) выполняется. Нетрудно видеть, что если обозначить $\delta = dx$, то мы получим $L(x, \delta) = f'(x)dx = df$.

Заметим, что если линейный оператор Δ_δ имеет параметром δ и действует из \mathcal{F} в \mathcal{F} , то оператор L , являясь линейным по приращению δ , имеет параметром функцию f и действует из пространства приращений, т.е. из A , в пространство K .

Пример 2: числовая функция $f(x, y) = e^{xy}$. Приращение вектора (x, y) обозначим $\delta = (\delta_x, \delta_y)$. Тогда

$$\Delta_\delta[f] = e^{(x+\delta_x)(y+\delta_y)} - e^{xy} = e^{xy}(e^{x\delta_y+y\delta_x+\delta_x\delta_y} - 1),$$

где с помощью разложения экспоненты в ряд Тейлора получаем линейную часть $L((x, y), \delta) = e^{xy}(x\delta_y + y\delta_x)$, которая представляет собой дифференциал $df = f'_x dx + f'_y dy$.

И вообще, в случае функции многих переменных дифференциал представляет собой скалярное произведение $\nabla f \cdot d\vec{x}$.²⁵ Если f такова, что существует ее

²⁵ $\nabla f = (f'_{x_1}, \dots, f'_{x_n})$ — вектор частных производных.

дифференциал (т. е. она имеет линейную часть по приращениям по формуле (5.9)), то f называется **дифференцируемой**.

Рассмотрим разность второго порядка для функции из предыдущего примера:

$$\Delta_\delta^2[f(x, y)] = \Delta_\delta[e^{(x+\delta_x)(y+\delta_y)} - e^{xy}] = e^{xy}(e^{2x\delta_y+2y\delta_x+4\delta_x\delta_y} - 2e^{x\delta_y+y\delta_x+\delta_x\delta_y} + 1),$$

где главная часть (по порядку малости при $\delta \rightarrow 0$) будет равна выражению

$$e^{xy}(2\delta_x\delta_y + y^2\delta_x^2 + x^2\delta_y^2 + 2xy\delta_x\delta_y) = f''_{xx}\delta_x^2 + 2f''_{xy}\delta_x\delta_y + f''_{yy}\delta_y^2$$

Здесь присутствуют только вторые степени приращений (первые степени взаимно сократились, а все высшие степени ушли в $\varepsilon(x, \delta)$). Такие функции называются **квадратичными формами** и в общем случае записываются суммой

$$Q(A, \vec{x}) = \sum_{i,j} a_{ij}x_ix_j$$

(в комплексном случае, как обычно, используется сопряжение: $a_{ij}x_i\overline{x_j}$), где матрица $A = (a_{ij})$. Квадратичная форма, в свою очередь, получается из **билинейной формы**:

$$BL(A, \vec{x}, \vec{y}) = \sum_{i,j} a_{ij}x_iy_j$$

путем приравнивания $\vec{x} = \vec{y}$ (в комплексном случае пишем $a_{ij}x_i\overline{y_j}$).

Билинейная форма от двух векторов линейна по каждому аргументу и может быть записана в матричном виде как

$$BL(A, \vec{x}, \vec{y}) = \vec{x}A(\vec{y})^T.$$

В нашем примере:

$$\Delta_\delta^2[f(x, y)] = (\delta_x, \delta_y) \begin{pmatrix} y^2e^{xy} & (1+xy)e^{xy} \\ (1+xy)e^{xy} & x^2e^{xy} \end{pmatrix} \begin{pmatrix} \delta_x \\ \delta_y \end{pmatrix}$$

Как видим, матрица A зависит от переменных x и y и не зависит от вектора приращений $\delta = (\delta_x, \delta_y)$.

В общем случае, если разность второго порядка для функции нескольких переменных можно представить в виде

$$\Delta_\delta^2[f(x)] = \delta A(x)\delta^T + \varepsilon(x, \delta),$$

где $\|\varepsilon\| = o(\|\delta\|^2)$ и матрица A не зависит от δ , то квадратичная форма $\delta A(x)\delta^T$ является **второй вариацией** (вариацией второго порядка) функции f . Вторая вариация обозначается $\delta^2 f$.

Как и в случае первой вариации, для числовых функций нескольких переменных наличие второй вариации означает дифференцируемость второго порядка, а сама вторая вариация совпадает с d^2f , причем матрица соответствующей билинейной формы будет не чем иным, как матрицей вторых частных производных ∇^2 :

$$d^2f = dx\nabla^2(dx)^T.$$

Конечно, и первая, и вторая вариации представляются в таком простом виде, только если мы имеем дело с числовыми функциями (от одной или нескольких переменных). В более сложных случаях все обстоит несколько иначе.

Пример 3: Рассмотрим функционал

$$f(x) = \int_0^1 (x(t) - c)^2 dt.$$

Здесь мы предполагаем, что A состоит из функций, непрерывных на отрезке $[0; 1]$, т. е. $A = C[0; 1]$. Функция f вычисляет квадрат среднего отклонения $x(t)$ от заданной константы c на этом отрезке. Приращением в этом случае будет служить любая функция $\delta(t)$ из этого же семейства. Приращение стремится к нулю, если $\sup |\delta(t)| \rightarrow 0$ (это \sup -норма на $C[a; b]$). Тогда

$$\Delta_\delta[f] = \int_0^1 (x(t) + \delta(t) - c)^2 dt - \int_0^1 (x(t) - c)^2 dt = \int_0^1 (\delta(t)^2 + 2(x(t) - c)\delta(t)) dt.$$

В данном случае вариацией будет интеграл

$$L(x, \delta) = 2 \int_0^1 (x(t) - c)\delta(t) dt,$$

линейный по δ . Невязка будет оцениваться величиной порядка $\|\delta\|^2$.

В нашем простом примере легко увидеть связь между обычным числовым дифференциалом и вариацией для функционалов: достаточно «продифференцировать» f по x под знаком интеграла, и мы получим подынтегральную функцию $2(x - c)$, а затем воспользоваться тем же правилом, что для функции многих переменных — построить «скалярное произведение», считая, что переменная t пробегает эти многие переменные, а интеграл отвечает за суммирование:

$$2(x - c) \cdot \delta = \int_0^1 2(x(t) - c)\delta(t) dt,$$

и вот вам первая вариация. Такой аналог скалярного произведения в случае континуального набора переменных обычно называется **сверткой** функций. Ясно, что интеграл при этом может быть по любой мере и с любым ядром, лишь бы он был сходящимся.

Аналогично числовым функциям для функционалов вычисляется и вторая вариация. В нашем примере она имеет тривиальный вид

$$\delta^2 f = 2 \int_0^1 \delta(t)^2 dt,$$

однако в общем случае это может быть громоздкое выражение, но суть та же — найти второй порядок приближения для исходного функционала при малых отклонениях аргумента.

5.3.5 Вариационное исчисление

Итак, выше мы определили понятие вариации функционала как главную линейную по приращению часть асимптотического приближения этого функционала при малом приращении аргумента. Кроме того, мы увидели, что в ряде случаев имеется возможность дополнить это приближение с помощью квадратичной формы или специального интеграла, так что мы получаем разложение вида

$$f(x + \delta) = f(x) + \delta f + \delta^2 f + o(\|\delta\|^2), \quad (5.10)$$

где δf — первая вариация, $\delta^2 f$ — вторая вариация, которая представляется квадратичной формой или ее интегральным обобщением.

В таблице 5.3 мы приводим сравнение случаев одномерной функции f , функции нескольких переменных и функционала от числовой функции. При этом функционал у нас устроен сам простым образом — он не зависит от производных этой числовой функции. Для более сложных функционалов, соответственно, и формулы будут сложнее.

Мы не включаем в эту таблицу случай дискретной $f(x)$, т. е. числовой последовательности. Это связано с тем, что при дискретном носителе (например, $\text{dom}(f) = \mathbb{N}$) нет возможности говорить о величинах малого порядка вроде $o(\|\delta\|^2)$, и поэтому дифференциальное исчисление, на котором зиждется вариационное исчисление, здесь не сработает.

Тем не менее, следует иметь ввиду, что в дискретном случае мы можем рассматривать $\delta = \pm 1$ и для каждой точки n вычислять соседние значения: $f(n-1)$, $f(n)$, $f(n+1)$. Критерием того, что мы имеем локальный строгий

*Упражнение 5.30.
Подумайте над таким обобщением.*

экстремум в точке n , является условие $\Delta_1[f](n)\Delta_{-1}[f](n) > 0$. Его можно обобщить на случай многомерной дискретной решетки, например, для чисел Гаусса.

Поэтому далее мы предполагаем, что работаем только с «хорошими» и «очень хорошими» функциями и функционалами, позволяющими пользоваться предельными переходами и брать различные производные. Кроме того, для простоты картины мы считаем, что $\mathbb{L} = \mathbb{R}$, чтобы не отвлекаться каждый раз на нюансы действий с комплексными числами (вроде умножения на сопряженное при вычислении скалярного произведения).

Таблица 5.3: Варианты вариаций.

аргумент	первая вариация	вторая вариация
$x \in \mathbb{R}$	$\delta f = f'(x)dx$	$\delta^2 f = f''(x)d^2x$
$\vec{x} \in \mathbb{R}^n$	$\delta f = (\nabla f) \cdot d\vec{x}$	$\delta^2 f = d\vec{x}(\nabla^2 f)(d\vec{x})^T$
$x \in C[0; 1],$	$\delta f =$	$\delta^2 f =$
$f(x) = \int_0^1 F(x(t), t)d\mu$	$\int_0^1 F'_x(x(t), t)\delta(t)d\mu$	$\int_0^1 F''_x(x(t), t)\delta^2(t)d\mu$

Основной задачей вариационного исчисления является поиск экстремальных значений аргумента функционала $f(x)$, которые в общем случае называются *экстремумами*, а также поиск значений x , доставляющих нулевое значение первой вариации δf , которые называются *экстремальными*.

Дадим определение. Точка x_0 называется (строгим) **локальным минимумом** функции f , если в некоторой окрестности $O \subseteq \text{dom}(f)$ ²⁶ этой точки выполняется условие:

$$\forall x \in O \setminus \{x_0\} : f(x) \geq f(x_0) \quad (f(x) > f(x_0)). \quad (5.11)$$

Под окрестностью O можно понимать шар $B(x_0)$ некоторого небольшого радиуса в смысле метрики, порожденной нормой на пространстве аргументов x . Неравенство $f(x) \geq f(x_0)$ можно также переписать в виде $\Delta_\delta[f(x_0)] \geq 0$, где вариация δ аргумента x уже должна будет находиться в некоторой окрестности нуля, т. е. $\|\delta\| < \varepsilon$.

Аналогично определяется (строгий) **локальный максимум**. Вместе минимум и максимум именуются **экстремумами** (экстремальями). Если же указанные неравенства выполняются всюду в области определения f , то экстремум называется **глобальным**.

Как и в случае с обычной числовой функцией, возможность представления функционала в виде (5.10) позволяет, во-первых, найти необходимое условие экстремума (а именно — равенство нулю первой вариации), а во-вторых, сформулировать достаточное условие через понятие положительной или отрицательной определенности второй вариации.

Если первая вариация δf существует (т. е. главная часть f линейна по приращению) в точке x и при этом *непрерывна* по приращению, то она называется также **дифференциалом Фрешё**, а сам функционал f в этом случае называется **сильно дифференцируемым** в точке x .

²⁶Это вложение означает, что x_0 — внутренняя точка области определения f .

Рассмотрим числовую функцию

$$J_{x,\delta}(\alpha) = f(x + \alpha\delta)$$

с параметрами x и δ , где α — действительное число (поскольку x, δ принадлежат линейному пространству над \mathbb{L} , такая комбинация допустима). При фиксированных параметрах $J_{x,\delta}$ является обычной числовой функцией, так что мы можем использовать стандартные средства анализа для ее изучения.

Если существует производная $\partial J_{x,\delta}/\partial\alpha$ в точке $\alpha = 0$, то она называется **дифференциалом Гато** для функционала f в точке x в направлении δ и обозначается $d_G f(x, \delta)$.²⁷

Дифференциал Гато является однородным по δ : $d_G f(x, \lambda\delta) = \lambda d_G f(x, \delta)$. Однако, в общем случае он не является аддитивным и, стало быть, линейным.

Если функционал f имеет дифференциал Гато в точке x , то он называется **слабо дифференцируемым** в точке x .

Нужно отметить, что термин «дифференциал» относится не к самому линейному оператору, а к его значению на приращении δ , в то время как сам по себе оператор принято называть «производной». Так, если у нас первая вариация представляется в виде $\delta f = L(x, \delta)$, где $L(x, \delta)$ — линейный непрерывный по δ оператор, то дифференциал есть значение $L(x, \delta)$ в точке δ , а производная (Фреше в данном случае) есть сам оператор $L(x, \cdot)$ с параметром x .

Аналогично, в n -мерном случае, если $df = \nabla f \cdot d\vec{x}$, то скалярное произведение $\nabla f \cdot d\vec{x}$ есть дифференциал Фреше, а оператор взятия производных (градиент) ∇f — производная Фреше. В одномерном случае производная выражается просто числом $f'(x)$, хотя подразумевается, что это есть одномерный линейный оператор $f'(x) \cdot$, где вместо точки могут подставляться любые числа.

Точно так же, если вторая вариация имеет представление $\delta^2 f = \delta M \delta^T$ с матрицей билинейной формы M , то полином $\delta M \delta^T$ есть второй дифференциал, а сама квадратичная форма M (читай — сама матрица M) есть вторая производная.

Эти терминологические тонкости могут ввести в заблуждение, поэтому мы будем использовать только термин «дифференциал» для единообразия языка.

Лемма 5.4. *Если существует дифференциал Фреше, то существует дифференциал Гато и они равны (в точке x).*

Доказательство. Поскольку существует дифференциал Фреше, имеет место

²⁷Заметим, что в аналитической геометрии используется понятие производной по направлению, которое определяется точно так же.

представление

$$\Delta_\delta[f(x)] = L(x, \delta) + o(\|\delta\|)$$

при $\|\delta\| \rightarrow 0$, где $L(x, \delta)$ — линейна и непрерывна по δ . Откуда получаем, что

$$\begin{aligned} \frac{J_{x,\delta}(\alpha) - J_{x,\delta}(0)}{\alpha} &= \frac{\Delta_{\alpha\delta}[f(x)]}{\alpha} = \frac{L(x, \alpha\delta) + o(\|\alpha\delta\|)}{\alpha} = \\ &= \frac{\alpha L(x, \delta) + o(\alpha\|\delta\|)}{\alpha} \rightarrow L(x, \delta) \end{aligned}$$

при $\alpha \rightarrow 0$ при фиксированных параметрах x, δ .

Таким образом, $J_{x,\delta}(\alpha)$ дифференцируема по α и ее производная (дифференциал Гато) совпадает с первой вариацией f , дифференциалом Фреше. \square

Лемма 5.5 (необходимое условие экстремума). *Если x_0 — экстремум f и f сильно дифференцируема в точке x_0 , то $\delta f = 0$ в точке x_0 (тождественно по приращению δ).*

Доказательство. Поскольку f сильно дифференцируема, она также слабо дифференцируема в точке x_0 . Тогда точка $\alpha = 0$ является экстремумом функции $J_{x_0,\delta}(\alpha)$ при любом допустимом приращении δ . Далее применяем стандартную теорему Анализа о необходимом условии экстремума — равенство производной нулю в точке $\alpha = 0$.

Равенство $J'_{x_0,\delta}(0) = 0$ означает, что дифференциал Гато, а следовательно, и Фреше, равен 0 тождественно по приращению δ . Следовательно, вариация $\delta f = 0$. \square

Очевидно, что верна лемма в более слабой формулировке: если x_0 — экстремум и f слабо дифференцируема в точке x_0 , то производная Гато равна 0 в этой точке.

Точка x , на которой достигается ноль первой вариации, называется **стационарной точкой** (иногда используется термин «подозрительная на экстремум»).

Как видим, лемма опирается на известный факт из одномерного вещественного анализа, который ни в коем случае нельзя выводить из данной леммы во избежание порочного круга. Однако уже для многомерного случая лемма дает необходимое условие экстремума — равенство нулю градиента f в точке x_0 .

Для случая простого функционала (не зависящего от производных x)

$$f(x) = \int_a^b F(x(t), t) d\mu,$$

где $F(x, t)$ — числовая функция от двух аргументов, необходимое условие экстремума выглядит следующим образом:

$$\int_a^b F'_x(x(t), t) \delta(t) d\mu = 0$$

тождественно по всем функциям $\delta(t)$ из области определения функционала f . При этом мы, конечно же, предполагаем, что функция F и мера μ достаточно «хороши», так что при вычислении пределов и производных можно переходить под знак интеграла (например, F непрерывно дифференцируема по x , а μ — мера Лебега, либо иная абсолютно непрерывная мера²⁸).

Чтобы не смущать читателя абстрактной мерой (тем более что в случае ее абсолютной непрерывности мы можем внести ее плотность в функцию F и тем самым прийти к интегралу Лебега), далее будем предполагать, что μ есть мера Лебега.

Более сложный случай представляет собой функционал

$$f(x) = \int_a^b F(x(t), x'(t), t) dt,$$

где x' обозначает производную функции $x(t)$, а $F : \mathbb{R}^3 \rightarrow \mathbb{R}$ дважды непрерывно дифференцируема по всем аргументам.

Стационарную точку данного функционала будем искать при следующих условиях: $x(a) = x_0$, $x(b) = x_1$. Соответственно, все вариации x также должны удовлетворять этим *граничным условиям*, т. е. $\delta(a) = 0 = \delta(b)$. Кроме того, предполагается непрерывная дифференцируемость x и δ на $[a; b]$.

Первая вариация этого функционала ищется довольно просто (опускаем аргументы там, где их легко восстановить из контекста):

$$\Delta_{\alpha\delta}[f] = \int_a^b (F(x + \alpha\delta, x' + \alpha\delta', t) - F(x, x', t)) dt,$$

откуда, пользуясь дифференцируемостью F , получаем разложение

$$\Delta_{\alpha\delta}[f] = \int_a^b (F'_x(x, x', t) \alpha\delta + o(\|\delta\|) + F'_{x'}(x, x', t) \alpha\delta' + o(\|\delta'\|)) dt$$

Здесь

мы сталкиваемся с необходимостью учитывать норму производной $\delta'(t)$, т. е. $\sup |\delta'(t)|$, поэтому в задачах, где функционал зависит от производных

²⁸ Абсолютная непрерывность меры означает наличие у нее функции плотности, или производной, которая позволяет свести интеграл по этой мере к интегралу Лебега. Подробнее об этом см. теорему Радона–Никодима, например, в [81].

своего аргумента, принято использовать следующую норму функции:

$$\|x\|^{(p)} = \sum_{k=0}^p \sup_{a \leq t \leq b} |x^{(k)}(t)|,$$

Упражнение 5.31.
Проверьте, что это норма.

где предполагается существование и непрерывность p -ой производной $x(t)$. Соответственно, и разложение (5.10) учитывает именно эту норму в остаточном члене.

В связи с разнообразием норм различают сильный и слабый экстремумы. Экстремум x_0 называется **сильным минимумом**, если в его определении (5.11) использовать обычную \sup -норму для определения окрестности точки x_0 , т. е. $\sup |x - x_0|$, и экстремум называется **слабым минимумом**, если использовать норму $\|x - x_0\|^{(p)}$ высших порядков ($p \geq 1$). Аналогично определяются сильный и слабый максимум.

Иначе говоря, утверждение, что неравенство $f(x) \geq f(x_0)$ выполняется для всех функций, близких к x_0 в смысле \sup -нормы, очевидно, сильнее утверждения, что данное неравенство выполняется для функций, близких не только в смысле \sup -нормы, но и в смысле близости производных. Поэтому если x_0 — сильный экстремум, то он же является и слабым экстремумом.

Осторожно! Не следует считать, что слабый минимум ищется с помощью слабого дифференциала, а сильный — с помощью сильного! Это не более чем терминологическое совпадение.

Легко видеть, что «обычные» \sup -нормы δ и δ' можно оценить сверху нормой $\|\delta\|^{(1)}$, откуда

$$\Delta_{\alpha\delta}[f] = \int_a^b F'_x(x, x', t) \alpha \delta dt + \int_a^b F'_{x'}(x, x', t) \alpha \delta' dt + o((b-a)\|\delta\|^{(1)})$$

Второй интеграл берем по частям:

$$\int_a^b F'_{x'}(x, x', t) \alpha \delta' dt = \alpha F'_{x'}(x, x', t) \delta \Big|_a^b - \alpha \int_a^b (F''_{xx'} x' + F''_{x'x'} x'' + F''_{x't}) \delta dt.$$

Таким образом, линейная по $\alpha\delta$ и главная по асимптотике часть $\Delta_{\alpha\delta}[f]$, т. е. вариация f , имеет вид

$$\delta f = \int_a^b (F'_x - F''_{xx'} x' - F''_{x'x'} x'' - F''_{x't}) \alpha \delta dt,$$

соответственно, дифференциал Гато равен

$$d_G f = \int_a^b (F'_x - F''_{xx'} x' - F''_{x'x'} x'' - F''_{x't}) \delta dt.$$

Его равенство нулю тождественно по δ равносильно дифференциальному **уравнению Эйлера**

$$F''_{x'x'}x'' + F''_{x'x}x' + F''_{x't} - F'_x = 0 \quad (5.12)$$

относительно x на отрезке $[a; b]$. Данный эквивалентный переход от равенства нулю интеграла к равенству нулю подынтегральной функции диктуется следующей леммой.

Лемма 5.6 (основная лемма вариационного исчисления). Пусть $\bar{\Omega}$ — замыкание открытого ограниченного множества в \mathbb{R}^n . Если

$$\int_{\Omega} F(\vec{x})\delta(\vec{x})d\vec{x} = 0$$

при любой непрерывной функции $\delta(\vec{x})$, то $F(\vec{x}) = 0$ на $\bar{\Omega}$.

Если заменить непрерывность δ на измеримость δ по Лебегу, то в конце формулировки нужно добавить слова «почти всюду».

Упражнение 5.32. Предлагаем читателю самостоятельно доказать эту лемму, которая лежит в основе всего вариационного исчисления по той причине, что позволяет от интегрального уравнения перейти к дифференциальному, избавившись при этом от линейного параметра — вариации аргумента функционала.

Уравнение (5.12) было сформулировано Эйлером в 1744 году и носит его имя. Если существует экстремум функционала $f(x)$ при заданных ограничениях на x , то он с необходимостью является решением уравнения (5.12).

Для иллюстрации изложенной теории приведем один простой пример. Пусть требуется построить линию наискорейшего спуска (под действием гравитации и без учета силы трения) из точки $(0, 0)$ в точку $(1, h)$, где $h > 0$ (ось ординат смотрит вниз, к центру гравитации). То есть мы ищем экстремали $x(t)$ с граничными условиями $x(0) = 0$ и $x(1) = h$. Переменная t пробегает горизонтальную ось, x отвечает за высоту тела над горизонтом. Линия наискорейшего спуска называется также **брахистрохрой**. Задача о брахистрохре была сформулирована Иоганном Бернулли в 1696 году и решена сразу несколькими учеными (в том числе Ньютоном и Лейбницем) в 1697 году.

Из закона сохранения энергии (равенство кинетической и потенциальной энергий) получаем выражение для абсолютного значения скорости:

$$v = \sqrt{2gx},$$

где g — ускорение свободного падения (считаем его постоянным). Отсюда находим проекцию скорости на горизонтальную ось:

$$v_t = \frac{v}{\sqrt{1 + (x')^2}} = \frac{\sqrt{2gx}}{\sqrt{1 + (x')^2}}.$$

Поскольку время на спуск равняется $\int_0^1 (1/v_t)dt$, задача сводится к минимизации функционала

$$f(x) = \frac{1}{\sqrt{2g}} \int_0^1 \sqrt{\frac{1 + (x'(t))^2}{x(t)}} dt.$$

Решение уравнения Эйлера для этой задачи удобно выражается в параметрическом виде

$$\begin{cases} x &= c(\theta - \sin(\theta)), \\ t &= c(1 - \cos(\theta)), \end{cases}$$

константа c определяется из условий $t(\theta_1) = 1, x(\theta_1) = h$ при некотором θ_1 .

Такая кривая представляет собой циклоиду, ее физическое воплощение можно встретить на крышах традиционных китайских домов, поскольку она обеспечивает наискорейший спуск воды, что весьма актуально в условиях тропических ливней.

На рисунке 5.2 красная линия представляет собой брахистохрону, свободный спуск по ней происходит заметно быстрее, чем по синим линиям.

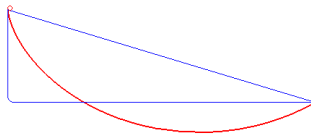


Рис. 5.2: Брахистохрона.

Уравнение (5.12) позволяет найти вид экстремали (если таковая существует), но при этом не гарантирует, что найденная стационарная точка действительно будет экстремумом. Достаточно для примера вспомнить функцию x^3 , производная которой обращается в 0 в нуле, однако экстремум в этой точке отсутствует. В многомерном анализе помимо экстремумов существуют седловые точки, которые в одном направлении являются точками условного максимума, а в другом — условного минимума.

Для однозначного установления того, что найденная из уравнения $df = 0$ стационарная точка x_0 является экстремумом, нужно понимать, является ли функционал f выпуклым в данной точке. Для этого существует критерий, связанный с вычислением второй вариации функционала. При этом, если существует второй дифференциал Гато, то он совпадает со второй вариацией, поэтому находим вторую производную функции $J_{x,\delta}(\alpha)$ при $\alpha = 0$ по направлению δ в точке x_0 :

$$d_G^2 f(x_0, \delta) = \frac{\partial^2 f(x_0 + \alpha\delta)}{\partial^2 \alpha}.$$

Далее, про $\delta^2 f = d_G^2 f(x_0, \delta)$ можно сказать следующее:

- положительно определена, если для всех $\delta \neq 0$: $d_G^2 f(x_0, \delta) > 0$;
- отрицательно определена, если для всех $\delta \neq 0$: $d_G^2 f(x_0, \delta) < 0$;
- не определена, если $d_G^2 f(x, \delta)$ меняет знак или равна нулю в зависимости от δ .

Условие $\delta \neq 0$ означает, что $\|\delta\| > 0$, а все δ при этом выбираются из некоторой окрестности нуля. При этом, как уже упоминалось ранее, в зависимости от структуры нормы, определяющей эту окрестность, экстремум называется сильным или слабым.

В случае положительной определенности второй вариации найденная из условия $\delta f = 0$ экстремаль x_0 является минимумом, а в случае отрицательной определенности — максимумом. В неопределенном случае могут быть разные варианты.²⁹

Вспомним, что в многомерном анализе есть такое свойство у квадратичных форм — положительная/отрицательная определенность. А поскольку у нас в случае n переменных вторая вариация представляется как квадратичная форма от n переменных, условие положительной/отрицательной определенности формы является достаточным условием минимума/максимума.

*и критерий
Сильвестра!*

В случае функционалов общего вида мы имеем ровно то же самое!

Рассмотрим снова функционал

$$f(x) = \int_a^b F(x, x', t) dt,$$

где поиск экстремума осуществляется при условиях $x(a) = x_0$, $x(b) = x_1$ (задача с «закрепленными концами»), а функция F трижды непрерывно дифференцируема в области определения.

Предположим, что $x_0(t)$ — некоторая экстремаль, т. е. решение уравнения Эйлера (5.12).

Находим второй дифференциал Гато с помощью $J_{x_0, \delta}(\alpha) = f(x_0 + \alpha\delta)$:

$$\frac{d^2}{d^2\alpha} J_{x_0, \delta}(\alpha) = \frac{d}{d\alpha} \int_a^b (F'_x \delta + F'_{x'} \delta') dt = \int_a^b (F''_{xx} \delta^2 + 2F''_{xx'} \delta \delta' + F''_{x'x'} (\delta')^2) dt.$$

Заметим, что $d\delta^2 = 2\delta\delta' dt$, так что второй интеграл можно взять по частям

$$\int_a^b 2F''_{xx'} \delta \delta' dt = F''_{xx'} \delta^2 \Big|_a^b - \int_a^b \delta^2 dF''_{xx'} = - \int_a^b \frac{d}{dt} F''_{xx'} \delta^2 dt,$$

²⁹На самом деле, все не так просто: к достаточным условиям экстремума следует еще отнести **условие Якоби** — см. ниже.

где мы воспользовались условием $\delta(a) = 0 = \delta(b)$. В итоге получаем

$$\frac{d^2}{d^2\alpha} J_{x_0, \delta}(0) = \int_a^b (F''_{xx} - \frac{d}{dt} F''_{xx'}) \delta^2 dt + \int_a^b F''_{x'x'} (\delta')^2 dt,$$

значения всех производных F считаются в точке $(x_0(t), x'_0(t), t)$, т. к. $\alpha = 0$.

Можно отсюда показать, что необходимым условием того, что x_0 представляет минимум (максимум), является выполнение неравенства $F''_{x'x'} \geq 0$ ($F''_{x'x'} \leq 0$) вдоль экстремали x_0 .³⁰

Условие для минимума $F''_{x'x'} \geq 0$ называется **условием Лежандра**, а условие $F''_{x'x'} > 0$ — **усиленным условием Лежандра** (аналогично — для максимума). Будем считать далее, что усиленное условие Лежандра для минимума выполнено для экстремали x_0 . Понятно, что для максимума все делается аналогично.

Рассмотрим теперь дополнительную вариационную задачу. Определим функционал

$$k(v) \Rightarrow \int_a^b (F''_{xx} - \frac{d}{dt} F''_{xx'}) v^2 dt + \int_a^b F''_{x'x'} (v')^2 dt,$$

где вариация $v = \delta$ является аргументом функционала, а в качестве $x(t)$ выбрана экстремаль x_0 .

Для краткости обозначим

$$\begin{aligned} S(t) &\Rightarrow F''_{xx}(x_0(t), x'_0(t), t) - \frac{d}{dt} F''_{xx'}(x_0(t), x'_0(t), t), \\ R(t) &\Rightarrow F''_{x'x'}(x_0(t), x'_0(t), t), \end{aligned}$$

тогда

$$k(v) = \int_a^b (Sv^2 + R(v')^2) dt.$$

Уравнение Эйлера для этого интеграла будет иметь вид

$$2Sv - 2\frac{d}{dt} Rv' = 0,$$

оно называется **уравнением Якоби** (для исходной задачи).

Так как мы предполагаем, что $R = F''_{x'x'} > 0$ (усиленное условие Лежандра для минимума) на экстремали x_0 , мы можем разделить уравнение Якоби на R , и в результате придем к новому уравнению вида

$$v'' + p(t)v' + q(t)v = 0, \quad p(t) = F'''_{x'x't}/R, \quad q(t) = -S/R,$$

³⁰Сравните два уравнения: 1. уравнение Эйлера для экстремали $F'_x - \frac{d}{dt} F'_{x'} = 0$, эквивалентное (5.12), 2. $F''_{xx} - \frac{d}{dt} F''_{xx'} = 0$ из первого интеграла выше.

причем, по условию задачи коэффициенты p и q непрерывны на $[a, b]$.

Пусть $v_0(t)$ — решение уравнения Якоби, удовлетворяющее условиям: $v_0(a) = 0$ и $v'_0(a) = 1$ (последнее условие нужно, чтобы исключить тождественный ноль и рассматривать только близкие к исходной экстремали кривые).

Существенным для дальнейшего будет тот факт, имеет ли решение $v_0(t)$ корни внутри $[a; b]$. Оказывается, что если такие корни имеются, то исследуемая экстремаль x_0 не может давать минимум функционалу $f(x)$.

Если $v_0(t) \neq 0$ при $a < x < b$, то говорят, что экстремаль $x_0(t)$ в интервале $(a; b)$ удовлетворяет **условию Якоби**, а если $v_0(t) \neq 0$ при $a < x \leq b$, то говорят, что экстремаль $x_0(t)$ удовлетворяет **усиленному условию Якоби**.

Имеет место следующая

Теорема 5.6. *Усиленные условия Лежандра и Якоби достаточны для того, чтобы экстремаль доставляла слабый (локальный) экстремум функционалу $f(x)$.*

Можно также показать, что если выполнены усиленные условия Лежандра и Якоби и, кроме того, $F''_{x'x'}(x, p, t)$ положительно (отрицательно) для всякого конечного p в некоторой области, содержащей экстремаль $x_0(t)$ внутри, то эта экстремаль дает сильный (абсолютный) минимум (максимум).

Выше мы рассмотрели случай, когда подынтегральная функция F функционала f зависит только от нулевой и первой производной функции x . А сама функция x закреплена на концах в заданных точках плоскости. Существует несколько обобщений данной задачи:

- F зависит от производных до порядка $p > 0$;
- x является вектор-функцией с m компонентами (x_1, \dots, x_m) ;
- x является функцией от n переменных $x(t_1, \dots, t_n)$;
- задача с «подвижными концами».

В совокупности первых трех обобщений подынтегральная функция имеет вид

$$F(t_1, \dots, t_n, x_1, \dots, x_m, (x_1)'_{t_1}, \dots, (x_m)'_{t_n}, \dots, (x_1)''_{t_1 t_1}, \dots, (x_m)^{(p)}_{t_n \dots t_n})$$

т. е. имеет порядка $n^{p+1}m$ аргументов.

Уравнение Эйлера, общее для первых трех обобщений, имеет вид системы

из m уравнений

$$\begin{aligned} \nabla_{\vec{x}} F - \sum_{j=1}^n \frac{\partial}{\partial t_j} \nabla_{\vec{x}'_{t_j}} F + \sum_{t_{j_1}, t_{j_2}} \frac{\partial^2}{\partial t_{j_1} \partial t_{j_2}} \nabla_{\vec{x}''_{t_{j_1} t_{j_2}}} F + \\ \dots + (-1)^p \sum_{t_{j_1}, \dots, t_{j_p}} \frac{\partial^p}{\partial t_{j_1} \partial t_{j_p}} \nabla_{\vec{x}^{(p)}_{t_{j_1} \dots t_{j_p}}} F = 0, \quad (5.13) \end{aligned}$$

где $\nabla_{\vec{x}^{(p)}_{t_{j_1} \dots t_{j_p}}}$ означает, что для каждой компоненты x_i вектора \vec{x} мы находим частные производные F по тем ее аргументам, которые обозначены как соответствующие частные производные x_i порядка p , т. е. на i -ой строке данного вектора стоит частичный градиент F следующего вида

$$\left(\dots \frac{\partial F}{\partial^p(x_i)/\partial t_{j_1} \dots \partial t_{j_p}} \dots \right),$$

который затем сворачивается с полной производной по тем же переменным t_{j_1}, \dots, t_{j_p} . Мы не можем здесь написать дифференциал, т. к. он предполагает нахождение производных вообще по всему вектору переменных \vec{t} , но в рамках указанного набора переменных это — дифференциал, т. е. для его вычисления указанные выше компоненты частичного градиента нужно дифференцировать по всем аргументам по правилам вычисления производной сложной функции.

Первое слагаемое при этом имеет простой вид:

$$\nabla_{\vec{x}} F = \begin{pmatrix} F'_{x_1} \\ \vdots \\ F'_{x_m} \end{pmatrix}$$

Символ ∇ здесь используется для того, чтобы подчеркнуть, что уравнение (5.13) на самом деле является системой уравнений для каждой компоненты \vec{x} .

Для более полного погружения в вариационное исчисление мы рекомендуем классическую книгу Понтрягина [62].

Принцип максимума Понтрягина

В конечномерном случае мы хорошо знаем, как находить условный экстремум функции при наличии связующих ограничений на ее переменные. Для этого строится функция Лагранжа, которая получается из исходной добавлением уравнений условий с некоторыми неизвестными коэффициентами.

Принцип максимума Понтрягина обобщает этот метод на случай континуума переменных, т. е. для функционалов.

Для простоты мы снова рассмотрим интегральный функционал от числовой функции одной переменной:

$$f(x) = \int_a^b F(x, u, t) dt,$$

где x, u — достаточно гладкие функции от t на отрезке $[a; b]$, связанные дополнительным условием

$$x' = \gamma(x, u, t)$$

(в частности, при $\gamma = u$ мы имеем тот интеграл, который рассматривали выше).

Тогда задача минимизации интеграла $f(x)$ сводится к задаче минимизации интеграла

$$l(x, u) = \int_a^b L(u, x, t) dt,$$

где

$$L(u, x, t) = F(x, u, t) + \lambda(t)(x' - \gamma(x, u, t)),$$

т. е. перед нами функция Лагранжа для континуум-мерного случая (переменная t пробегает континуум и индексирует уравнения ограничений).

Далее задача поиска минимума решается так. Во-первых, составляется уравнение Эйлера для функции L , откуда получается уравнение для функции $\lambda(t)$. Во-вторых, при каждом t выбирается $u(t)$ так, чтобы она доставляла минимум функции $L(u)$ (с учетом найденного $\lambda(t)$).

Найденная $u(t)$ и будет искомой экстремалью с условием $x' = \gamma(x, u, t)$.

Как видим, с помощью функции u мы можем свести классическую задачу вариационного исчисления к задаче на условный экстремум, причем выбором зависимости между u и x (или ее производными) мы управляем так, чтобы было удобнее решать задачу. Собственно, традиционное обозначение для нового параметра u и происходит от русского слова «управление».

Посмотрим, как этот метод работает на классической задаче Ньютона об обтекаемой поверхности.

Задача, сформулированная Ньютоном в книге «Начала натуралистической философии», вышедшей в 1687 году, строго формулируется следующим образом. Дан круг известного радиуса, нужно сделать для него обтекатель в виде тела вращения, симметричного относительно центральной нормали этого круга, так, чтобы сопротивление встречному потоку (вдоль той же нормали) однородной разреженной стационарной среды было минимальным.



*Лев
Семенович
Понтрягин*

Конечно, все сразу подумают о торпедах и подводных лодках. Как бы не так! Для упрощения задачи предполагается именно сильная разреженность среды, чтобы не возникало турбулентных эффектов, а также ее неподвижность, чтобы скорости всех частиц были равны по модулю и направлению (т. е. противоположно направлены движению этого обтекателя). Кроме того, предполагается, что соударение частицы с искомой поверхностью является абсолютно упругим и не приводит к ее повторным соударениям ни с другими частицами, ни с поверхностью.



Исаак
Ньютон

Думается, Ньютон выбрал такие условия не для того, чтобы рассчитать геометрию капсулы спускаемого аппарата или сверхзвуковых стратосферных ракет, он лишь искал простые подступы к задаче об оптимальном обтекателе. Тем не менее, задача именно в таком виде пригодилась как раз в космонавтике и высотных сверхзвуковых полетах.

Поскольку мы ищем тело вращения, то, как в задаче о брахистохроне, нужно соединить точку радиуса данного круга с точкой на оси вращения на удалении h от этого круга, так что тело будет иметь высоту h , а в основании круг радиуса r (на рис. 5.3 искомая кривая отмечена красным).

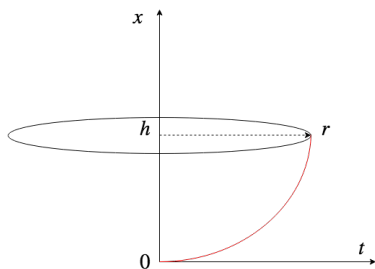


Рис. 5.3: Обтекатель — постановка задачи.

Для данной вариационной задачи можно составить (с помощью закона сохранения импульса) целевой функционал

$$f(x) = \int_0^T \frac{tdt}{1 + (x')^2}$$

с краевыми условиями $x(0) = 0$, $x(r) = h$ и ограничением на производную $x' \geq 0$.

Чтобы найти решение с помощью метода максимума Понтрягина, вводим управление $u = x'$ с ограничением $u \geq 0$ и получаем функцию Лагранжа

$$L(u) = \frac{t}{1 + u^2} + \lambda(t)(x' - u).$$

Уравнение Эйлера для данной функции имеет чрезвычайно простой вид:

$$\lambda' = 0,$$

т. е. $\lambda(t) \equiv \lambda_0$ — константа по t .

Следовательно, в каждой точке t нужно найти точку минимума функции $L(u) = t/(1 + u^2) + \lambda_0(x' - u)$ при указанных ограничениях.

Легко видеть, что при $\lambda_0 \geq 0$ минимум равен $-\infty$, поэтому полагаем $\lambda_0 < 0$, при котором для малых t функция $L(u)$ достигает минимума в нуле. Далее, начиная со значения $t = -2\lambda_0$ минимум уходит от нуля вверх и достигается в точке u , удовлетворяющей дифференциальному уравнению

$$-\lambda_0 - \frac{2tu}{(1 + u^2)^2} = 0,$$

откуда, учитывая равенство $x' = u$, находим, что

$$x = \frac{-\lambda_0}{2} \left(-\ln u + u^3 + \frac{3}{4}u^4 \right) + \frac{7}{8}\lambda_0,$$

$$t = \frac{-\lambda_0}{2} \left(\frac{1}{u} + 2u + u^3 \right),$$

т. е. управляющий параметр становится параметром получаемой кривой. Величина λ_0 находится из начального условия $x(r) = h$.

Таким образом, обтекатель получился тупоносым! См. рис. 5.4

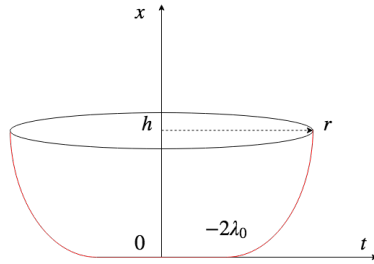


Рис. 5.4: Обтекатель — решение.

Здесь уместно вспомнить о форме капсулы спускаемого аппарата корабля «Союз».

На самом деле, можно немного улучшить сопротивляемость плоского носа обтекателя, если его сделать зубчатым. Если каждый зубец есть треугольник с углом около 30 градусов, то давление на плоскую часть обтекателя существенно уменьшается просто за счет разложения импульса молекул на вертикальную и горизонтальную составляющую (последние при этом симметрично друг друга компенсируют, и столкновения капсулы с траектории не происходит).

Но и это решение, как выясняется, не оптимальное. Здесь нужно немного вспомнить топологию. Ранее мы предполагали, что капсула должна быть гомотопична шару, т. е. представлять собой односвязный объем пространства. Но что если ее сделать гомотопичной тору (с внутренностью)?

Решение было найдено лишь в 2009-м году А. Ю. Плаховым [86]. Оказывается, что если выбрать тор правильной формы, а именно: вращать равнобедренный треугольник вокруг оси, параллельной основанию треугольника, то при правильно подобранных углах треугольника мы получим фигуру (внимание!) с нулевым сопротивлением!

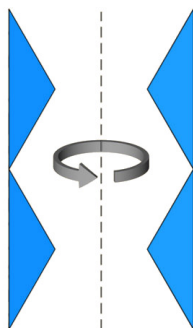


Рис. 5.5: Обтекатель — треугольный тор. Права на изображение принадлежат Etudes.ru

Конечно, физически оно будет не нулевым, а просто очень маленьким. Однако, если вместо потока частиц разреженного газа мы возьмем поток света (например, солнечный ветер), такая фигура станет практически прозрачной вдоль своей оси симметрии. От полностью прозрачной фигуры ее отличает лишь перевернутое изображение в центре. Но и это поправимо: достаточно взять два таких одинаковых тора, поставить друг на друга, и никакого перевернутого изображения не будет — фигура почти полностью исчезнет, если смотреть на нее вдоль оси симметрии (см. рис. 5.5).

Анимацию на эту тему можно посмотреть [здесь](#).

При этом, внутри треугольных торов можно получить почти половину объема охватывающего цилиндра, т. е. 45% полезного объема при почти полной прозрачности!

Оказывается также, что если вместо прямолинейных треугольников выбрать параболические, то коэффициент полезного объема можно увеличить аж до 67% !

Как видим, задачи вариационного исчисления выглядят совсем уж прикладными на фоне той теории, которая излагалась выше, хотя и вырастают непосредственно из нее. Вообще, удивительное свойство математики состоит в том, что она плодит несметное количество логического и вычислительного инструментария, не утруждая себя проверкой его на практике, но тем не менее рано или поздно ему находитесь применение, причем порой весьма неожиданное. Означает ли это, что мир устроен математически или что мы сами

интерпретируем физические явления на некотором языке, который записываем математическими формулами? Ответы на подобные вопросы выходят за рамки нашей книги, но должны волновать любого, кто соприкасается с этой наукой.

Заметим, что строгий континуальный анализ, который, казалось бы, не имеет никакого отношения к реальности (ведь вещество невозможно дробить бесконечно, а число π не реализуется в материи с бесконечной точностью), играет очень мощную роль в обосновании используемых математических методов. Так, если мы решаем обыкновенное дифференциальное уравнение вроде $y' = y$, то мы на 100% уверены в том, что его решение существует и единственно (если задана начальная точка). Но это знание есть плод строгой теории, связанной аксиоматикой вещественных чисел! Чего уж говорить о квантовой теории, которая непостижимым образом работает на практике, хотя никто не может объяснить с материальной точки зрения, почему это происходит.

Тем не менее, иногда математика «не поспевает». Например, физики уже не одно десятилетие пользуются уравнением Навье-Стокса в практических целях, хотя строго математически не доказано существование и единственность его решения при заданных граничных условиях. И это один из примеров того, что математика является живой наукой, требующей развития, ставящей много новых интересных вопросов, дающей порой непредсказуемые, но всегда точные ответы.

5.4 Вероятности

В этом разделе мы сделаем шаг назад и вернемся к мере. Как уже отмечалось, если ограниченная мера на пространстве X такова, что $\mu(X) = 1$, то она называется **вероятностной мерой** и обычно обозначается буквой P .

Это — самое простое, короткое и самое непонятное определение вероятности. Действительно, трудно представить себе, какое отношение мера (объем, площадь) имеет к вероятностям. Тем не менее, именно к такому определению сводится аксиоматика Колмогорова, и именно оно является наиболее математическим определением вероятности из всех ранее существовавших.



*Андрей
Николаевич
Колмогоров*

5.4.1 Классическая вероятность

С точки зрения бытовой логики вероятность встретить крокодила на улицах Москвы равна 0.5, поскольку либо мы его встретим, либо нет. Тем не менее, все прекрасно понимают, что шансов встретить его все-таки гораздо меньше, чем не встретить. Просто потому, что таков наш опыт. Иначе гово-

ря, если за тысячи выходов на улицу в своей жизни мы ни разу не встретили крокодила, то, скорее всего, не встретим его и в следующий раз.

Именно это — *частотное* — определение вероятности положено в основу математической статистики. Именно так и следует интерпретировать слово *вероятность* в повседневной жизни.

Каким же образом это связано с мерой?

Проведем мысленный эксперимент. На этот раз с монеткой, чтобы ни одно животное не пострадало. Допустим, мы подбросили ее 100 раз, и в каждом исходе этой операции мы получили либо 1 (орёл), либо 0 (решка). Пусть $100p$ — количество орлов, $100q$ — количество решек (ясно, что $p + q = 1$ и $p, q \geq 0$). Теперь эти 100 подбрасываний мы представим себе как некоторый единичный объем, в котором есть два подобъема: все нулевые и все единичные исходы. Это — разбиение объема размером (мерой) 100 на две части, каждая из которых измерима и равна $100p$ и $100q$, а если рассматривать их как доли единичного объема, то эти доли равны p и q , соответственно.

Таким образом, у нас начинает складываться представление о том пространстве, где мы работаем с частотами–долями–вероятностями. Это — пространство всех возможных исходов при некоторой процедуре испытаний. Под испытаниями при этом можно понимать любые однородные действия, результат которых можно измерить числом (как мы понимаем теперь, под числом можно понимать довольно широкий класс математических объектов, но как правило это действительное или комплексное число или вектор).

Заметим, что пространство исходов — воображаемое, у него нет физических аналогов, поэтому и физическое понятие объема к нему неприменимо. «В физике» реализоваться может один или несколько исходов, а все одновременно варианты существуют лишь виртуально.

Коль скоро мы можем измерить результаты испытаний, мы имеем некую числовую функцию, определенную на пространстве испытаний. Пользуясь этой функцией, мы можем выделять в этом пространстве измеримые части (подобъемы) и рассчитывать их долю от общего объема пространства. Доля при этом может вычисляться пропорционально количеству элементов, а может и как-то иначе, особенно, если пространство бесконечное. Важно при этом, что доля всегда неотрицательная и не превосходит 1, а доля дизъюнктной суммы подобъемов равна сумме их долей.

Вот и связь частоты (вероятности) с мерой. Здесь мы абстрагировались от понятия пространства как некоего физического протяженного объекта и рассмотрели пространство событий, растянутое во времени (хотя можно и в пространстве, если у вас рота солдат по команде подбрасывает монетки, но этот эксперимент получится дороже).

Пространство с вероятностной мерой называется **вероятностным пространством**, его элементы — *элементарными событиями* (они неделимы, как кварки в физике), а те их совокупности, которые можно так или ина-

че измерить с помощью частот — просто *событиями*. Обычно вероятностное пространство (его носитель) обозначается Ω .

Для понимания механики вероятностей практически всегда можно держать в голове схему с подбрасыванием монеты. Но в ряде случаев полезно также иметь следующую картинку, которую мы позаимствуем из физики. Пусть у нас имеется источник электронов (электронная пушка), а перед ней установлена шторка с узкой щелью (единственной!), за которой находится экран–детектор. В данном эксперименте, как известно, электроны ведут себя как частицы: пролетая через щель, они попадают на экран практически в одну и ту же точку в центре, но с некоторым рассеянием (т. к. щель имеет ненулевую ширину, а кроме того, электроны могут отражаться от стенок и краев этой щели). Чем дальше от центра экрана, тем ниже частота попадания туда электронов. В итоге на экране мы наблюдаем плотность (насыщенность) пучка электронов, очень похожую на гауссово нормальное распределение.

Вероятностная механика!

Здесь событием является попадание электрона на детектор и, вообще говоря, таких событий континуум, если координату точки попадания измерять вещественным числом (что наиболее естественно). Поэтому вероятностным пространством является \mathbb{R} (считаем, что электрон может отклониться на любой угол и улететь достаточно далеко от центра), а наблюдаемая картина частот — плотностью вероятности попадания электрона в точки на \mathbb{R} . Дело в том, что попадание в каждую конкретную координату происходит с нулевой вероятностью, но попадание в любой непустой интервал имеет вполне определенную положительную частотность, которая определяется шириной этого интервала и его местоположением на прямой. Если длину интервала делать бесконечно малой, то колебанием плотности вероятности можно пренебречь, и тогда вероятность попадания в интервал длины dx в точке x будет выражаться формулой $\rho(x)dx$, где $\rho(x)$ и называется плотностью вероятности. Величина плотности определяет плотность потока электронов в данной точке (точнее, в данном бесконечно малом интервале). При этом плотность вероятности всегда нормируется условием

$$\int_{\mathbb{R}} \rho(x)dx = 1,$$

где интегрирование производится по всем элементарным событиям (в нашем случае по \mathbb{R}).

Заметим, что если в случае конечного пространства элементарных событий мы можем считать алгебру событий полной (включающей все подмножества), то в случае континуального пространства для корректного определения вероятностной меры обычно требуется борелевская сигма-алгебра.

Это связано вот с чем. Для погружения теории вероятностей в Анализ нам необходимо оперировать числовыми функциями. Один пример такой функ-

ции мы уже привели выше — это плотность вероятности. Но плотность существует не всегда. Более общим понятием является функция распределения

$$F(x) \doteq P(-\infty, x),$$

представляющая собой меру левого луча на \mathbb{R} (в случае, когда x является вектором из \mathbb{R}^n , берется вероятность прямоугольного конуса с вершиной в точке x , который является аналогом луча в n -мерном пространстве). И вот тут мы видим прямую связь с борелевскими множествами, поскольку, как мы помним, лучи образуют предбазу интервальной топологии. Это значит, что сигма-алгебра событий должна быть как минимум борелевской. В то же время, попытки расширить алгебру могут привести к нарушению счетной аддитивности меры (достаточно вспомнить пример множества Витали, неизмеримого по Лебегу). Таким образом, если речь идет о вещественных или комплексных пространствах, то обычно предполагается, что алгебра событий является борелевской относительно стандартной топологии. Это же обеспечивает измеримость функции распределения $F(x)$ относительно меры Лебега.

Вернемся теперь к подбрасываниям монетки и вспомним, что при каждом исходе мы производили некоторое измерение: если выпадал орёл, мы записывали 1, а если решка — 0. Иначе говоря, мы построили функцию $\xi : \Omega \rightarrow \{0, 1\}$. Эта функция, очевидно, измерима относительно меры P на Ω и лебеговой меры в \mathbb{R} , поскольку мы предположили, что алгебра на Ω является полной.

Помимо упомянутой ξ мы могли бы рассмотреть и другие измеримые функции, например, если занумеровать испытания, то можно выдавать номер испытания, если выпал орел, и минус номер, если решка. Однако за случайность (непредсказуемость) результата измерения все равно будет отвечать бинарная монетка.

В примере с электроном измеримой функцией от события будет координата этого события, т. е. точка попадания электрона в экран. При этом мы можем рассмотреть и другие измеримые функции от x , случайность значений которых при каждом попадании электрона на детектор будет определяться плотностью $\rho(x)$.

Таким образом, если мы проводим серию измерений некоторой величины, значения которой выпадают с той или иной вероятностью, то мы имеем дело с измеримой числовой функцией, определенной на пространстве элементарных событий.

Измеримая числовая функция, определенная на вероятностном пространстве, называется **случайной величиной**.

Случайная величина индуцирует вероятностную меру на то числовое пространство, в котором она действует. Например, если $\xi : \Omega \rightarrow \mathbb{R}$ есть случайная

величина, то для всякого борелевского множества $A \subseteq \mathbb{R}$ можно определить счетно-аддитивную меру по правилу

$$P_{\xi}(A) = P\{w \in \Omega \mid \xi(w) \in A\}, \quad (5.14)$$

если, конечно, вероятность на Ω счетно-аддитивна. Множество, стоящее в правой части, обычно записывается короче: $\{\xi \in A\}$. Это позволяет избежать упоминания исходного вероятностного пространства Ω и полностью «работать» в \mathbb{R} . Соответственно, функция индуцированного распределения будет

$$F_{\xi}(x) = P\{\xi < x\} = P_{\xi}(-\infty; x).$$

Подводя итоги вводной части, еще раз отметим, что под вероятностью мы чаще всего понимаем частотность события. При этом события могут быть описаны как подмножества множества произвольной природы. Обычно событие описывается некоторой истинностной формулой, определяющей множество элементарных событий. Например, событие, связанное с подбрасыванием монетки, может быть описано формулой $x = 1$, где x — верхняя сторона монетки после падения, а 1 обозначает орла. В случае пучка электронов, если r означает отклонение от центра экрана, формула может быть записана как $r \leq 1$, а событие будет состоять в том, что электрон отклонился от центра не более чем на 1. Если мы обозначаем событие буквой A , а множество всех возможных событий Ω , то $A \subseteq \Omega$. Предполагается, что A должно быть измеримым относительно заданной на Ω вероятностной меры.

Измеримость события означает, что существует мера P_A (частота встречаемости событий, описанных формулой, задающей событие A). Поскольку над событиями нужно часто производить теоретико-множественные операции (пересечение, разность и т.д.), их принято погружать в фигурные скобки, т.е. P_A и $P\{A\}$ — это равные термы.

Если, кроме того, на событиях задана некоторая числовая функция ξ (например, различные статистические метрики, которые считает Google Analytics), то мы говорим о случайной величине, для которой можно измерить частотность попадания в определенный интервал через частотность тех событий, которые она меряет. Так у нас появляется индуцированная в \mathbb{R} вероятностная мера P_{ξ} и соответствующая функция распределения F_{ξ} .³¹

Наконец, в ряде задач приходится кластеризовать исходное пространство Ω на более мелкие фрагменты и изучать те или иные события внутри одного из фрагментов. Например, считать вероятность попадания электрона слева от центра при условии, что он попал в единичный интервал $r \leq 1$. Тем самым, мы, с одной стороны, исключаем статистические выбросы (далеко улетевшие электроны), с другой стороны — проверяем симметричность пучка. Для этих

³¹На самом деле, для определения меры P_{ξ} достаточно знать F_{ξ} , которая задает меру левых лучей на \mathbb{R} .

целей существует такое понятие как условная вероятность. Мы берем подмножество $\Omega' \subset \Omega$ и хотим его рассматривать как новое пространство. Для этого нам нужно все вероятности нормировать так, чтобы у нас получилось вероятностное подпространство. Это делается следующим образом:

$$P\{A|\Omega'\} = \frac{P\{A \cap \Omega'\}}{P\{\Omega'\}}.$$

То есть все события A мы редуцируем до интересующего нас события Ω' , рассматривая пересечение $A \cap \Omega'$, а затем нормируем на вероятность $P\{\Omega'\}$, чтобы получить вероятностное пространство. Символ $P\{A|\Omega'\}$ называется **условной вероятностью** события A при условии Ω' .

Здесь как A , так и Ω' могут быть заданы непосредственно как часть Ω , а могут индуцировать условие через некоторые случайные величины, заданные на Ω . Это позволяет, например, рассматривать условные случайные величины.

Комментарий 22. О пользе формулы Байеса

Напомним знаменитую формулу английского священника Томаса Байеса:

$$P\{A|B\} = \frac{P\{B|A\} P\{A\}}{P\{B\}}$$

для любых событий $A, B \subseteq \Omega$ (предполагается, что $P\{B\} > 0$).

Формула выводится тривиально из определения условной вероятности, но получена была еще до определения как условной вероятности, так и вообще вероятности как строгого математического понятия. Чем и ценна.

Вернемся на минутку к моделям машинного обучения (см. комментарий 8). Предположим, что у нас есть обучающая выборка, в которой находятся объекты со свойствами C_1, \dots, C_n , разделенные на два класса A и B .³² Мы можем посчитать следующие эмпирические вероятности (частоты):

$$P\{A\}, P\{B\}, P\{C_k|A\}, P\{C_k|B\}, P\{C_1, \dots, C_n|A\}, P\{C_1, \dots, C_n|B\}.$$

Некоторые из них можно посчитать быстро, некоторые могут потребовать много ресурсов для расчета.



Томас Байес

³² C_k в данном случае нужно рассматривать как измеримые характеристики, могущие принимать различные значения. В обучающей выборке много разных объектов, отличающихся различными комбинациями значений этих характеристик.

Пользуясь формулой Байеса, имеем:³³

$$\begin{aligned} P\{A|C_1, \dots, C_n\} &= \frac{P\{C_1, \dots, C_n|A\} P\{A\}}{P\{C_1, \dots, C_n\}}, \\ P\{B|C_1, \dots, C_n\} &= \frac{P\{C_1, \dots, C_n|B\} P\{B\}}{P\{C_1, \dots, C_n\}} \end{aligned} \quad (5.15)$$

Далее, пусть появляется новый объект, у которого мы измерили характеристики C_1, \dots, C_n , т.е. получили их конкретные значения, и нам требуется отнести его к классу A или классу B (предполагается, что других классов нет). Пользуясь этими формулами, мы можем:

- оценить вероятности принадлежности нового объекта к классу A и классу B (на основе эмпирических вероятностей, полученных на обучающей выборке);
- принять решение об отнесении нового объекта к тому или иному классу в зависимости от величины соответствующей вероятности (принцип максимума правдоподобия);
- вычислить информационную энтропию (по Шённону)

$$\begin{aligned} H &= -(P\{A|C_1, \dots, C_n\} \log_2 P\{A|C_1, \dots, C_n\} + \\ &+ P\{B|C_1, \dots, C_n\} \log_2 P\{B|C_1, \dots, C_n\}). \end{aligned}$$

Заметим, однако, что на практике обычно все не так хорошо, как в теории, поэтому сложную вероятность $P\{C_1, \dots, C_n|A\}$ требуется как-то декомпозировать, чтобы не хранить вероятности всех возможных комбинаций значений C_1, \dots, C_n , да еще в сочетании с признаком разделения на классы A и B .

Поэтому мы «наивно» (откуда и название метода: Naïve Bayes) предполагаем условную независимость характеристик C_1, \dots, C_n ³⁴, что упрощает формулы (5.15) до

$$\begin{aligned} P\{A|C_1, \dots, C_n\} &= \frac{P\{C_1|A\} \dots P\{C_n|A\} P\{A\}}{P\{C_1, \dots, C_n\}}, \\ P\{B|C_1, \dots, C_n\} &= \frac{P\{C_1|B\} \dots P\{C_n|B\} P\{B\}}{P\{C_1, \dots, C_n\}}. \end{aligned} \quad (5.16)$$

³³Перечисление событий через запятую означает их одновременное выполнение.

³⁴Заметим, что здесь мы постоянно «путаем» характеристики с событиями. Более правильно было бы говорить о наблюдаемых значениях характеристик c_k , а под событиями C_k понимать некоторое соотношение вроде $a_k < c_k < b_k$ или $c_k = r$ и т.п. Но пока это не вводит в заблуждение, принято отождествлять события и сами величины. Здесь мы видим очередное проявление архетипа **базового множества**, но в применении к событиям.

Условную независимость можно показать на следующем примере. Пусть римские легионы во главе с императором атакуют варваров в окрестностях Рима. Если мы ничего не знаем об исходе битвы, то судьба Рима (C_1) и судьба императора (C_2) вполне могут быть зависимыми. Однако, если битва проиграна (событие A), то судьба Рима уже никак не зависит от судьбы императора — он неминуемо падёт под натиском варваров. Таким образом, условные события $C_1|A$ и $C_2|A$ становятся независимыми, и их совместная вероятность раскладывается в произведение.

Вообще, если выше мы говорили о том, что вероятность — это частота, то при изучении условных вероятностей правильнее понимать их как меру нашего незнания. Чем меньше мы знаем о монетке, тем более случайным кажется нам ее поведение. Такой взгляд на вероятность помогает развить интуицию при работе с байесовскими вероятностями и случайными процессами во времени.

Глядя теперь на формулы (5.16), мы видим, что оценка класса для нового объекта по его наблюдаемым характеристикам сводится к знанию декомпозированных *априорных* условных вероятностей $P\{C_k|A\}$, $P\{C_k|B\}$, $P\{A\}$, $P\{B\}$, полученных на обучающей выборке или заданных моделью. Вероятность $P\{C_1, \dots, C_n\}$ при этом является лишь нормировочным множителем и не играет никакой роли при принятии решения (кроме того, ее легко получить как сумму числителей в дробях (5.16)).

Вычисляя *апостериорные* вероятности $P\{A|C_1, \dots, C_n\}$ и $P\{B|C_1, \dots, C_n\}$ (без нормировки), мы получаем два числа, вообще говоря, разных. И если первая вероятность больше второй, то мы решаем, что вновь прибывший объект принадлежит классу A , иначе — классу B .

Ясно, что байесовский подход масштабируется на случай, когда классов может быть много. Больше того, наблюдая параметры C_k , которые легко измерить, мы можем оценить недостающие параметры A_k , рассматривая их в данной модели как классы. Получив апостериорное распределение для $\{A_k\}$, мы уже можем рассматривать набор $\{A_k, C_k\}$ как наблюдаемые характеристики, и с их помощью оценивать какие-то третьи величины B_k , и т.д. Конечно, с каждой итерацией точность метода падает, но вычислительная скорость такого итерационного подхода в ряде задач оказывается важнее лишних 2-3-х процентов точности ответа.

Удивительно, но даже при таком наивном подходе в последние годы (с 1990-х до нашего времени) в ряде задач машинного обучения байесовский метод классификации оказывается эффективнее метода SVM и некоторых других. Прежде всего, он выигрывает по скорости (из-за декомпозиции условных вероятностей), но при этом не сильно теряет в точности.

Энтропия H , приведенная выше, используется для понимания того, насколько полученный результат информативен. Энтропия считается мерой недостатка информации, поэтому, чем больше энтропия полученного распределения, тем хуже качество полученного распределения (в случае байесовского метода вы-

сокая энтропия означает, что мы почти наугад выдаем ответ классификатора, и тогда непонятно, зачем он вообще нужен). Действительно, для сильно «размазанного» распределения с более-менее одинаковыми вероятностями энтропия растет со скоростью $\ln n$, где n — количество этих вероятностей. В то же время, если распределение схлопывается в одну-две точки, то энтропия получается достаточно близкой к нулю. Иначе говоря, энтропия есть некий интегральный показатель детерминированности распределения.

Информационная энтропия используется в математической теории связи и многих других вероятностных приложениях, например, при построении деревьев решений требуется, чтобы с ростом дерева росла информация (убывала энтропия), тогда классификатор на основе этого дерева будет работать корректно.

О характеристических функциях

Ранее мы упоминали термин «характеристическая функция» применительно к интегральным преобразованиям. Заметим, что случайные величины и вероятности настолько хорошо ложатся в теорию меры и интеграла, что мы естественным образом можем применять формализм интегральных преобразований и в этой области.

Как мы видели выше, случайная величина $\xi : \Omega \rightarrow \mathbb{R}$ индуцирует меру P_ξ на \mathbb{R} . Рассмотрим производящую функцию

$$F(\xi, z) \doteq \int_{-\infty}^{\infty} z^x dP_\xi, \quad (5.17)$$

где интегрирование ведется по переменной x , а z является параметром. Это не что иное как интегральное преобразование вероятностного распределения P_ξ с ядром z^x или, проще говоря, **производящая функция случайной величины** ξ . Подставляя $z = e^{it}$, мы получаем **характеристическую функцию случайной величины** ξ . Стандартное обозначение для нее:

$$\phi(\xi, t) \doteq F(\xi, e^{it}).$$

Интересно, что если мы вернемся в исходное пространство Ω , то можно заметить, что

$$F(\xi, z) = \int_{\Omega} z^\xi dP, \quad \phi(\xi, t) = \int_{\Omega} e^{it\xi} dP. \quad (5.18)$$

Отсюда, в частности, видно, что производящая и характеристическая функции нелинейны относительно алгебраических операций над случайными величинами, как бы нам того ни хотелось. Напомним, что интегральные преобразования линейны относительно подынтегральной функции и меры.

Но в представлении (5.17) выражение z^x является ядром преобразования, а подынтегральная функция есть тождественная единица, которая не зависит от выбора ξ . В то же время, очевидно, что $P_{\xi+\eta} \neq P_\xi + P_\eta$, иначе это не было бы вероятностью.

Несмотря на то, что представление в виде (5.18) кажется наиболее естественным для случайных величин, как правило, на практике мы работаем с индуцированным в \mathbb{R} распределением, и потому пользуемся стандартной для интегральных преобразований формой (5.17). Например, распределение Бернулли B_p задается двумя точками: 0 с вероятностью q , и 1 с вероятностью p . В этом случае

$$F(B_p, z) = q + zp, \quad \phi(B_p, t) = q + pe^{it}.$$

В двух основных случаях (а также для их арифметических комбинаций) — дискретном и абсолютно непрерывном, — мы видим уже прямую связь интегральных преобразований с вероятностями. Действительно, обобщим случай распределения Бернулли на произвольное дискретное распределение $p = (p_0, p_1, \dots)$ (имеется ввиду, что $P\{\xi = k\} = p_k$). Тогда мы получим, что

$$F(\xi, z) = \sum_k z^k p_k, \quad \phi(\xi, t) = \sum_k e^{itk} p_k,$$

т. е. производящая и характеристическая функции являются преобразованиями последовательности p .

Если же распределение ξ абсолютно непрерывно, т. е. по определению $dP_\xi = f(x)dx$, где $f(x)$ называется **плотностью распределения** и интегрируема на всем \mathbb{R} , то мы получим, что

$$F(\xi, z) = \int_{-\infty}^{\infty} z^x f(x)dx, \quad \phi(\xi, t) = \int_{-\infty}^{\infty} e^{itx} f(x)dx,$$

т. е. производящая и характеристическая функции являются преобразованиями плотности f относительно меры Лебега.

Соответственно, эти преобразования линейны относительно p и f . Однако стоит заметить, что произвольная линейная комбинация $ap + b\tilde{p}$ или же $af + b\tilde{f}$ не имеет никакого практического смысла для вероятностей. Смысл будет лишь в том случае, если коэффициенты при распределениях (плотностях) являются весовыми, т. е. они положительны и их сумма равна 1. В этом случае мы будем иметь дело с некоторой взвешенной комбинацией вероятностей (плотностей), которая не будет соответствовать линейной комбинации исходных случайных величин.

Заметим также, что в принципе любая вероятностная мера может быть представлена как взвешенная сумма абсолютно непрерывной, дискретной и сингулярной меры. Точнее, если $F(x) = P(-\infty; x)$ — функция распределения меры P , то

$$F(x) = (aF_{disc} + bF_{cont} + cF_{sing})(x),$$

где $a, b, c \geq 0$, $a + b + c = 1$ и F_{disc} — дискретная мера, F_{cont} — абсолютно непрерывная, F_{sing} — сингулярная.

Сингулярной называется мера, полностью сосредоточенная на множестве лебеговой меры ноль (например, на канторовом совершенном множестве). К сингулярной мере можно отнести и дискретную, если задача не требует выделения ее в виде отдельного компонента исходной меры. В то время как абсолютно непрерывная мера не сводится к сингулярной в силу теоремы Радона–Никодима [81].

Из такого разложения меры следует, что практически всегда (т. е. когда можно пренебречь сингулярной компонентой) производящую (и характеристическую) функцию можно записать как взвешенную сумму интеграла от плотности и суммы с дискретными вероятностями, т. е. как взвешенную сумму интегральных преобразований плотности и вектора дискретных вероятностей:

$$F(\xi, z) = a \int_{-\infty}^{\infty} z^x f(x) dx + b \sum_k z^k p_k.$$

Таким образом, мы видим полное погружение вероятностей в теорию интегральных преобразований, однако извлечь из этого погружения мы можем только то, что не связано напрямую с алгеброй пространства случайных величин и распределений. Например, мы можем извлечь формулы для обратного преобразования, т. е. научиться получать дискретные вероятности p_k и плотность абсолютно непрерывного распределения через известные характеристические функции (**формула обращения**):

$$p_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itk} \phi(\xi, t) dt, \quad f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi(\xi, t) dt.$$

Еще одним существенным следствием того, что характеристические функции погружаются в теорию линейных операторов, является теорема непрерывности.

Теорема 5.7 (Леви о непрерывности). *Пусть задана последовательность $\{\xi_n\}$ случайных величин, определенных на пространстве $\langle \Omega, \mathcal{P} \rangle$. Данная последовательность слабо сходится (относительно меры \mathcal{P}) к случайной величине ξ тогда и только тогда имеет место поточечная сходимость*

$$\phi(\xi_n, t) \rightarrow \phi(\xi, t)$$

в каждой точке $t \in \mathbb{R}$ к некоторой характеристической функции $\phi(\xi, t)$.

Данная теорема широко используется для получения предельных теорем теории вероятностей. Например, **центральная предельная теорема** утверждает, что сумма $\xi_1 + \dots + \xi_n$ независимых одинаково распределенных случайных величин с мат.ожиданием a и конечной дисперсией $\sigma^2 > 0$ слабо сходится

к нормальному закону распределения с мат.ожиданием na и дисперсией $n\sigma^2$. Это утверждение достаточно просто выводится с помощью характеристических функций и теоремы о непрерывности.

Комментарий 23. О последовательности Коллатца.

Ранее (стр. 132) мы определили последовательность Коллатца (она же си-ракузская последовательность), которая, стартуя с произвольного натурального числа $x_0 > 0$, вычисляется итеративно следующим способом: если предыдущее число x_{n-1} четное, то $x_n = x_{n-1}/2$, иначе $x_n = 3x_{n-1} + 1$. Гипотеза Коллатца состоит в том, что независимо от стартового числа x_0 данная последовательность рано или поздно придет к циклу $4 \rightarrow 2 \rightarrow 1$.

Мы приведем сейчас оценочное доказательство гипотезы Коллатца, основанное на предположениях о вероятности встретить четное число на каждом шаге. Если рассуждать совсем отвлеченно, то мы можем считать, что поскольку четных и нечетных чисел поровну в любом достаточно большом отрезке натуральных чисел, вероятность встретить четное число равна $1/2$.

Теперь запишем процесс построения последовательности знаками $+$ и $-$, отмечая плюсом рост последовательности, а минусом — убывание. Например:

$$\begin{aligned} 7 \xrightarrow{+} 22 \xrightarrow{-} 11 \xrightarrow{+} 34 \xrightarrow{-} 17 \xrightarrow{+} 52 \xrightarrow{-} 26 \xrightarrow{-} 13 \xrightarrow{+} 40 \\ \xrightarrow{-} 20 \xrightarrow{-} 10 \xrightarrow{-} 5 \xrightarrow{+} 16 \xrightarrow{-} 8 \xrightarrow{-} 4 \xrightarrow{-} 2 \xrightarrow{-} 1. \end{aligned}$$

Далее заметим, что всякий раз после $+$ идет $-$, поскольку если k нечетное, то $3k + 1$ заведомо четное. Поэтому блоки вида $(+-)$ нужно рассматривать как безусловное явление в последовательности. Перепишем 7-последовательность с учетом блоков:

$$(+-)(+-)(+-) - (+-) - -(+-) - - -$$

Теперь, считая, что блоки $(+-)$ и $-$ каждый раз возникают случайно и независимо от предыдущих испытаний (именно это и есть слабое место вероятностной проверки гипотезы Коллатца), мы получаем последовательность испытаний Бернулли с генерирующей случайнов величиной ξ , которая равна 1, если встречается блок $(+-)$, и 0, если $-$. Оба значения имеют вероятность $1/2$. Иначе говоря, мы имеем дело с суммой независимых одинаково распределенных случайных величин

$$S_n = \xi_1 + \dots + \xi_n,$$

и значение S_n этой суммы равно количеству блоков $(+-)$ в знаковой нотации последовательности Коллатца. При этом $ES_n = n/2$, $D S_n = n/4$.

Для удобства теперь будем считать, что шагом последовательности Коллатца является либо деление на 2, либо умножение на 3 с добавлением 1 и делением на 2, т. е. блок $(+-)$ мы рассматриваем как один шаг:

$$7 \xrightarrow{+-} 11 \xrightarrow{+-} 17 \xrightarrow{+-} 26 \xrightarrow{-} 13 \xrightarrow{+-} 20 \xrightarrow{-} 10 \xrightarrow{-} 5 \xrightarrow{+-} 8 \xrightarrow{-} 4 \xrightarrow{-} 2 \xrightarrow{-} 1$$

Оценим элемент x_n . Заметим, что если x_{n-1} был четный, то $x_n = x_{n-1}/2$, а если нечетный, то $x_n = x_{n-1} * 1.5 + 0.5 = (1.5 + 0.5/x_{n-1})x_{n-1}$. Чтобы оценить довесок $0.5/x_{n-1}$, приведем следующие соображения. Мы можем считать, что x_{n-1} (и вообще все члены последовательности) больше или равны некоторого числа K_0 . Дело в том, что если мы вручную (или на компьютере) проверили³⁵ истинность гипотезы Коллатца для стартовых чисел от 1 до $K_0 - 1$, то как только в какой-то новой последовательности встретится элемент $x_k < K_0$, последовательность сойдется к циклу $4 \rightarrow 2 \rightarrow 1$ (по проверенному вручную). Это значит, что мы можем рассматривать только такие последовательности, у которых заведомо все члены $\geq K_0$.

Итак, предполагая, что $x_{n-1} \geq K_0$, находим оценку для нечетного случая: $x_n \leq (1.5 + 0.5/K_0)x_{n-1}$. Таким образом, каждый шаг дает либо увеличение последовательности не более чем в $1.5 + 0.5/K_0$ раз, либо уменьшение ровно в 2 раза. Если при построении элемента x_n использовалось k шагов $(+)$, то

$$x_n \leq x_0(1.5 + 0.5/K_0)^k / 2^{n-k} = x_0(3 + 1/K_0)^k / 2^n.$$

Вот здесь мы и делаем недоказанное допущение. Мы предположим теперь, что k — это число успехов в схеме Бернулли, т.е. заменим его на случайную величину S_n , в результате чего будем оценивать детерминированное число x_n случайной величиной

$$X_n = x_0(3 + 1/K_0)^{S_n} / 2^n,$$

предполагая, что вероятность события $\{x_n > X_n\}$ ничтожно мала (стремится к нулю с ростом n).

Оценим вероятность того, что $X_n > x_0(3 + 1/K_0)^A$ для некоторого $A \in \mathbb{R}$:

$$\begin{aligned} P\{X_n > x_0(3 + 1/K_0)^A\} &= P\{(3 + 1/K_0)^{S_n} / 2^n > (3 + 1/K_0)^A\} = \\ &= P\left\{S_n - n \frac{\ln 2}{\ln(3 + 1/K_0)} > A\right\} = P\left\{\frac{S_n - n/2}{0.5\sqrt{n}} > \frac{A + n(\frac{\ln 2}{\ln(3 + 1/K_0)} - \frac{1}{2})}{0.5\sqrt{n}}\right\}. \end{aligned}$$

Последняя вероятность в силу центральной предельной теоремы стремится к нулю, поскольку $\frac{S_n - n/2}{0.5\sqrt{n}}$ асимптотически нормально распределена с параметрами $(0; 1)$ и величина $\frac{\ln 2}{\ln(3 + 1/K_0)} - \frac{1}{2}$ положительна при $K_0 > 1$.

Заметим, что $\frac{\ln 2}{\ln(3 + 1/K_0)} - \frac{1}{2} > 0$ уже при $K_0 = 2$. Но большие значения K_0 позволяют расширить наше предположение о распределении генерирующей случайной величины. Если вероятность нечетного $P\{\xi = 1\} = p$, то $ES_n = np$, $DS_n = np(1 - p)$ и

$$P\{X_n > (3 + 1/K_0)^A\} = P\left\{\frac{S_n - np}{\sqrt{np(1 - p)}} > \frac{A + n(\frac{\ln 2}{\ln(3 + 1/K_0)} - p)}{\sqrt{np(1 - p)}}\right\}.$$

³⁵На 2019 год проверено свыше 10^{18} первых чисел натурального ряда, т.е. можно смело считать, что $K_0 = 10^{18}$, но для наших целей годится любое $K_0 > 1$.

Данная вероятность стремится к нулю в том случае, когда $\frac{\ln 2}{\ln(3+1/K_0)} - p > 0$. Это последнее неравенство выполняется при достаточно больших K_0 и p заведомо меньше $(\ln 2)/(\ln 3) = 0.6309 \dots$. Иначе говоря, интересующая нас вероятность стремится к нулю и в том случае, если немного ослабить представления о вероятности выпадения нечетного числа, считая их частотность немного выше частотности четных чисел.

Ясно, что, выбирая отрицательное A , мы можем сделать величину $x_0(3 + 1/K_0)^A$ сколь угодно малой. А это значит, что X_n сходится к нулю по вероятности. Конечно, это не означает, что она перестанет принимать сколь угодно большие значения после некоторого номера n (как это было бы в случае обычного предела), но частота ухода X_n от нуля будет становиться ничтожно малой. Но тогда это значит, что X_n достаточно часто будет принимать нулевые значения, что в отношении последовательности Коллатца выглядит абсурдно.

Отсюда следует, что, по крайней мере, одно из наших исходных предположений неверно. Например, то, что $x_n \geq K_0$ даже при $K_0 = 2$. Но тогда мы получаем, что x_n должна принять значение 1 при некотором n , и это доказывает гипотезу Коллатца.

Второй вариант разрешения противоречия состоит в том, что наша оценка $x_n < X_n$ ошибочна и что пользоваться бернуллиевской вероятностной схемой здесь неуместно.

Тем не менее, данный пример рассуждений хорошо иллюстрирует как методику использования вероятностных методов в комбинаторике, так и необходимость проверять адекватность используемой модели.

Помимо возможности рассматривать производящие и характеристические функции как линейные непрерывные операторы, мы имеем достаточно весомый набор их собственных свойств, не связанных с интегральными преобразованиями:

ChF1 $\phi(\xi, t)$ непрерывна по $t \in \mathbb{R}$;

ChF2 $\phi(\xi, 0) = 1$ и $|\phi(\xi, t)| \leq 1$ для всех $t \in \mathbb{R}$;

ChF3 $\phi(\xi, t)$ является положительно определенной, т.е. матрица $\Phi = (\phi_{ij})$ положительно определена, где $\phi_{ij} = \phi(\xi, t_i - t_j)$, при любом выборе переменных t_1, \dots, t_n и любом $n \geq 1$;

ChF4 $\phi(t)$ — хар.функция случайной величины тогда и только тогда, когда она удовлетворяет свойствам ChF1–ChF3 (теорема Бохнера–Хинчина, см. [97]);

ChF5 $\phi(\xi, t)$ однозначно определяет распределение P_ξ , которое случайная величина ξ индуцирует на \mathbb{R} ;

CHF —
швейцарский
франк??

ChF6 хар. функция суммы независимых случайных величин есть произведение хар.функций этих величин;

$$\text{ChF7 } \phi(a\xi, t) = \phi(\xi, at)$$

$$\text{ChF8 } \phi(\xi, t) = \overline{\varphi(\xi, -t)} \text{ (эрмитовость хар.функции);}$$

ChF9 моменты ξ выражаются через производные соответствующих порядков:

$$i^n E\xi^n = \frac{d^n}{dt^n} \phi(\xi, t) \Big|_{t=0}.$$

Вспоминая еще раз о производящей функции, отметим, что *факториальные моменты* ξ выражаются через производные $F(\xi, z)$ соответствующих порядков:

$$E\xi^n = \frac{d^n}{dz^n} F(\xi, z) \Big|_{z=0}.$$

Как видим, несмотря на отсутствие линейности по случайной величине, мы все-таки наблюдаем некоторое сохранение алгебры оператором характеристической функции. Именно, если случайные величины ξ и η независимы(!), то

$$\phi(a\xi + b\eta; t) = \phi(a\xi, t)\phi(b\eta, t) = \phi(\xi, at)\phi(\eta, bt),$$

т.е. для независимых величин их сумма отображается в произведение хар. функций, а коэффициенты меняют масштаб аргумента хар. функции. Так себе линейность, но некоторую аналогию провести можно. Обычно переход суммы в произведение сопряжен с экспоненциальным преобразованием, и оно здесь прямо присутствует в определении хар. функции.

Мы не будем далее углубляться в сторону всем хорошо известных свойств случайных величин и распределений, т.к. их можно найти в любом учебнике по теории вероятностей (например, [77]) и/или мат.статистике (например, [82]). Вместо этого мы, как обычно, займемся возведением межотраслевых мостов.

Алгебраическая интерпретация

Рассмотрим вероятностное пространство $\langle \Omega, \mathcal{P} \rangle$ (мы опускаем значок для алгебры множеств, т.к. она просто совпадает с $\text{dom}(\mathcal{P})$). Пусть далее

$$\Xi \rightleftharpoons \{ \xi \mid \xi : \Omega \rightarrow \mathbb{R} \text{ — измерима} \},$$

т.е. множество всех случайных величин.

Нетрудно видеть, что Ξ является линейным пространством над \mathbb{R} , более того, оно является коммутативной алгеброй над \mathbb{R} , либо же его можно рассматривать как множество самосопряженных элементов более широкой алгебры измеримых функций из Ω в \mathbb{C} . Далее выделим в Ξ подалгебру $\mathcal{A} \subseteq \Xi$, элементы ξ которой обладают тем свойством, что интеграл

$$E_P \xi = \int_{\Omega} \xi dP$$

сходится и конечен. Иначе говоря, алгебра \mathcal{A} содержит только те случайные величины, которые имеют конечное математическое ожидание.

Заметим, что функционал $E_P \xi$ линеен и, кроме того, непрерывен относительно сходимости по вероятности (если для Ξ выполняется свойство банаховости).

Таким образом, для произвольного вероятностного пространства мы получаем коммутативную алгебру случайных величин с заданным на ней линейным (непрерывным) функционалом $\varphi(\xi) = E\xi$, обладающим свойством $\varphi(1) = 1$. Запомним такую алгебраическую интерпретацию вероятностного пространства до конца главы, а именно, до таблицы 5.4!

Еще один пассаж, который нам стоит совершить, находясь в рамках классической теории вероятностей, заключается в следующем. Мы уже ранее видели, что можно по-разному подходить к определению вероятностного распределения. Именно, мы можем рассматривать некое вероятностное пространство $\langle \Omega, P \rangle$ с заданной мерой P как первичный объект, а случайные величины вводить как (вторичные) измеримые функции из Ω в \mathbb{R} (или в \mathbb{R}^n). В этом случае у нас вероятности «заморожены», а случайные величины могут быть разными. Назовем такой подход *функциональным*.

Второй подход заключается в том, что случайная величина ξ индуцирует некоторое распределение P_ξ на множество \mathbb{R} (или \mathbb{R}^n) по формуле (5.14) и, соответственно, вместо случайной величины ξ мы можем изучать это индуцированное распределение, а еще лучше — его функцию распределения $F(x) = P\{\xi < x\}$. Тогда переменным объектом у нас уже становится распределение, но не на абстрактном пространстве Ω , а на \mathbb{R} (или \mathbb{R}^n). Первичным объектом у нас становится \mathbb{R} с мерой Лебега. Назовем такой подход *дистрибутивным*.

Во втором подходе возникает некоторая потеря информации о вероятностях, если функция ξ не является инъективной. Например, если $\Omega = \{1, \dots, 2d\}$, $P = (p_1, \dots, p_{2d})$ а функция $\xi(k) = k \bmod 2$, то распределение P_ξ будет биномиальным: $P_\xi\{0\} = p_2 + p_4 + \dots + p_{2d}$, $P_\xi\{1\} = p_1 + p_3 + \dots + p_{2d-1}$. Распределение P_ξ не может дать адекватное представление об исходном распределении P на Ω . С другой стороны, если мы рассматриваем все индуцированные распределения, то по их совокупности уже получаем более-менее

полную картину об исходном пространстве (но если мощность Ω превосходит $2^{\mathfrak{c}}$, то потери информации о распределении неизбежны).

В дистрибутивном подходе у нас нет случайной величины, заданной на вероятностном пространстве \mathbb{R} , хотя «овеществить» ее достаточно просто. Случайной величиной в этом случае можно назвать тождественную функцию на \mathbb{R} , т. е. id .³⁶ При этом, как легко видеть,

$$\int_{\mathbb{R}} \text{id} dF(x) = \int_{\mathbb{R}} x dF(x) = \int_{\Omega} \xi dP = \mathbb{E}_P \xi,$$

т. е. среднее значение id по распределению $F(x)$ совпадает со средним значением ξ в функциональном подходе.

Наконец, отметим, что теория вероятностей не была бы полноценной математической дисциплиной, если бы не имела обобщений в области комплексных чисел и не имела бы универсальной алгебраической модели. К этим обобщениям мы теперь и перейдем.

5.4.2 Квантовая вероятность

Рассмотрим для начала конечное вероятностное пространство $\Omega = \{1, \dots, d\}$ с вероятностями $P = (p_1, \dots, p_d)$ и случайную величину $X : \Omega \rightarrow \mathbb{R}$. Заметим, что вектор P можно рассматривать как элемент пространства \mathbb{R}^d , удовлетворяющий следующим условиям:

$$p_k \geq 0 \ (k = \overline{1, d}), \quad \sum_{k=1}^d p_k = 1.$$

Иначе говоря, P является элементом единичного симплекса в пространстве \mathbb{R}^d . Случайную величину X также можно рассматривать как вектор (x_1, \dots, x_d) из \mathbb{R}^d , но без дополнительных ограничений на координаты. Таким образом, перед нами — заданные двумя векторами дискретное распределение и дискретная случайная величина.

Произведем следующее обобщение. Рассмотрим вектор, составленный из корней вероятностей:

$$\psi \rightleftharpoons (\sqrt{p_1}, \dots, \sqrt{p_d}) \rightleftharpoons (\psi_1, \dots, \psi_d).$$

*PSI —
Probability
Square root
(Imaginary)*

Ясно, что ψ принадлежит единичной сфере S^{d-1} в \mathbb{R}^d , т. к. $\|\psi\| = 1$.

Рассмотрим также матрицу

$$X = \begin{pmatrix} x_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & x_d \end{pmatrix}$$

³⁶Такой трюк используется для определения случайного комбинаторного объекта, например, графа.

Матрица X определяет линейный непрерывный самосопряженный (эрмитов) оператор \hat{X} , для которого вектор (x_1, \dots, x_d) является набором собственных чисел, а единичные векторы стандартного орто-нормированного базиса являются собственными векторами оператора \hat{X} .

Для классических случайных величин определяется линейный функционал среднего значения (или математического ожидания) по правилу:

$$\mathbb{E}_P X = \sum_{k=1}^d x_k p_k$$

Этот же функционал мы получим через матрицу:

$$\mathbb{E}_P X = \psi \cdot X \psi = \psi^T X \psi,$$

где первое выражение записано как скалярное произведение векторов ψ и $X\psi$, а второе — как произведение матриц.

Заметим далее, что, меняя базис пространства \mathbb{R}^d при помощи ортогонального преобразования U , мы не меняем действие линейного оператора \hat{X} , хотя его матрица меняется по правилу $U^{-1}XU$. Скалярные произведения такие преобразования также не меняют (см. раздел 3.5). То есть, если в новом базисе вектор ψ принимает вид $\tilde{\psi} = U^{-1}\psi$, то

$$\tilde{\psi}^T (U^{-1}XU) \tilde{\psi} = \psi^T U U^{-1} X U U^{-1} \psi = \psi^T X \psi,$$

поскольку, как мы знаем, $U^T = U^{-1}$ для ортогональных матриц.

Таким образом, математическое ожидание не зависит от выбора базиса пространства \mathbb{R}^d , т. е. является собственным свойством пары (ψ, \hat{X}) распределения и оператора. Понятно, что и все остальные функционалы от случайной величины X , которые могут быть выражены через скалярные произведения векторов, также не зависят от выбора конкретного базиса.

Такое замечательное свойство линейных пространств позволяет придти к обобщению классической вероятности на случай комплексных величин.

Пусть теперь вектор ψ является элементом \mathbb{C}^d (по-прежнему, с ограничением $\|\psi\| = 1$, но уже без требования положительности компонент!), а \hat{X} — линейный самосопряженный (эрмитов) оператор на \mathbb{C}^d . Если его матрица в некотором базисе равна X , то требование самосопряженности выражается равенством $\overline{X}^T = X$. Часто для обозначения матрицы сопряженного оператора используют символ X^\dagger , чтобы не ставить два символа — транспонирование и комплексное сопряжение.³⁷

³⁷ Другим распространенным обозначением является «звездочка» X^* , причем ее применяют и к матрице, и к вектору, и к числу. В последнем случае она просто обозначает комплексное сопряжение, т. к. число само себе транспонировано (заметьте, что это работает и в том случае, когда комплексное число записано матрицей).

В случае \mathbb{C}^d , как мы знаем, унитарное преобразование, заданное унитарной матрицей U , также сохраняет скалярные произведения и является изометрией. Поэтому функционал

$$\langle \hat{X} \rangle_\psi \Rightarrow \psi \cdot X\psi = \psi^\dagger X\psi$$

не зависит от выбора базиса и является собственной характеристикой пары (ψ, \hat{X}) .

Это значит, что в подходящем базисе (а именно: в базисе из собственных векторов оператора \hat{X}) матрица оператора \hat{X} принимает диагональный вид

$$X = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_d \end{pmatrix},$$

где $\lambda = (\lambda_1, \dots, \lambda_d)$ — спектр оператора \hat{X} , т. е. набор собственных чисел. Кроме того, из эрмитовости оператора следует, что все его собственные числа вещественные, т. е. $\lambda \in \mathbb{R}^d$. Требование эрмитовости оператора связано именно с тем, чтобы его спектр был вещественным, поскольку он должен отражать реально измеряемые физические параметры. Поэтому в случае комплексного обобщения мы имеем почти такую же картину, как в случае описания

Слово «почти» — не то, которое в теории меры!

классической случайной величины. «Почти» — потому что вектор ψ теперь является комплексным. И это, как мы увидим далее, является очень существенным отличием квантовой вероятности от классической.

Дадим теперь сопутствующую терминологию. Вектор ψ принято называть **состоянием** (квантовой системы),³⁸ оператор \hat{X} — **наблюдаемой** (величиной), собственные значения оператора \hat{X} — **значениями** наблюдаемой, собственные векторы оператора \hat{X} — **базисными состояниями** наблюдаемой, функционал $\langle \hat{X} \rangle_\psi$ — **ожидаемым значением** наблюдаемой \hat{X} в состоянии ψ .

В квантовой физике широко используется нотация Дирака для записи векторов:

$$\langle \alpha | \Rightarrow (\bar{\alpha}_1 \dots \bar{\alpha}_d) = \alpha^\dagger, \quad |\alpha\rangle \Rightarrow \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_d \end{pmatrix},$$

³⁸В общей канве книги предпочтительнее было бы назвать их *предвероятностями*, *предраспределениями* или *псевдораспределениями*, но термин «состояние» пришел из физики и укоренился, как это часто бывает.

где символ $\langle \alpha |$ называется **бра-символом**, а символ $|\alpha\rangle$ — **кет-символом**. Нетрудно видеть, что бра-символ — это вектор-строка, или сопряженный вектор, а кет-символ — это обычный вектор-столбец. Для удобства записи также принято сокращать различные обозначения умножения и двойные вертикальные черточки. Так, скалярное произведение $\alpha \cdot \beta$ в символах Дирака будет записано как $\langle \alpha | \beta \rangle$. Обычная матричная запись, соответствующая этому символу, выглядит так: $\alpha^\dagger \beta$.

совместно
именуемые
брэкет-
символами
от англ.
bracket.

Следует обратить внимание на то, что запись вектора как упорядоченного набора с помощью запятых (*comma*-нотация) — тоже строковая, однако это аналог записи вектора-столбца, в то время как вектор-строка (транспонированный столбец) записывается без запятых.

Как и в обычной теории вероятностей, вместо буквы внутри брэкет-символа может стоять высказывание, определяющее вектор. Например, мы записываем $P\{\xi \text{ равно нормали к плоскости } \kappa\}$ и $\langle \text{нормаль к плоскости } \kappa |$. Часто это позволяет сделать текст более читабельным. Брэкет-символы удобны также и тем, что их часто можно рассматривать как независимо перемножаемые матрицы (из-за свойств линейности операторов и скалярного произведения), соблюдая при этом правила действий с матрицами.

Так,³⁹

$$\langle \alpha | \beta \rangle \cdot \langle \zeta | \eta \rangle = \langle \alpha | \cdot | \beta \rangle \langle \zeta | \cdot | \eta \rangle.$$

В соответствии с вышеизложенным *ожидаемое значение наблюдаемой* \hat{X} в состоянии ψ задается по формуле

$$\langle \hat{X} \rangle_\psi \Rightarrow \psi \cdot X \psi = \langle \psi | X | \psi \rangle.$$

Здесь, как и в классике, принято пропускать символ состояния (распределения), если это не приводит к неточностям. Так что, вместо $\langle \hat{X} \rangle_\psi$ пишут $\langle \hat{X} \rangle$.

Вспомним теперь нашу подбрасываемую монетку и предположим, что измерения результатов ее подбрасывания принимают значения либо 1, либо -1. Соответствующий оператор имеет матрицу

$$X = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Он имеет собственные числа 1 и -1, и соответствующие им собственные векторы $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ и $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

³⁹По правилам перемножения матриц выражение $|\beta\rangle\langle\zeta|$ представляет собой матрицу всех возможных попарных произведений компонентов векторов β и ζ . Данная матрица является матрицей оператора проекции на направление β .

В свободном полете монетка имеет состояние $\psi = (\sqrt{p}, \sqrt{q})$, пока мы не произведем измерение, т. е. пока она не упадет на стол. И поэтому ее ожидаемое значение будет равно $\langle \psi | X | \psi \rangle = p - q$. Но, упав на стол, она придет в одно из своих базисных состояний — либо в $\psi_0 = (1, 0)$, либо в $\psi_1 = (0, 1)$. В первом из них монетка стабилизируется в значении $\langle \psi_0 | X | \psi_0 \rangle = 1$, во втором — в значении $\langle \psi_1 | X | \psi_1 \rangle = -1$.

Похожим образом ситуация развивается в случае с электроном, который до всякого измерения может иметь некоторое состояние $\psi = (z, w)$, где $z, w \in \mathbb{C}$ и $|z|^2 + |w|^2 = 1$. Электрон можно представить себе в виде волчка, т. е. монетки с центральной осью, и эта ось может смотреть либо вверх, либо вниз, т. е. иметь два состояния 1 и -1 вдоль оси Oz нашего пространства. Направление оси вращения в данном случае называется *спином* электрона.⁴⁰

Оператором, описывающим наблюдаемую спина относительно оси Oz принято считать

$$\sigma_z \equiv \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

с теми же собственными числами и собственными состояниями, что у «монетного» оператора (коэффициент 1/2 выбирается из физических соображений и для нас не играет роли).

При этом произвольное состояние $\psi = (z, w)$ можно представить как комбинацию базисных состояний:

$$\begin{pmatrix} z \\ w \end{pmatrix} = z \begin{pmatrix} 1 \\ 0 \end{pmatrix} + w \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

или, в нотации Дирака,

$$|\psi\rangle = z|\psi_0\rangle + w|\psi_1\rangle.$$

Такое представление произвольного состояния называется **суперпозицией**.

Заметим, что «весовые» комплексные коэффициенты z, w здесь можно вычислить как скалярные произведения

$$z = \langle \psi_0 | \psi \rangle, \quad w = \langle \psi_1 | \psi \rangle$$

и считать некой характеристикой перехода из состояния ψ в соответствующее базисное состояние.

Измерение спина электрона производится в известном опыте Штерна–Герлаха (рис. 5.6).

⁴⁰На самом деле все немного сложнее, и речь идет конечно же о собственном магнитном моменте электрона. Спин не связан с геометрией «вращения» электрона, которой попросту нет, но само понятие возникло именно из первоначальных (неверных) представлений об электроны, как о вращающейся частице.

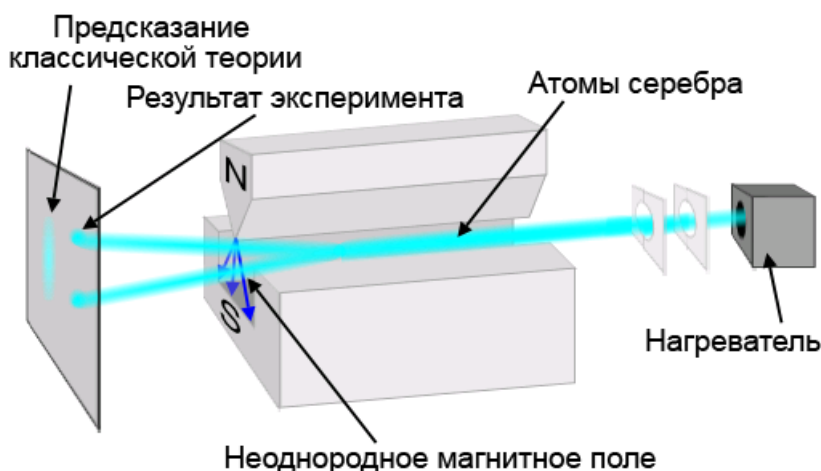


Рис. 5.6: Опыт Штерна–Герлаха.

Права на изображение принадлежат [Theresa Knott](#), CC BY-SA 3.0

Электрон (или атом серебра во избежание воздействия электрического заряда, как в оригинале этого опыта), пролетая через магнитную щель с неоднородным магнитным полем, отклоняется этим полем вверх или вниз на некоторый угол. В том случае, если бы это была намагниченная *макροчастица*, она бы отклонялась на произвольный угол при каждом новом испытании (угол зависит от угла наклона магнитного поля самой частицы) и засвечивала бы сплошное пятно на экране, образуя что-то вроде гауссовой плотности распределения. *Однако с квантами это не так.* Электрон отклоняется строго либо на определенный угол вверх, либо на симметричный ему вниз (независимо от ориентации его спина до прохождения через прибор), т. е. ведет себя как бернуллиевская монетка, игнорируя топологию евклидова пространства.

Такое поведение электрона описывается не классической, а квантовой вероятностью, когда спин электрона считается наблюдаемой величиной, записанной в виде эрмитова оператора.

Если гипотетическое ожидаемое значение спина электрона до измерения (т. е. до прохождения через прибор и экран) можно вычислить как $\langle \psi | \sigma_z | \psi \rangle = (|z|^2 - |w|^2)/2$, то в состоянии ψ_0 это будет $\langle \psi_0 | \sigma_z | \psi_0 \rangle = 1/2$, а в состоянии ψ_1 — $\langle \psi_1 | \sigma_z | \psi_1 \rangle = -1/2$.

Более интересной ситуация становится, когда мы ставим последовательно два прибора Штерна–Герлаха и смотрим, что происходит после прохождения электроном их обоих. В этой ситуации первый прибор *подготавливает* состояние, так что нам становится известно распределение (z, w) , а второй его

модифицирует, прежде чем экран произведет *измерение*. Таким образом, в первом приборе мы получаем одно базисное состояние, во втором — другое. Переход из одного базисного состояния в другое носит дискретный вероятностный характер, а переходные вероятности (в классическом их смысле) считаются через скалярные произведения состояний. Скалярные произведения состояний называются **амплитудами вероятностей**.

Точнее, пусть электрон после прохождения первого прибора получил состояние ψ_0 (т. е. мы отсекали электроны, получившие состояние ψ_1). Амплитудой вероятности того, что электрон получит состояние ψ_0 после прохождения второго прибора, является скалярное произведение $\langle \psi_0 | \psi_0 \rangle$, а вероятность перехода из состояния ψ_0 в состояние ψ_0 равна квадрату модуля амплитуды, т. е. $|\langle \psi_0 | \psi_0 \rangle|^2$.

В нашем примере амплитуда равна $\langle \psi_0 | \psi_0 \rangle = 1$, как и соответствующая вероятность. т. е., отклонившись в первом приборе вверх, электрон сделает то же самое и во втором с вероятностью 1. Соответственно, $\langle \psi_0 | \psi_1 \rangle = 0$.

До сих пор мы рассматривали в качестве наблюдаемой только положение спина электрона относительно оси Oz пространства. На самом деле, в трехмерном пространстве имеется три взаимно ортогональных оси Ox , Oy , Oz , относительно каждой из которых можно рассматривать направление спина, а также комбинацию относительно всех трех осей.

Для каждой из осей существует своя матрица оператора наблюдаемой величины:

$$\sigma_x \Rightarrow \frac{1}{2} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y \Rightarrow \frac{1}{2} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z \Rightarrow \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Данные матрицы называются **матрицами Паули** и образуют полный базис в пространстве всех линейных эрмитовых операторов с нулевым следом над \mathbb{C}^2 . Если к ним добавить еще единичную матрицу, то полученный набор из 4-х матриц является полным базисом в пространстве всех линейных эрмитовых операторов над \mathbb{C}^2 .

Чтобы объяснить выбор именно таких матриц в качестве базисных, можно вспомнить матричное представление кватернионов. Кватернион $z + wj$ представляется матрицей $\begin{pmatrix} z & w \\ -\bar{w} & \bar{z} \end{pmatrix}$, а базисные кватернионы $1, i, j, k$ имеют представление

$$1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad i = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad j = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad k = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}.$$

Нетрудно видеть, что матрицы $\sigma_x, \sigma_y, \sigma_z$ можно получить из матриц для кватернионов k, j, i формальным делением их компонент на $2i$.⁴¹

⁴¹ Деление формальное, т. к. матрицы Паули не являются матричным представлением кватернионов, их нельзя разделить на $2i$ как кватернион на кватернион. В то же время, матрицы кватернионов также не являются эрмитовыми матрицами.

Так что матрицы Паули находятся в точном алгебраическом соответствии с базисом i, j, k в мнимом подпространстве кватернионов, которое, как мы знаем, хорошо моделирует \mathbb{R}^3 при построении группы вращений (см. раздел 3.6.2).

Отсюда следует, что полное трехмерное представление спина электрона в трехмерном пространстве является линейной комбинацией матриц Паули:

$$\sigma = a\sigma_x + b\sigma_y + c\sigma_z.$$

Теперь мы можем вычислить, например, вероятность того, что электрон, стартовав с первого прибора в состоянии i перейдет после второго прибора в состояние j (предполагается, что первый прибор ориентирован осью N-S в направлении оси i , а второй — в направлении оси j). Состояние i — это собственный вектор матрицы σ_z , соответствующий собственному числу 1 и равный $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, а состояние j — это собственный вектор матрицы σ_y , соответствующий собственному числу 1 и равный $\begin{pmatrix} 1 \\ i \end{pmatrix}/\sqrt{2}$. Тогда амплитуда вероятности перехода из i в j равна

$$\langle j|i \rangle = 1/\sqrt{2},$$

а сама вероятность перехода равна $1/2$. То есть ровно половина электронов, прошедших первый вертикально ориентированный прибор в первом базисном состоянии по оси Oz , пройдут второй, горизонтально ориентированный прибор, в первом базисном состоянии по оси Oy .

Заметим, что если бы мы производили аналогичный опыт с классическими намагниченными частицами, то даже после первого прибора Штерна–Герлаха, распределение частиц на экране было бы похоже на плотность гауссовского распределения, но никак не двухточечного.

Для того, чтобы лучше понять правила оперирования с амплитудами, рассмотрим пример с двумя щелями (опыт Юнга, рис. 5.7).

Если бы частицы, испускаемые источником, вели себя как макрочастицы, то картина была бы следующей. Обозначим через η случайную величину, которая равна $+a/2$, если частица проходит через щель S_1 , и $-a/2$, если частица проходит через щель S_2 . Далее, обозначим через ξ случайную координату x на экране, в которую попадет частица в отсутствие щелей, т. е. в чистом случае. При этом предполагается, что точка $x = 0$ находится непосредственно напротив источника частиц. При наличии щелей мы предполагаем, что частица, пройдя сквозь щель, ведет себя также, как если бы она была испущена источником, находящимся в этой щели. Наконец, обозначим через ζ координату попадания частицы в экран при условии прохождения одной из щелей. Тогда $\zeta = \xi + \eta$.

Предполагая, что координата x дискретна (принимает значения x_0, x_1, x_{-1} и т.д.), мы можем записать вероятность попадания частицы в точку x при

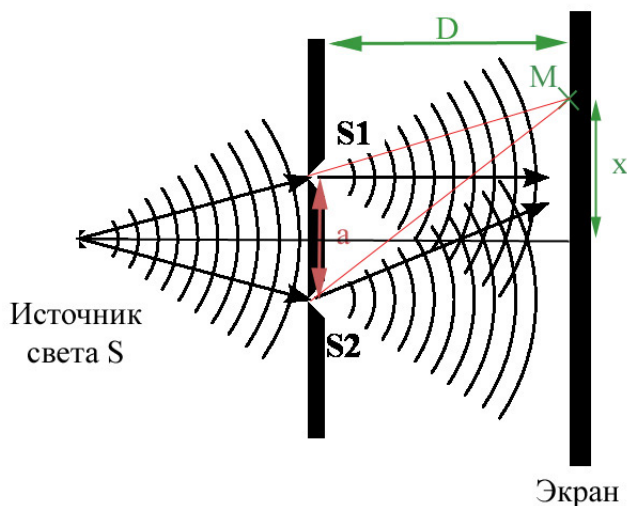


Рис. 5.7: Опыт Юнга.

Права на изображение принадлежат Савенок Д., CC BY-SA 3.0

условии прохождения ею одной из щелей. По формуле Байеса имеем:

$$\begin{aligned}
 P\{\zeta = x\} &= \\
 &P\{\zeta = x - a/2 | \eta = a/2\} P\{\eta = a/2\} + \\
 &+ P\{\zeta = x + a/2 | \eta = -a/2\} P\{\eta = -a/2\} = \\
 &= P\{\xi = x - a/2\} P\{\eta = a/2\} + P\{\xi = x + a/2\} P\{\eta = -a/2\}
 \end{aligned}$$

Здесь мы в очень упрощенном виде наблюдаем фейнмановский принцип суммирования по независимым траекториям: вдоль траектории вероятности перемножаются, а вероятности параллельных траекторий с общими началом и концом складываются.

Байесовская формула может быть обобщена на случай непрерывного распределения ξ , что даст формулу плотности вероятности. Как видим, фактически она дает сумму смещенных одинаковых плотностей с весовыми коэффициентами $P\{\eta = a/2\}$ и $P\{\eta = -a/2\}$, которые в данном опыте можно считать равными $1/2$ (из-за симметрии отверстий). Итоговая плотность будет иметь вид

$$r(x) = \frac{D/2\pi}{D^2 + (x - a/2)^2} + \frac{D/2\pi}{D^2 + (x + a/2)^2},$$

и при разном соотношении a/D график может иметь вид двухвершинной плотности или же одновершинной (см. рис 5.8)

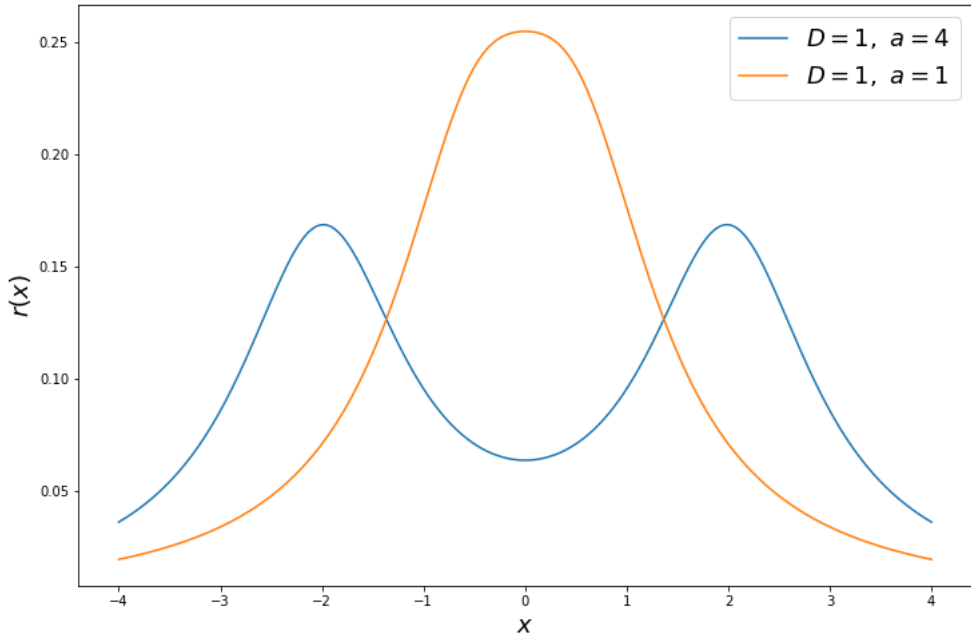


Рис. 5.8: Опыт Юнга. Классическое распределение.

Однако в квантовом случае мы наблюдаем интерференционную картину. Для того, чтобы ее вычислить, нужно вместо вероятностей в формуле Байеса подставлять амплитуды вероятностей. Это общее правило: *амплитуды вероятностей альтернативных возможностей складываются*, т.е. в отличие от классики, мы оперируем амплитудами до самого конца вычислений, и только в конце для перехода к вероятности находим квадрат модуля результирующей амплитуды (это называется **правилом Борна**). При этом, как и раньше, прохождение через щель подготавливает частицу (фотон), т.е. задает ее базовое состояние $+a/2$ или $-a/2$, так что далее нам нужно складывать амплитуды перехода из базовых состояний в состояние x на экране: $\langle x|+a/2\rangle$ и $\langle x|-a/2\rangle$ с теми же весовыми коэффициентами, отвечающими за вероятность попадания в точки $+a/2$ и $-a/2$, которые также можно положить равными.

Таким образом, плотность вероятности регистрации фотона в точке x экрана равна

$$\left| \langle x|-a/2\rangle \frac{1}{\sqrt{2}} + \langle x|+a/2\rangle \frac{1}{\sqrt{2}} \right|^2.$$

Обычно в данном опыте амплитуды выбираются в виде $\exp(ikr)/r$ с соот-

ветствующей нормировкой, где r означает расстояние от точки x на экране до соответствующей щели. Поэтому плотность распределения мы получаем в виде

$$r(x) = \frac{D}{2\pi} \left| \frac{\exp(ik\sqrt{D^2 + (x - a/2)^2})}{\sqrt{D^2 + (x - a/2)^2}} + \frac{\exp(ik\sqrt{D^2 + (x + a/2)^2})}{\sqrt{D^2 + (x + a/2)^2}} \right|^2.$$

График полученной плотности распределения представлен на рис. 5.9.

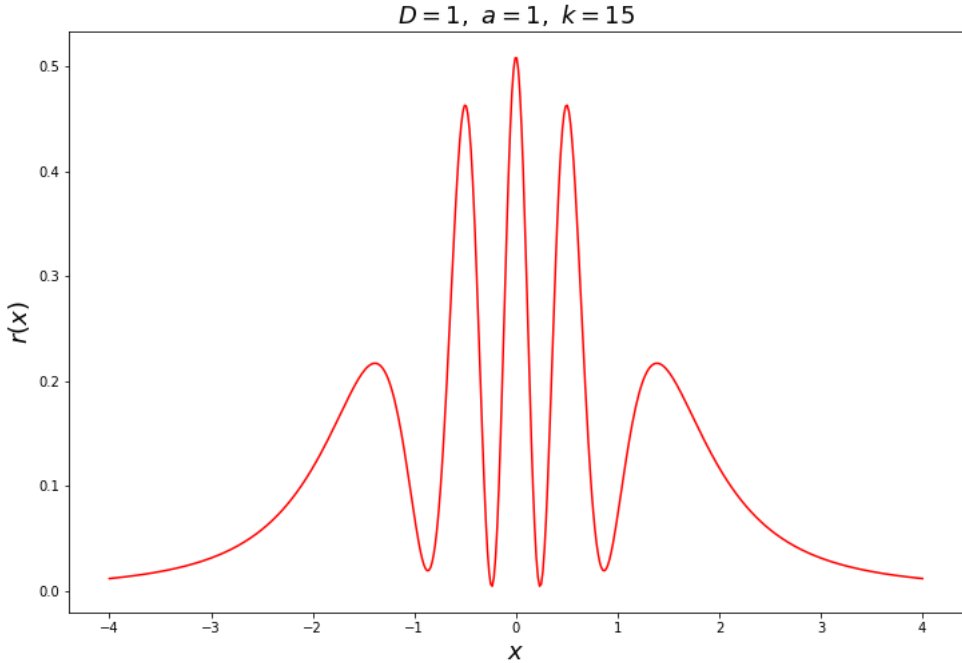


Рис. 5.9: Опыт Юнга. Квантовое распределение. $D = 1 = a, k = 15$.

В ходе рассмотрения примеров мы незаметно перешли от дискретных наблюдаемых к непрерывным, полагая число точек на экране бесконечно большим, а расстояние между соседними — бесконечно малым. При этом вместо вероятностей возникла плотность распределения. Чаще всего такой переход оправдан в силу «хороших» свойств изучаемых случайных величин и наблюдаемых.

В общем случае ситуация выглядит следующим образом. Каждая квантовая система соответствует некоторому гильбертову пространству \mathcal{H} , которое отвечает за состояния этой системы. Точнее, за них отвечают векторы $\psi \in \mathcal{H}$, имеющие норму $\|\psi\| = 1$. Наблюдаемые — это самосопряженные (эрмитовы) линейные операторы на \mathcal{H} . Поскольку в гильбертовом пространстве задано скалярное произведение, мы также можем определить амплитуды вероятно-

стей вида $\langle \alpha | \beta \rangle$ и, например, ожидаемое значение $\langle A \rangle_\psi = \langle \psi | A | \psi \rangle$.

Размерность \mathcal{H} при этом может быть какой угодно — все зависит от конкретной физической модели. Так, **кубиту** соответствует пространство \mathbb{C}^2 , это *двухуровневая* квантовая система, описывающая спин электрона, направление фотона и т.д. Базовые состояния кубита описываются матрицами Паули 2×2 .

Также встречаются многоуровневые системы с дискретным спектром, системы с непрерывными спектрами операторов, а также системы, эволюционирующие во времени.

Об эволюции во времени

До сих пор мы ничего не говорили об изменении наблюдаемых и состояний во времени. Между тем, как мы знаем, в классической теории вероятностей и приложениях очень важную роль играют случайные процессы, развивающиеся во времени.

Например, винеровские процессы.

Поэтому предположим теперь, что состояние ψ есть функция от времени $t \in \mathbb{R}$. Кроме того, значения $\psi(t)$ являются элементами гильбертова пространства \mathcal{H} , не обязательно конечномерного, над полем \mathbb{C} . Например, это могут быть суммируемые с квадратом модуля на отрезке функции вида $\mathbb{R}^n \rightarrow \mathbb{C}$.

Эволюцией состояния называется зависимость вида

$$\psi(t) = \hat{U}(t)\psi(0),$$

где $\hat{U}(t)$ есть оператор над \mathcal{H} , удовлетворяющий условиям:

EVOL1 $\hat{U}(t)$ сохраняет скалярное произведение, заданное в \mathcal{H} ;

EVOL2 $\hat{U}(t)$ есть линейный оператор;

EVOL3 $\hat{U}(t)$ дифференцируема в точке $t = 0$;

EVOL4 $\hat{U}(t)\hat{U}(s) = \hat{U}(t+s)$, т. е. \hat{U} осуществляет групповой гомоморфизм из $(\mathbb{R}, +)$ в группу операторов с операцией композиции.

Первые два условия, как мы знаем из раздела 3.5, означают, что $\hat{U}(t)$ есть унитарный оператор, т. е. $\hat{U}^\dagger(t)\hat{U}(t) = \text{id}$.

Из третьего условия следует представление

$$\hat{U}(dt) = \text{id} - (i/\hbar)\hat{H}dt + o(dt)$$

при бесконечно малых dt , где коэффициент перед \hat{H} был вынесен с определенной целью, а именно:

$$\begin{aligned} 0 &= \hat{U}^\dagger(t)\hat{U}(t) - \text{id} = (\text{id} + (i/\hbar)\hat{H}^\dagger dt + o(dt))(\text{id} - (i/\hbar)\hat{H}dt + o(dt)) = \\ &= (i/\hbar)(\hat{H}^\dagger - \hat{H})dt + o(dt), \end{aligned}$$

откуда следует, что $\hat{H}^\dagger - \hat{H} = 0$, т. е. оператор \hat{H} является эрмитовым. Оператор \hat{H} называется **гамильтонианом**.

Из четвертого условия следует всюду дифференцируемость

$$i\hbar\hat{U}'(t) = \hat{H}U(t).$$

(5.19)

Упражнение
5.33.
Докажите
(5.19)!

Наконец, отсюда мы получаем знаменитое **уравнение Шрёдингера**

Теорема 5.8 (Шрёдингера). Для эволюции состояния $\psi(t)$ справедливо уравнение:

$$i\hbar\frac{d}{dt}|\psi(t)\rangle = \hat{H}|\psi(t)\rangle, \quad (5.20)$$

где \hat{H} — эрмитов оператор.

Мы получили уравнение Шрёдингера с независимым от времени гамильтонианом (поскольку он получен как производная ψ в нуле), однако в общем случае предполагается, что в уравнении (5.20) гамильтониан может зависеть от времени.

Пусть E_0 и ψ_0 — некоторые соответствующие друг другу собственное значение и собственный вектор гамильтониана, т. е. $\hat{H}\psi_0 = E_0\psi_0$, $E_0 \in \mathbb{R}$. Заметим, что из разложения в любом из собственных состояний

$$i\hbar d\psi(t)|_{\psi=\psi_0} = \hat{H}\psi_0 dt = E_0\psi_0 dt$$

и того, что вектор ψ не имеет единиц измерения, следует, что число E_0 должно измеряться в единицах, обратных времени (т. к. умножается на dt), т. е. быть пропорциональным частоте квантовой частицы с состоянием $\psi(t)$. Известно, что $\hbar\nu$ — энергия частицы с частотой ν . Поэтому принято считать, что собственные числа E_0 оператора \hat{H} — это энергетические уровни квантовой системы, описываемой функцией $\psi(t)$, а сам оператор \hat{H} является оператором энергии. Иначе говоря гамильтониан — это наблюдаемая величина энергии квантовой системы.

Функцию состояния $\psi(t)$, удовлетворяющую уравнению Шрёдингера, принято называть **волновой функцией**.

Помимо эволюции состояния можно рассматривать **эволюцию наблюдаемой**:

$$\hat{X}(t) = \hat{U}(t)^\dagger \hat{X}(0) \hat{U}(t),$$

где $\hat{U}(t)$ — тот же самый унитарный дифференцируемый оператор.

Дифференцируя это равенство по t и пользуясь соотношением (5.19), нетрудно получить **уравнение Гейзенберга**:

$$i\hbar\frac{d}{dt}\hat{X}(t) = \hat{H}(t)\hat{X}(t) - \hat{X}(t)\hat{H}(t) \equiv [\hat{H}(t), \hat{X}(t)],$$

где $\hat{H}(t) = \hat{U}(t)^\dagger \hat{H} \hat{U}(t)$ — эволюция гамильтониана, а $[\hat{H}(t), \hat{X}(t)]$ называется **коммутатором** $\hat{H}(t)$ и $\hat{X}(t)$.

Полученное уравнение интересно тем, что дает критерий постоянства наблюдаемой. $\hat{X}(t)$ сохраняется во времени тогда и только тогда, когда она коммутирует с наблюдаемой энергией. В частности, поскольку гамильтониан сам с собой коммутирует, энергия остается постоянной во времени.

Самое время вспомнить о том «небольшом» отличии квантовых вероятностей от классических, о котором мы упоминали ранее. Дело в том, что в комплексном случае операторы над гильбертовым пространством не обязаны коммутировать, т. е. $\hat{X}\hat{Y}$ не всегда совпадает с $\hat{Y}\hat{X}$. Их $[\hat{X}, \hat{Y}] = \hat{X}\hat{Y} - \hat{Y}\hat{X}$ не обязан быть нулевым оператором.

Известная следующая

Теорема 5.9 (Гейзенберга). Пусть заданы эрмитовы операторы \hat{X}, \hat{Y} на пространстве \mathcal{H} и состояние $\psi \in \mathcal{H}$. Кроме того, определены средние квадратические отклонения операторов:

$$\Delta\hat{X} = \sqrt{\langle\psi|(\hat{X} - \langle\hat{X}\rangle)^2|\psi\rangle}, \quad \Delta\hat{Y} = \sqrt{\langle\psi|(\hat{Y} - \langle\hat{Y}\rangle)^2|\psi\rangle}.$$

Тогда

$$\Delta\hat{X}\Delta\hat{Y} \geq \frac{1}{2} \left| \langle\psi|[\hat{X}, \hat{Y}]|\psi\rangle \right|$$

Иначе говоря, для некоммутирующих операторов невозможно одновременно (в одном состоянии) получить точные собственные значения: если одна наблюдаемая коллапсирует в одно из своих собственных значений в состоянии ψ , то вторая наблюдаемая не коллапсирует, и наоборот.

Предположим, что гильбертово пространство \mathcal{H} устроено некоторым классическим способом, т. е. его элементы суть функции от вещественного аргумента x (дискретного или непрерывного числа или вектора). Таким образом, состояние ψ будет функцией от этого аргумента x и, возможно, от времени t .

Определим оператор пространственной координаты:

$$\hat{Q}[\psi(x, t)] \Rightarrow x\psi(x, t)$$

и оператор импульса:

$$\hat{P}[\psi(x, t)] \Rightarrow -i\hbar\nabla\psi(x, t).$$

Легко видеть, что

$$\hat{P}\hat{Q}[\psi] - \hat{Q}\hat{P}[\psi] = -i\hbar\nabla(\hat{Q}[\psi]) - x\hat{P}[\psi] = -i\hbar\nabla(x\psi) + xi\hbar\nabla\psi = -i\hbar[\psi],$$

т. е. коммутатор $[\hat{P}, \hat{Q}] = -i\hbar$. Отсюда по теореме Гейзенберга получаем известное соотношение неопределенности Гейзенберга:

$$\Delta\hat{Q}\Delta\hat{P} \geq \frac{\hbar}{2}.$$

Что означает принципиальную невозможность одновременного точного измерения координаты и импульса квантовой частицы.

Отметим, что нечто похожее мы имеем в классической математической статистике. Имеется ввиду **неравенство Рао–Крамера**

$$D\hat{\theta}(x) \geq \frac{1}{I_n(\theta)},$$

означающее, что дисперсию несмещенной оценки параметра θ невозможно занулить (коллапсировать оценку в точку), если информация Фишера $I_n(\theta)$ конечна. Примечательно, что существует еще и *квантовое неравенство Рао–Крамера*, являющееся обобщением обычного неравенства Рао–Крамера, но не имеющее прямого отношения к неравенству Гейзенберга. Подробнее см. в [82, 96].

Оператор энергии можно выразить через операторы кинетической и потенциальной энергии:

$$\hat{H}[\psi(x, t)] = \frac{\hat{P}^2}{2m} + \hat{V}(x, t),$$

где первое слагаемое отвечает за кинетическую энергию, а второе — за потенциальную, m — масса частицы. При этом \hat{P}^2 следует считать квадратом оператора \hat{P} , т. е. двукратным его применением, а функцию состояния $\psi(x, t)$ — дифференцируемой второго порядка, т. е. уже достаточно гладкой.

Подставляя такое представление гамильтониана, а также определение оператора импульса в полученное выше уравнение Шрёдингера, получаем классическое уравнение Шрёдингера:

$$i\hbar \frac{d}{dt} |\psi(x, t)\rangle = \left(-\frac{\hbar^2}{2m} \nabla^2 + V(x, t) \right) |\psi(x, t)\rangle. \quad (5.21)$$

В том случае, когда потенциальная энергия $V(x, t)$ не зависит от времени, можно найти частное решение уравнения Шрёдингера в виде

$$|\psi(x, t)\rangle = e^{-itE/\hbar} |\psi(x)\rangle, \quad (5.22)$$

где E — какое-либо собственное значение (комплексное число) оператора \hat{H} , $\psi(x)$ — собственный вектор, соответствующий собственному значению E , т. е. $\hat{H}\psi(x) = E\psi(x)$. Раскрывая \hat{H} , перепишем уравнение $\hat{H}\psi(x) = E\psi(x)$ в виде

$$-\frac{\hbar^2}{2m} \nabla^2 \psi(x) + V(x)\psi(x) = E\psi(x).$$

Заметим, что данное уравнение не содержит переменной времени, поэтому оно называется *стационарным уравнением Шрёдингера*.

Выражение (5.22) является частным решением уравнения (5.21), общее решение представляется как линейная комбинация всех частных решений, соответствующих всем собственным значениям \hat{H} .

Обычно в таких задачах бывает счетный набор собственных значений энергии, которые нумеруются числами $1, 2, \dots$, но иногда встречается непрерывный спектр оператора \hat{H} и даже смешанный случай с дискретным и непрерывным спектром одновременно.

Например, пусть частица колеблется на отрезке $[0; \pi]$ так, что длина волны колебаний волновой функции кратна π . В этом случае получаем простое уравнение

$$-\psi_n''(x) = E_n \psi_n(x),$$

где $x \in [0; \pi]$, а $\psi(x)$ отвечает за состояния колеблющейся в этом интервале частицы. Решением данного уравнения будет $\sin(nx)$ с соответствующей нормировкой, число n определяет уровень энергии или, что то же самое, частоту колебаний. Данное уравнение описывает также колебания струны с закрепленными концами с частотой n .

Интересно, что в многомерном случае мы имеем уравнение

$$-\nabla^2 \psi_n(x) = E_n \psi_n(x),$$

где x лежит в кубе $[0; \pi]^d$, и в его углах значение $\psi(x) = 0$. Решением такого уравнения будет $\sin(n \cdot x)$ с соответствующей нормировкой, причем $n = (n_1, \dots, n_d)$ и

$$n_1^2 + \dots + n_d^2 = n^2$$

при подходящих натуральных значениях n .

Здесь мы видим необходимость решать диофантовые уравнения для представления квадратов натуральных чисел в виде суммы квадратов (обобщение пифагоровых троек!), чтобы вычислить уровни энергии и описать состояние частицы.

Для более полного изучения квантовой механики мы рекомендуем видеокурс «[Элементарное введение в квантовую механику](#)» [xLfFWXUNJ_I] от пользователя [LightCone](#) и классические книги [83, 94].⁴²

5.4.3 Общий алгебраический подход

Рассмотрим теперь $*$ -алгебру \mathcal{A} (не обязательно коммутативную) над полем \mathbb{C} . Символ $*$ означает, что над этой алгеброй задана **операция сопряжения**, которая подчиняется следующим свойствам:⁴²

$$(\alpha x + \beta y)^* = \bar{\alpha} x^* + \bar{\beta} y^*, \quad (xy)^* = y^* x^*, \quad 1^* = 1,$$

⁴²Здесь мы возвращаемся к традиции обозначать сопряжение звездочкой вместо креста, а комплексное сопряжение — чертой.

кроме того, операция $*$ является **инволюцией**, т. е. $x^{**} = x$.

Часто рассматриваются банаховы $*$ -алгебры,⁴³ у которых операция сопряжения связана с нормой по правилу:

$$\|xx^*\| = \|x\|\|x^*\|,$$

в этом случае такую банахову алгебру называют C^* -алгеброй. Если в алгебре имеется единица, то алгебру называют **унитальной**.

Например, поле \mathbb{C} с операцией комплексного сопряжения является C^* -алгеброй над самим собой. Пространство квадратных комплексных матриц с операцией $A^* = \overline{A}^T$ (транспонирование + комплексное сопряжение элементов) также является C^* -алгеброй над \mathbb{C} (норма матрицы задается как норма соответствующего линейного оператора). Наконец, пространство линейных ограниченных операторов над гильбертовым комплексным пространством с операцией сопряжения⁴⁴ также является C^* -алгеброй.

Характером алгебры (унитальной) \mathcal{A} над полем \mathbb{C} называется отображение $\chi : \mathcal{A} \rightarrow \mathbb{C}$, которое:

Char1 линейно;

Char2 сохраняет умножение;

Char3 $\chi(e) = 1$.

Поскольку характер — это, в частности, гомоморфизм алгебр \mathcal{A} и \mathbb{C} , уместно рассмотреть его ядро $\ker \chi$. Оказывается, что ядро характера является максимальным идеалом алгебры \mathcal{A} , а факторалгебра $\mathcal{A}/\ker \chi$ является полем.

Если исходная алгебра \mathcal{A} является унитарной коммутативной банаховой алгеброй над \mathbb{C} , то фактор $\mathcal{A}/\ker \chi$ является банаховой алгеброй над \mathbb{C} , которая в силу теоремы Гельфанда–Мазура изоморфна \mathbb{C} . Поэтому каждому максимальному идеалу I можно поставить в соответствие единственный характер χ такой, что $\ker \chi = I$. Этот характер определяется как композиция факторотображения и изоморфизма \mathcal{A}/I в \mathbb{C} . Таким образом, между множеством характеров и множеством максимальных идеалов существует естественная биекция.

Множество всех характеров называется **пространством максимальных идеалов** или **спектром алгебры** \mathcal{A} и обозначается $\text{Spec } \mathcal{A}$. Для банаховой унитарной коммутативной алгебры \mathcal{A} спектр является компактным хаусдорфовым пространством (в топологии поточечной сходимости характеров).

Не вдаваясь сильно в детали, приведем две основные теоремы о $*$ -алгебрах.

⁴³Напомним, что в банаховой алгебре задана норма, относительно которой алгебра полна, а норма подчиняется условию субмультипликативности.

⁴⁴В общем случае оператор A^* называется сопряженным к оператору A , если $A(x) \cdot y = x \cdot A^*(y)$ для всех x, y , где точка означает скалярное произведение.

Теорема 5.10 (первая теорема Гельфанда–Наймарка). Пусть \mathcal{A} — коммутативная C^* -алгебра с единицей. Тогда существует изометрический $*$ -изоморфизм (преобразование Гельфанда) из алгебры \mathcal{A} в пространство $C(\text{Spec } \mathcal{A})$ всех непрерывных комплекснозначных функционалов на спектре алгебры \mathcal{A} .

Преобразование Гельфанда каждому элементу $a \in \mathcal{A}$ ставит в соответствие непрерывную функцию $F_a(\chi)$ и задается по правилу:

$$F_a(\chi) = \chi(a), \quad \chi \in \text{Spec } \mathcal{A}$$

Нужно подчеркнуть, что здесь элемент алгебры индексирует функции (не обязательно все), заданные на характерах.

Рассмотрим, например, в качестве алгебры \mathcal{A} поле \mathbb{C} над самим собой. У такой алгебры существует единственный характер $\chi = \text{id}$, соответствующий максимальному идеалу $I = \{0\}$, т.е. $\text{Spec } \mathbb{C} = \{\text{id}\}$ и $C(\text{Spec } \mathbb{C}) = \{ \{(\text{id}, z)\} \mid z \in \mathbb{C} \}$. Соответственно, отображение Гельфанда имеет вид $F_z(\chi) = z$, т.е. тоже совпадает с id .

Рассмотрим пространство \mathbb{C}^n как алгебру функций, определенных на множестве $\{1, \dots, n\}$, с поточечным сложением, умножением и сопряжением. Очевидно, что это коммутативная C^* -алгебра, полная относительно стандартной нормы и обладающая свойством субмультипликативности этой нормы. Обозначим эту алгебру $\mathcal{A}\mathbb{C}^n$.

Спектр $\text{Spec } \mathcal{A}\mathbb{C}^n$ состоит из n характеров, являющихся проекциями на координатные оси:

$$\chi_k(x_1, \dots, x_n) = x_k.$$

Упражнение 5.34. Проверьте, что χ_k является характером.

Таким образом, $\text{Spec } \mathcal{A}\mathbb{C}^n$ — конечное множество мощности n , а пространство $C(\text{Spec } \mathcal{A}\mathbb{C}^n)$ является множеством всех функций из этого множества в \mathbb{C} . Очевидно, что это ровно то же самое, что множество всех n -мерных комплексных векторов, т.е. \mathbb{C}^n .

Посмотрим, как в данном случае устроено преобразование Гельфанда. Пусть $x \in \mathcal{A}\mathbb{C}^n$. Тогда

$$F_x(\chi_k) = \chi_k(x) = x_k,$$

т.е. преобразование Гельфанда каждому вектору $x = (x_1, \dots, x_n)$ ставит в соответствие функцию $F_x = \{(\chi_1, x_1), \dots, (\chi_n, x_n)\}$.⁴⁵

На примере алгебры $\mathcal{A}\mathbb{C}^n$ мы наглядно видим работу теоремы Гельфанда–Наймарка.

Итак, имея коммутативную банахову C^* -алгебру \mathcal{A} , мы можем пройти такой путь

$$\mathcal{A} \rightarrow \text{Spec } \mathcal{A} \rightarrow C(\text{Spec } \mathcal{A}) \cong \mathcal{A} \rightarrow L(C(\text{Spec } \mathcal{A}), \mathbb{C}) \cong L(\mathcal{A}, \mathbb{C}),$$

⁴⁵В данном случае очень удобно записывать функцию в формализме ZF.

где L означает еще один этаж надстройки, а именно — пространство линейных непрерывных функционалов на исходной алгебре, которое будет изоморфно пространству линейных непрерывных функционалов на топологическом пространстве $C(\text{Spec } \mathcal{A})$.

Заметим, что в случае унитарности \mathcal{A} пространство $\text{Spec } \mathcal{A}$ компактно (см., например, в [89]), что существенно упрощает работу с $C(\text{Spec } \mathcal{A})$. Если же алгебра \mathcal{A} неунитарна, ее можно пополнить единицей следующим способом. Построим алгебру

$$\mathcal{A}^+ = \{(a, \lambda) \mid a \in \mathcal{A}, \lambda \in \mathbb{C}\},$$

определив на ней покомпонентное сложение и сопряжение, а умножение зададим по правилу:

$$(a, \lambda)(b, \mu) = (ab + \mu a + \lambda b, \lambda\mu).$$

В алгебре \mathcal{A}^+ единицей будет пара $(0, 1)$, а все основные свойства она унаследует от алгебры \mathcal{A} .

Как видим, унитаризация алгебры, в общем-то, процесс конструктивный и довольно простой. Поэтому для удобства можно сразу предполагать, что мы работаем с унитарной банаховой алгеброй, наделенной инволюцией.

Переходя к пространству линейных функционалов $L(\mathcal{A}, \mathbb{C})$, мы на самом деле определяем то, что выше называлось состояниями.

Скажем, что линейный функционал $f : \mathcal{A} \rightarrow \mathbb{C}$ **положителен**, если $f(a^*a) \geq 0$ для любого $a \in \mathcal{A}$. В случае унитарной алгебры любой линейный положительный функционал на ней ограничен и $\|f\| = f(e)$, где e — единица алгебры \mathcal{A} . **Состоянием** называется положительный линейный функционал f на \mathcal{A} со свойством $f(e) = 1$. Здесь уместно вспомнить о классической теории вероятностей, а точнее, об интеграле

$$f(\xi) = \int_{\Omega} \xi dP,$$

который является линейным функционалом на пространстве вещественных случайных величин (со сходящимся первым моментом), если мы используем функциональный подход, и об интеграле

$$f(l) = \int_{-\infty}^{\infty} l(x) dF(x),$$

который является линейным функционалом от функции l , если мы используем дистрибутивный подход к вероятностям (см. стр. 507). Оба интеграла можно рассматривать как состояния, или как распределения вероятностей. Нормировка $f(e) = 1$ отвечает за то, что здесь используется вероятностная мера.

В приведенном выше примере алгебры \mathcal{AC}^n состояниями являются линейные функционалы вида $f_v(x) = x \cdot (v^*v)$, где $v \in \mathbb{C}^n$ и $\|v\| = 1$.

В данном случае удобно функционал f_v отождествлять с вектором v^*v , компоненты которого положительны, а сумма координат дает 1. Мы уже встречались с такими состояниями, когда описывали классическую дискретную вероятность в терминах состояний (см. раздел 5.4.2).

Поскольку, имея только $*$ -алгебру, мы не можем выделить «в натуре» исходное распределение (состояние), мы поступаем проще, говоря, что состоянием является линейный функционал с определенными свойствами (вместо P рассматриваем интеграл по мере P). В примере выше мы смогли легко перейти от функционала к определяющему его вектору v^*v . На самом деле, зная, что алгебра \mathcal{A} изоморфна (в коммутативном случае) пространству $C(\text{Spec}\mathcal{A})$, мы можем сказать, что состоянием является положительный линейный функционал на пространстве непрерывных функций $C(\text{Spec}\mathcal{A})$ со свойством $\|f\| = 1$.

Кроме того, зная, что $\text{Spec}\mathcal{A}$ — топологическое пространство (с поточечной сходимостью функций), мы можем ввести на нем борелевскую сигма-алгебру и взять некоторую меру μ . Тогда интеграл

$$E_\mu f = \int_{\text{Spec}\mathcal{A}} f d\mu, \quad f \in C(\text{Spec}\mathcal{A}),$$

будет представлять собой пример линейного функционала на $C(\text{Spec}\mathcal{A})$ (а, следовательно, и на \mathcal{A}), а условие его положительности и единичной нормы сведется к тому, что мера μ должна быть вероятностной (полная аналогия с дискретным случаем). В этом случае пространство $\text{Spec}\mathcal{A}$ становится вероятностным пространством.

Отметим также, что аргумент f этого линейного функционала есть то, что ранее мы называли *наблюдаемой*, а в классическом случае — *случайной величиной*.⁴⁶

Таким образом, мы видим, что в коммутативном случае C^* -алгебры дают нам классические вероятностные состояния, а математическое ожидание играет роль линейного функционала на наблюдаемых.

Теорема 5.11 (вторая теорема Гельфанда–Наймарка). *Любая C^* -алгебра \mathcal{A} изоморфна замкнутой подалгебре в алгебре $L(\mathcal{H})$ линейных ограниченных операторов, действующих на некотором гильбертовом пространстве \mathcal{H} .*

Как видим, в некоммутативном случае мы также имеем возможность изоморфно отобразить алгебру на некоторое пространство функций, однако здесь нам уже приходится использовать операторы вместо функционалов

⁴⁶Надо отметить, что случайная величина в общем случае определяется как измеримая функция, здесь же мы ведем речь только о непрерывных. Однако в силу известной теоремы Лузина это отличие не так уж существенно — см. [91].

в качестве эквивалента элементов алгебры. Как мы помним, в квантовом случае в качестве наблюдаемых мы рассматривали эрмитовы (т.е. самосопряженные) операторы. В исходной алгебре \mathcal{A} им соответствуют самосопряженные элементы.

Линейный функционал на алгебре соответствует линейному функционалу на операторах и служит для определения понятия состояния в некоммутативном случае. На самом деле, определение здесь ровно то же самое: состоянием называется положительный линейный функционал f на алгебре \mathcal{A} со свойством $f(e) = 1$, что соответствует положительному линейному функционалу на $L(\mathcal{H})$, принимающему значение 1 на единичном операторе.

Так, выбирая вектор единичной длины $\psi \in \mathcal{H}$, мы можем задать функционал

$$\langle X \rangle_\psi \equiv \langle \psi | X | \psi \rangle,$$

который будет линеен на операторах X из $L(\mathcal{H})$, причем $\langle \psi | E | \psi \rangle = 1$. Этот функционал соответствует некоторому состоянию f для исходной алгебры \mathcal{A} . А поскольку он полностью определяется вектором $\psi \in \mathcal{H}$, то термин «состояние» переносится и на сам этот вектор гильбертова пространства.

Мы не будем останавливаться на разборе примеров для второй теоремы Гельфанда–Найма́рка, поскольку выше уже подробно рассмотрели квантовые вероятности. Вместо этого сведем теперь в общую схему классические и квантовые вероятности. Сопоставление понятий приведено в таблице 5.4.

Таблица 5.4: Квантовые и классические вероятности

Термины	Общие	Классические	Квантовые
Пространство	*-алгебра \mathcal{A}	$C(\Omega)$	$L(\mathcal{H})$
Наблюдаемая	a^*a	$\xi : \Omega \rightarrow \mathbb{R}$	л.о. \hat{X} на \mathcal{H}
Состояние	л.ф. $\phi : \mathcal{A} \rightarrow \mathbb{C}$, $\phi(e) = 1$	E_μ , $\mu\Omega = 1$	$\langle \psi \cdot \psi \rangle$, $\ \psi\ = 1$
Ожидаемое	$\phi(a^*a)$	$E_\mu \xi$	$\langle \psi X \psi \rangle$

Итак, мы видим, что:

- *-алгебра может быть интерпретирована как пространство наблюдаемых или случайных величин;
- линейные функционалы на ней, нормированные в единице, — как состояния или распределения вероятностей;
- значение линейного функционала на наблюдаемой — как ожидаемое значение этой наблюдаемой в этом состоянии.

Причем, коммутативная алгебра порождает классические вероятности, в то время как некоммутативная — квантовые.

5.5 Несколько слов об общей картине

Отметим, что текущая глава стоит особняком по сравнению с другими главами, поскольку в ней мы практически не вводили новых архетипов (кроме архетипа **связности–непрерывности, компактности и неподвижной точки**). По сути эта глава есть большая строительная площадка, на которую мы заранее привезли материалы, технику, инструменты, рабочих и проектную документацию. Здесь мы занимались преимущественно строительством тех объектов, которые приближают фундаментальную математику к так называемой прикладной. Причем, мы только в конце главы добрались до действительно прикладных теорий, имеющих прямое отношение к реальной жизни (теория вероятностей) и к физике (квантовая механика).

Эта глава сама по себе является неким глобальным архетипом математики, демонстрирующим, каким витиеватым и сложным может быть путь от чистого разума к прикладным вещам. Вместе с тем, этот путь показывает нам, как все устроено в мире точных наук, почему оно так устроено и где, возможно, стоит поискать новые пути развития.

Без столь глубинного понимания устройства математики в целом невозможно полноценно ею пользоваться, создавать новые аналитические инструменты и теории, которые могут быть полезными физике, информатике и другим наукам, которые обустроивают нашу жизнь в техническом плане.

Мы видим, что математика есть огромное хранилище всевозможных инструментов, опыта и знаний. Оно столь огромно, что нет никакой возможности для постороннего человека зайти сюда и правильно выбрать нужный ему инструмент. Математики, как архивариусы и мастера, упорядочивают, систематизируют и пополняют это хранилище, не всегда основываясь на внешних запросах, часто следуя только внутреннему творческому позыву. В итоге получается колоссальная глубоко продуманная и структурированная система знаний, которую уместно было бы назвать «империей математики».

В главах, посвященных числам и структурам, мы сравнивали математику с городом, в котором есть здания, башни, дороги, деревья, коммуникации. А в главах об исчислениях мы увидели инфраструктуру этого города, поскольку исчисления — это технологии, позволяющие по определенным правилам (алгоритмам) извлекать новые знания из уже построенных чисел и структур и применять их на практике.

Наконец, коснувшись теории вероятностей и квантовой физики, мы увидели жизнь в этом могучем и древнем городе. И хотя данные разделы не имеют прямого отношения к заявленным целям книги, они, тем не менее, демонстрируют, если угодно, глубинный смысл и силу математического знания. Это необходимо знать и понимать каждому математику, сидящему в своем «замке из словной кости», выражаясь словами Клайна [3], а также всем тем, кто хоть в какой-то мере использует математику для получения результатов

в других областях знаний, прежде всего, конечно, в физике и информатике. Целостное представление об огромном складе математических инструментов позволит человеку ищущему и трудолюбивому найти необходимое орудие и вдохновение использовать его на практике.

Пожалуй, самым существенным архетипом, который можно вынести из этой главы, является способность математики инкапсулировать в себя различные не вполне формализованные идеи, гипотезы, теории, находя для них строгие формальные конструкции и их изоморфные образы. Так, пространство наблюдаемых величин пришло к нам из физики и получило несколько интерпретаций на различных уровнях математического знания. Это, во-первых, эрмитовы матрицы, во-вторых, эрмитовы операторы над гильбертовым пространством, наконец, это самосопряженные элементы в алгебре с инволюцией. Примечательно то, что такая инкапсуляция идей в виде различных математических конструкций выдерживает стресс-тест на согласованность. Иначе говоря, хоть мы и вводим понятие различными способами (порой вообще независимыми и принадлежащим к разным областям математики), тем не менее, мы находим изоморфное соответствие между ними, что позволяет отвлечься от технических деталей и сосредоточиться на изучении нового понятия как такового.

Графы

Граф — один из новейших (в контексте всей истории математики) и один из самых мощных архетипов математики. Концепцию графа можно поставить в один ряд с концепцией произвольного множества или алгебраической структуры. В наше время очень много задач могут быть рассмотрены с точки зрения теории графов. Прежде всего, это касается различных структур данных, возникающих в естественно-научных (химия, биология, медицина), финансовых и социальных (экономические и социальные связи) моделях. Немало времени уделяется на изучение графов интернет-типа, описывающих сеть Интернет и аналогичные структуры. С наступлением эры BigData «графические» методы выходят на передний план в задачах искусственного интеллекта и машинного обучения. Кроме того, графы играют значительную роль в теории алгоритмов и исследовании операций.

6.1 Подходы к определению

Определение графа дается, как правило, в контексте рассматриваемой теории. Оно может варьироваться от книги к книге, хотя в целом все эти вариации укладываются в общую концепцию о том, что граф — это множество вершин, некоторые из которых соединены ребрами (они могут быть направленными и ненаправленными, а также петлями). Иначе говоря, граф G это пара (V, E) , где V — множество **вершин**, E — множество **ребер**. Варианты того, как именно E связано с V , мы приводим в следующей таблице 6.1.

Рассмотрим подробнее эти определения. Во-первых, как видим, ребра **неориентированного** графа представляются неупорядоченными парами вершин, в то время как ребра **ориентированного** графа (или **орграфа**) — упорядоченными парами вершин. Упорядоченные пары вершин также называют *дугами*. Вершины, определяющие ребро, называются концами этого ребра, а в ориентированном случае выделяют начало и конец дуги. Если концы ребра совпадают, то оно называется **петлей**. Если у двух и более ребер совпадают множества концов (т. е. ребра соединяют одну и ту же пару вершин) и направление (если ребра ориентированы), то такие ребра называются **кратными**.

Во-вторых, простой граф не содержит петель и кратных ребер.

В-третьих, легко видеть, что ориентированный граф — это алгебраическая

Таблица 6.1: Виды определения графа.

	Неориентированный	Ориентированный
Простой граф	$E \subseteq \{\{a, b\} \mid a, b \in V, a \neq b\}$	$E \subseteq \{(a, b) \mid a, b \in V, a \neq b\}$
Граф	$E \subseteq \{\{a, b\} \mid a, b \in V\}$	$E \subseteq \{(a, b) \mid a, b \in V\}$
Псевдограф	$E \rightarrow \{\{a, b\} \mid a, b \in V, a \neq b\}$	$E \rightarrow \{(a, b) \mid a, b \in V, a \neq b\}$
Мультиграф	$E \rightarrow \{\{a, b\} \mid a, b \in V\}$	$E \rightarrow \{(a, b) \mid a, b \in V\}$

структура, в которой множеством-носителем является множество вершин, а единственным отношением на нем — множество ориентированных ребер, которое представляет собой бинарное отношение. Простой орграф характеризуется тем, что это отношение антирефлексивно. Запомним основательно это место, т. к. в формализме теории множеств это есть самое «правильное» из «простых» и самое «простое» из «правильных» определений графа.

Как правило, четырех кейсов (простой/непростой и ориентированный/неориентированный) вполне достаточно, чтобы изучать теорию графов. Объединим эти кейсы под общим названием **обычный граф**. Более того, они все сводятся к одному случаю — орграфу (т. е. алгебраической структуре с единственным отношением). Действительно, факторизуя множество орграфов по отношению эквивалентности, «забывающему» ориентацию ребер, мы получаем неориентированные графы как классы эквивалентности. Отметим, что в простом орграфе допустима ситуация, когда вершины соединены двумя ребрами, только в этом случае они должны быть противоположно направленными!

Если не говорить о факторизации, то неориентированный граф можно определить как структуру с симметричным отношением (да, при этом у нас будет по два противоположно направленных ребра между связанными вершинами, но и такое определение не лишено смысла). Тем не менее, избавление от ориентации путем перехода к фактор-множеству является более алгебраическим и более продуктивным способом. Как вы могли заметить на протяжении всей книги, факторизация — излюбленный (*архетипический*) прием математиков, прежде всего алгебраистов и логиков.

В качестве обобщения принято также рассматривать графы смешанного типа, в которых есть как ориентированные, так и неориентированные ребра.

Наконец, в-четвертых, **мультиграф** есть уже некая функциональная

структура, где связь между V и E устанавливается не отношением, а функцией (инцидентности) из некоторого множества E в множество упорядоченных или неупорядоченных пар вершин (**псевдограф** отличается, опять-таки, отсутствием петель, но допускает кратные ребра). Если функция $I : E \rightarrow V^2$ инъективна, то она определяет обычный граф. Однако, если данная функция не инъективна, то попадают случаи, когда $I(e_1) = I(e_2)$ при $e_1 \neq e_2$. В этом случае у нас возникают кратные ребра: e_1 и e_2 можно отождествлять с обычным ребром $\{a, b\}$, но взятым дважды (помеченным один раз символом e_1 , второй раз — e_2).

В принципе, если мы принимаем аксиому выбора или работаем с конечными множествами, мы можем прообраз $I^{-1}\{a, b\} = \{e_1, e_2, \dots\}$ упорядочить и найти его мощность $k\{a, b\} = \|I^{-1}\{a, b\}\|$, в результате чего от мультиграфа вернемся к обычному графу, у которого на ребрах будет задана функция кратности (если только нам не требуется различать кратные ребра по каким-то их дополнительным свойствам, например, по длине ребра-дороги или величине электрического сопротивления ребра-проводника). В таком случае можно говорить о том, что мультиграф — это граф, в котором отношение E является мультимножеством.

Ясно, что функциональная конструкция $(E \rightarrow V^2)$ обобщает графы и мультиграфы как ориентированные, так и неориентированные, и смешанные.

Комментарий 24.

В программировании графы также задаются по-разному, наиболее употребимые представления — это:

- вершины пронумерованы числами $0, 1, \dots, n-1$, ребра задаются списком или словарем из двухэлементных списков;
- матрица смежности, где вершины пронумерованы числами $0, 1, \dots, n-1$, а наличие единицы в ячейке (ij) матрицы означает наличие ребра (i, j) в графе (вместо единиц можно задавать кратности или метки);
- граф задается словарем, в котором индексами выступают номера вершин, а значениями — списки смежных с ними вершин (с учетом направления ребра и кратности)

Существует еще одно обобщение графа, когда функция действует не в V^2 , а в произвольную степень V^α , т.е. ребро в данном случае связывает не 2, а несколько вершин графа, и является мультиребром.¹ Такое обобщение графа называется гиперграфом. Ясно, что гиперграф можно рассматривать

¹По нашему мнению, лучше было бы такие ребра называть «гранями» или «мембранами».

как алгебраическую структуру с некоторым α -арным отношением, которое само по себе еще и является мультимножеством.

Наконец, можно взять структуру с отношениями различной арности и рассматривать гиперграф, в котором одно ребро может соединять произвольное подмножество вершин. При этом становится неважным, ориентирован такой граф или нет, поскольку для каждого случая упорядочения вершин мы можем запастись достаточным количеством элементов множества E . Более того, точно так же, как в случае мультиграфа, мы можем развернуть функцию в обратную сторону — от вершин к ребрам, и либо для произвольного набора вершин задавать количество ребер, их соединяющих (кратных ребер), либо указывать конкретное множество ребер, каждое из которых может обладать какими-то отличительными признаками (весом, цветом, направлением и т.д. — это уже задается функцией на ребрах или структурой самого ребра).

Таким образом, квинтэссенцией всех описанных нами конструкций графов, на наш взгляд, является следующее определение. **Гиперграфом** будем называть всякую функцию $G : \mathcal{P}(V) \rightarrow \mathcal{P}(E)$, где V — произвольное непустое множество, именуемое множеством вершин графа G , а E — произвольное множество, именуемое множеством ребер графа G . При этом элементы множества $G(v)$ ($v \subseteq V$, $G(v) \subseteq E$) называются инцидентными элементами множества v , если $G(v)$ непусто, кроме того, все вершины из v смежны друг другу, если $G(v)$ непусто. Функцию G будем называть также **функцией смежности**. Кроме того, что функция G не может быть какой угодно, она должна удовлетворять некоторым аксиомам (сравните с аксиомами меры):

$$\text{Graph1 } G(\emptyset) = E;$$

$$\text{Graph2 } G(v \cup w) = G(v) \cap G(w).$$

Неформально эти аксиомы можно понимать следующим образом. **Graph1** говорит, что пустому множеству вершин соответствуют все ребра графа. **Graph2** говорит, что чем больше множество вершин, тем меньше у этих вершин общих («накрывающих») ребер. На самом деле, если мы рассмотрим дополнительную функцию $H = E \setminus G$, то для нее будем иметь равенство $H(v \cup w) = H(v) \cup H(w)$, т.е. она сохраняет операцию объединения между множествами вершин и ребер. Тривиальным примером является постоянная функция $H(v) = \emptyset$, т.е. $G(v) = E$, для всех $v \subseteq V$. Это значит, что все ребра из E связывают сразу все вершины графа.

Заметим также, что если множество вершин графа G конечно, то достаточно определить функцию G лишь для одноточечных множеств, полагая $G(\{v_k\}) = \{e_{k,1}, \dots, e_{k,n_k}\}$, то есть для каждой вершины указать множество исходящих из нее ребер, после чего продолжить G на весь булеан $\mathcal{P}(V)$ с помощью аксиомы **Graph2**. То же самое можно сделать и в случае бесконечного

количества вершин для ребер, соединяющих одновременно не более чем конечный набор вершин. Для продолжения G на бесконечные подмножества V нам потребуется усиление правила **Graph2**, напоминающее аксиому счетной аддитивности меры.

Вообще, как видим, с теоретико-множественной точки зрения, у меры и графа может быть много общего, если то и другое мы определяем с помощью абстрактных инструментов **ZF**. Хотя заранее предположить такое сходство кажется невозможным.

Значение $G(v)$ ($v \subseteq V$) можно интерпретировать как множество всех ребер, «накрывающих» одновременно(!) все вершины множества v (и, возможно еще какие-то другие вершины). В случае обычного графа мы можем рассмотреть такой пример: вершины $V = \{1, 2\}$ и ребра $e_1 = (1, 2)$, $e_2 = (2, 1)$, $l = (1, 1)$ (петля). В этом случае мы будем иметь: $G(\{1\}) = \{l, e_1, e_2\}$, $G(\{2\}) = \{e_1, e_2\}$, $G(V) = \{e_1, e_2\}$. Как видим, $G(\{1\} \cup \{2\}) = G(\{1\}) \cap G(\{2\})$, а также $G(\{2\} \cup \{1, 2\}) = G(\{2\}) \cap \{1, 2\}$ и т.д.

Предположим, что мы взяли одно из классических определений графа, например, обычного графа, ребра в котором являются неупорядоченными парами, т.е. каждое ребро e имеет вид множества $\{a, b\}$, где $a, b \in V$. Далее мы определим

$$G(v) = \{e \in E \mid v \subseteq e\}. \quad (6.1)$$

Ясно, что в случае $e = \{a, b\}$ имеем $G(\{a, b\}) = \{e\}$, а также $G(\{a\})$ включает e и другие инцидентные a ребра. а также, возможно, петлю $\{a\}$, если она есть в графе, т.к. $\{a\} \subseteq \{a, b\}$. Для любого множества v из более чем двух вершин мы получим $G(v) = \emptyset$. В то же время, для пустого v , очевидно, $G(v) = E$, т.к. $\emptyset \subseteq e$ для любого ребра e . Таким образом, **Graph1** выполняется для G , определенной в (6.1).

Легко проверить, что функция G удовлетворяет и **Graph2**. Действительно,

$$\begin{aligned} G(v \cup w) &= \{e \in E \mid v \cup w \subseteq e\} = \{e \in E \mid (v \subseteq e) \wedge (w \subseteq e)\} = \\ &= \{e \in E \mid v \subseteq e\} \cap \{e \in E \mid w \subseteq e\} = G(v) \cap G(w). \end{aligned}$$

Таким образом, классическое определение графа включается в наше определение гиперграфа. Аналогично можно проверить остальные стандартные определения, только вместо отношения $v \subseteq e$ потребуется использовать предикат «все вершины из v инцидентны ребру e ».

Посмотрим теперь, как, наоборот, из общего определения задать классические графы. Пусть $V^{[2]}$ обозначает множество всех неупорядоченных пар вершин из V . Ясно, что $V^{[2]}$ — это подмножество $\mathcal{P}(V)$, элементы которого имеют мощность 1 или 2.

Для того, чтобы определить обычный (т.е. без кратных ребер и мультиребер) неориентированный граф на множестве вершин V , нам достаточно определить функцию инцидентности $G : V^{[2]} \cup \{\emptyset\} \rightarrow \mathcal{P}(V^{[2]})$ следующим

образом: $G(\{a, b\}) = \{\{a, b\}\}$, либо \emptyset ($a \neq b$) — определит все ребра с двумя вершинами, а $G(\{a\})$ должно включить $\{a\}$, если в графе должна быть петля в вершине a , а также включить все ребра вида $\{a, c\}$, если в графе должны быть ребра $\{a, c\}$. Равенство $G(\{a\}) = \emptyset$ означает, что вершина a изолированная, равенство $G(\{a, b\}) = \emptyset$ означает, что между вершинами a и b нет ребра.

Похожим образом задается и орграф как функция $G : V^{[2]} \cup \{\emptyset\} \rightarrow \mathcal{P}(V^2)$. При этом значение $G(\{a, b\})$ является подмножеством множества $\{(a, b), (b, a)\}$ и задает дуги, $G(\{a\})$ включает все ребра, входящие/исходящие из a , а также петлю (a, a) , если таковая нужна в графе.

Смешанный граф будет определяться функцией вида $G : V^{[2]} \cup \{\emptyset\} \rightarrow \mathcal{P}(V^2) \cup \mathcal{P}(V^{[2]})$ (отметим, что $V^{[2]} \subseteq \mathcal{P}(V)$ и $E = V^2 \cup V^{[2]}$).

Как видим, простая теоретико-множественная конструкция (функция из одного булеана в другой) доставляет нам все возможные определения графов.

Более того, имея функцию G , нам не нужно отдельно задавать множество вершин и ребер, т. к. они однозначно определяются из самой функции, а именно:

$$V(G) \rightleftharpoons \cup \text{dom}(G), \quad E(G) \rightleftharpoons \cup \text{ran}(G).$$

Кроме того, для ребра $e \in E(G)$ через $|e|$ обозначим его носитель, т. е. множество инцидентных ему вершин:

$$|e| \rightleftharpoons \{a \in V(G) \mid \exists v \subseteq V(G) : (a \in v) \wedge (e \in G(v))\},$$

например, в обычном графе $|\{a, b\}| = \{a, b\}$, в орграфе $|(a, b)| = \{a, b\}$.

Скажем, что граф G' является **подграфом** графа G (пишут $G' \subseteq G$), если

1. $V(G') \subseteq V(G)$;
2. $E(G') \subseteq E(G) \cap \{e \in E(G) \mid |e| \subseteq V(G')\}$;
3. $\forall v \subseteq V(G') : G'(v) \subseteq G(v)$.

Здесь первое условие означает, что у подграфа множество вершин является подмножеством множества вершин графа, второе условие — что множество ребер подграфа является подмножеством множества ребер графа, причем заведомо исключаются ребра, инцидентные вершинам, не входящим в $V(G')$, а третье условие гарантирует сохранение инцидентности ребер и вершин (поскольку ребра не обязаны быть жестко привязаны к вершинам в силу своей конструкции).

Графы G и G' назовем **изоморфными**, если существуют биекции $f_v : V(G) \leftrightarrow V(G')$ и $f_e : E(G) \leftrightarrow E(G')$ такие, что

$$G(v) = e \Leftrightarrow G'(f_v[v]) = f_e[e],$$

т. е. изоморфизм графов устанавливает взаимно однозначное соответствие между их ребрами и вершинами, сохраняя инцидентность. Образно говоря, изоморфные графы отличаются всего лишь обозначениями вершин и ребер. Отметим, что в общем случае **изоморфизмом** является пара функций (f_v, f_e) , хотя в ряде случаев может быть достаточно только f_v и требования сохранения смежности вершин. Так, для обычных орграфов достаточно потребовать биекцию $f : V(G) \leftrightarrow V(G')$ и чтобы в орграфе G существовало ребро (a, b) тогда и только тогда, когда в орграфе G' существует ребро $(f(a), f(b))$ (для неориентированного графа используем неупорядоченные пары). Если мы допускаем кратные ребра или мультиребра, то таких требований для изоморфизма недостаточно.

Если на графах вводится дополнительная структура, например, веса или иные метки ребер, то изоморфизм должен сохранять и эту структуру.

Кроме того, скажем, что $f = (f_v, f_e)$ является **автоморфизмом** графа G , если f есть изоморфизм графа G на себя. Ниже мы рассмотрим автоморфизмы чуть подробнее для обычных графов.

Отметим, что абстрактная теория графов более пригодна для сложных гиперграфов, где требуются нетривиальные зависимости вершин и ребер. Она позволяет вводить на таких графах различные теоретико-множественные, алгебраические и топологические конструкции, оперировать сопряженным графом, где ребра и вершины меняются ролями (вершину можно рассматривать как мультиребро), и т.д.

Заканчивая лирическое повествование о самых общих подходах к определению понятия «граф», скажем, что изучение столь глубоких конструкций не является целью данной книги, поэтому, чтобы не усложнять себе жизнь, в дальнейшем мы ограничимся только понятием обычного графа. При этом мы сразу договоримся, что базовой структурой для нас является **орграф**, заданный как алгебраическая структура (V, E) с единственным бинарным отношением E на множестве V и, возможно, какими-то дополнительными числовыми функциями на множествах V и/или E . А все остальное мы получаем методом факторизации выбранного множества орграфов, отождествляя те или иные признаки, которые будем считать избыточными. В ряде случаев будет полезно производить двойную факторизацию, т. е. факторизовать фактор-множество. К таким фактор-башням мы уже привыкли, когда строили системы чисел $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$, а также в алгебраических конструкциях вроде субнормального ряда подгрупп.

Условимся также, что термин «граф» будет по-умолчанию означать неориентированный граф, если иное не следует из контекста. Таким образом, граф — это класс эквивалентных орграфов, отличающихся только направлением отношения E на одной или более парах вершин.

6.2 Обычные графы

В теории графов сложился достаточно большой корпус терминов, которые необходимо вводить с самого начала, поскольку без них хоть какое-то существенное продвижение становится невозможным. И хотя мы приводим здесь довольно много терминов и определений, для составления более полной картины рекомендуем обратиться к видеокурсу д.ф.-м.н. А. Омельченко «Основы теории графов» [https://stepik.org/126] и классическим книгам [115, 120].

Называя мощность $\|V(G)\|$ множества вершин графа его **порядком**, мы можем рассмотреть класс всех простых графов порядка n . Точнее, мы ограничимся² графами с вершинами $1, \dots, n \in \omega$.

Обозначим множество всех простых орграфов с вершинами $1, \dots, n$ через $\bar{\mathcal{G}}_n$, а множество всех простых графов с вершинами $1, \dots, n$ через

Упражнение 6.2. Проверьте формулу! \mathcal{G}_n . Последнее является фактор-множеством над $\bar{\mathcal{G}}_n$. Очевидно, что $\|\mathcal{G}_n\| = 2^{n(n-1)/2}$, в то время как множество $\bar{\mathcal{G}}_n$ простых орграфов порядка n имеет мощность $2^{n(n-1)}$.

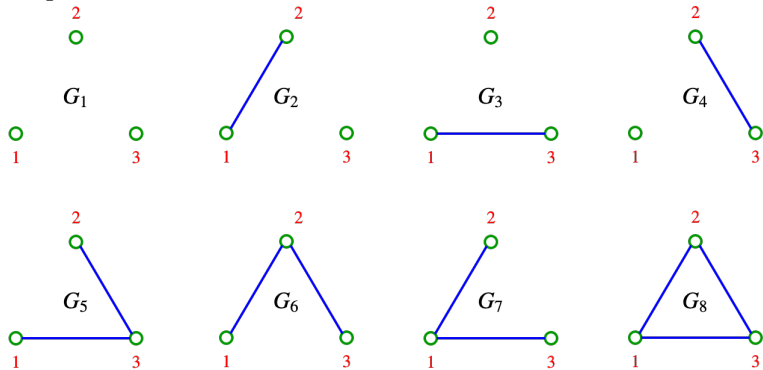


Рис. 6.1: Трехвершинные неориентированные графы.

Среди графов из \mathcal{G}_n могут быть изоморфные. Например, на рис. 6.1 представлены все простые (неориентированные) графы порядка $n = 3$ (соответствующие орграфы мы предлагаем домыслить читателю самостоятельно), среди которых есть три изоморфных графа с одним ребром и три изоморфных графа с двумя ребрами.

Вот он — фактор над фактором Множество \mathcal{G}_n факторизуется отношением изоморфизма графов на классы изоморфных графов. Внутри каждого класса эквивалентности графы отличаются только обозначениями своих вершин. Таким образом, класс изоморфных графов можно представить как

²Если строго не задать множество вершин, то им может быть любое множество мощности n , что породит бесконечные классы изоморфных графов, не являющиеся множествами, и нам каждый раз придется оговариваться, как устроен граф, выбирая изоморфный ему с вершинами $1, \dots, n$, чтобы иметь возможность сравнивать эти классы.

«обезличенный» граф, т. е. такой граф, у которого стёрли названия вершин.

Обычно такой граф принято называть **непомеченным**, а, соответственно, граф с помеченными вершинами — **помеченным графом**. Теоретико-множественная конструкция предполагает, что вершины графа помечены, т. к. элементы множества вершин различны по определению множества.

Поэтому для создания непомеченного графа приходится вновь прибегать к факторизации.

Архетип факторизации!

Имея теоретико-множественную конструкцию графа, мы закладываем в нее максимум различий между графами, поскольку вершины и ребра являются попарно различными множествами.³ Но в ряде задач нам неинтересна избыточно точная модель данных, поэтому от точной модели мы переходим к несколько огрубленной, объединяя в классы эквивалентности графы какого-либо одного вида. Например, рассматривая множество \mathcal{G}_n помеченных графов порядка n , мы можем перейти от них к непомеченным графам, факторизуя это множество по изоморфизму графов. Если же вместо инцидентности вершин и ребер для определения эквивалентности использовать иные структурные показатели графа, например, упорядочение вершин, имеющих общую смежную вершину, то можно получить другое разбиение множества \mathcal{G}_n на классы и работать уже с какой-то другой абстракцией графа. Подробнее мы остановимся на этом при изучении деревьев.

Как видим, непомеченных графов с тремя вершинами существует ровно 4 (т. е. 4 класса изоморфных графов с 3 вершинами). В общем случае количество непомеченных простых графов порядка n не представляется формулой в замкнутом виде или простым рекурсивным соотношением. Вообще, числовые характеристики для непомеченных графов, как правило, получить значительно труднее, чем для помеченных. Некоторую помощь в этом способна оказать теория вероятностей, позволяющая в ряде случаев отыскать предельную асимптотику числовых характеристик графов при $n \rightarrow \infty$. Так, число простых непомеченных графов порядка n асимптотически равно $2^{n(n-1)/2}/n!$, т. е. с ростом n сближается с отношением числа всех помеченных графов к числу перестановок номеров вершин.

Ниже мы еще коснемся темы решения перечислительных задач теории графов асимптотическими методами.

Аutomорфизмы графа G образуют группу с операцией композиции. Эта группа обозначается $\text{Aut}G$. Легко видеть, что если $G \in \mathcal{G}_n$, то $\text{Aut}G$ является подгруппой симметрической группы S_n . Как следствие, *порядок группы автоморфизмов графа порядка n делит число $n!$* .

Рассмотрим, например, группу автоморфизмов графа G_5 на рисунке 6.1.

³Стоит, однако, вспомнить, что и равенство множеств тоже есть некоторое соглашение об эквивалентных записях. Этот вопрос мы подробно разбирали в разделе 1.1.5.

Помимо тривиального автоморфизма существует автоморфизм (12).⁴ Так что, $\text{Aut}(G_5) = \{e, (12)\}$. У треугольника G_8 автоморфизмов существенно больше: $\text{Aut}(G_8) = S_3$. Последняя группа, как мы знаем, является группой симметрий треугольника. На самом деле, если посмотреть на автоморфизмы произвольного графа, то все они так или иначе представляют собой симметрии, производимые над различными частями графа. Поэтому автоморфизмы графа называют также симметриями графа. Свойства симметрий графа используются, например, при изучении молекул.

Заметим, что группа автоморфизмов графа далеко не всегда нормальна в S_n . В качестве примера достаточно взять граф вида $1 - 3 - 2$, для которого перестановка (12) является (единственным) автоморфизмом. Взяв перестановку (13), составим композицию

$$(13)(12)(13)^{-1} = (23),$$

которая также не является автоморфизмом, а значит, условие нормальности ($ghg^{-1} \in H$) для $\text{Aut}(1 - 3 - 2)$ не выполняется.

Рассмотрим обратную задачу. Пусть нам дан простой непомеченный неориентированный граф G порядка n . Сколько существует таких же, но помеченных графов? Для получения разметки таких графов нам нужно выполнить все возможные перестановки нумераций их вершин, т. е. $n!$ штук, и сократить на количество изоморфных графов, т. е. на порядок группы $\text{Aut}G$. В итоге получим, что из графа G можно получить $n!/|\text{Aut}G|$ различных (в смысле изоморфизма) помеченных графов.

Несмотря на простоту формулы, вычислить ее знаменатель в общем случае затруднительно. Вообще, нужно отметить, что поиск автоморфизмов и установление изоморфности графов — задача сложная, и на сегодняшний день не существует удовлетворительных (относительно) легкого проверяемых критериев (авто|изо)морфизма. Не найдено и быстро сходящихся алгоритмов для их машинного вычисления в общем случае.

Маршрут длины n в графе G (в том числе ориентированном) — это упорядоченный набор вершин и ребер этого графа $(v_0, e_1, v_1, e_2, \dots, e_n, v_n)$ такой, что ребро e_k инцидентно вершинам v_{k-1}, v_k , $k = \overline{1, n}$. То есть, ребро (или петля), стоящее в этом наборе между двух вершин, инцидентно данным вершинам. Такое определение маршрута исключает, например, случай $(1, e, 1, e, 1)$, где одно и то же ребро $e = \{1, 2\}$ проходится дважды, но при этом мы не проходим через вершину 2, хотя все вершины и ребра последовательно инцидентны.

Отметим, что маршрут не обязан соблюдать ориентацию ребер, и потому данный термин можно применять как к графам, так и орграфам.

⁴Напомним, что (12) обозначает перестановку, при которой 1 переходит в 2 и наоборот, а остальные точки неподвижны.

В случае орграфа возникает определение **пути** длины n как последовательности ребер (e_1, \dots, e_n) , где все ребра имеют направление от начала пути к концу, т. е. $\text{pr}_2(e_k) = \text{pr}_1(e_{i+k})$. Путь можно считать частным случаем маршрута, т. к. промежуточные вершины пути однозначно восстанавливаются по смежным направленным ребрам.

Соответственно, можно говорить о **склейке** путей или маршрутов (при этом последняя вершина первого пути должна быть первой вершиной второго пути), о **цепи** или **орцепи** (все ребра попарно различны), о **простой** цепи/орцепи (цепь/орцепь, у которой все промежуточные вершины попарно различны). Также вводится понятие **замкнутого маршрута** или **контура** (первая вершина совпадает с последней), **цикла** или **орцикла** (замкнутая цепь/орцепь) и **простого цикла/орцикла** (цикл без повторяющихся промежуточных вершин).

Две вершины в графе называются **связанными**, если существует маршрут, соединяющий их, и **сильно связанными** — если существуют два ориентированных пути, один из которых начинается в первой вершине и заканчивается во второй, а второй путь, наоборот, начинается во второй вершине, а заканчивается в первой. Вершина сама с собой связана и сильно связана по определению. Соответственно, **граф связан**, если любые две его вершины связаны, и **сильно связан** — если они сильно связаны. Обычно под связностью в (ор)графе понимается связность без учета ориентации.

Число c_n связанных помеченных неориентированных графов с n вершинами удовлетворяет соотношению

$$c_n = 2^{\binom{n}{2}} - \frac{1}{n} \sum_{k=1}^{n-1} k \binom{n}{k} 2^{\binom{n-k}{2}} c_k. \quad (6.2)$$

Отношение связности на вершинах графа является отношением эквивалентности. Это отношение эквивалентности разбивает множество вершин на классы эквивалентности, которые называются **компонентами (сильной) связности** графа.

*И снова факторизация
:)*

Так же, как в топологии, для графов существует понятие одно-, дву- и более связности. Будем говорить, что граф $G - a$ получен из графа G удалением вершины a , если $G - a$ отличается от G отсутствием вершины a и всех инцидентных ей ребер. Аналогично, удаление ребра e приводит к графу $G - e$, в котором удалено ребро e (а инцидентные ему вершины остаются).

Если граф G имеет k компонент связности, а граф $G - a$ имеет $k + 1$ компоненту связности, то вершина a называется **точкой сочленения** или **шарниром** графа G . Если граф G имеет k компонент связности, а граф $G - e$ имеет $k + 1$ компоненту связности, то ребро e называется **мостом** графа G .

Граф G называется **вершинно k -связным**, если удаление из него любых $k - 1$ вершин приводит к невырожденному (т. е. $k < n$) связанному графу. Об-

разно говоря, это означает, что между любыми вершинами графа существует как минимум k различных связующих путей. Граф G называется **реберно k -связным**, если удаление из него любых $k - 1$ ребер приводит к невырожденному связному графу.⁵

На основе этих определений вводятся две числовые характеристики графа G : вершинная связность $\kappa(G)$ и реберная связность $\lambda(G)$. Обе определяются как максимальное число k такое, что граф k -связен (в первом случае вершинно, во втором — реберно). Несвязный граф, очевидно, 0-связен. Наличие циклов в связном графе означает, что он является **локально реберно двухсвязным**, т. е. удаление циклового ребра не нарушает связности графа. Более того, имеет место

Теорема 6.1. *Граф G (неориентированный) является 2-связным тогда и только тогда, когда любые его два различных ребра принадлежат какому-то циклу.*

Упражнение
6.3.

Кроме того, нетрудно доказать, что

$$\kappa(G) \leq \lambda(G), \quad (6.3)$$

поскольку удаление вершин влечет и удаление ребер.

Как видим, на графах мы снова сталкиваемся с топологической номенклатурой понятий: путь, связность, многосвязность. Через них в дискретные объекты, которыми являются графы, проникает топологический архетип связности–непрерывности. Можно также говорить о топологии на графе, точнее, на множестве его вершин, индуцированной множеством ребер.

Например, на множестве вершин вводится расстояние как минимальная длина маршрута между ними. Такое расстояние является метрикой, если дополнительно сказать, что расстояние от a до a равно нулю, а расстояние между вершинами из разных компонент связности равно ∞ . Метрика, как и положено, индуцирует топологию на множество $V(G)$, но эта топология дискретна, и поэтому неинтересна для исследования.

Однако существует способ задать топологию, основываясь на путях в графе G . Именно, для всякой вершины $v \in V(G)$ обозначим через Ov множество всех вершин, достижимых из v (вершина w **достижима** из v , если $w = v$ или существует путь с началом в точке v и концом в точке w). Легко проверить, что такое семейство окрестностей образует базу топологии на множестве вершин V .

Например, в графе $a \rightarrow b$ (две вершины и одна стрелка) открытыми множествами будут $\emptyset, \{b\}, \{a, b\}$. Это не что иное, как связное двоеточие.

⁵Требование невырожденности оставшегося графа связано с тем, что любой граф порядка n после удаления из него $n - 1$ вершины становится тривиальным связным графом. А в случае удаления ребер вырожденность оставшегося графа означает, что он был таковым изначально.

Наконец, для неориентированного графа топологию можно задать погружением в \mathbb{R}^3 . При этом вершины графа переходят в точки, а ребра — в гладкие пути, соединяющие эти точки (пути не должны пересекаться!) Заметим, что для некоторых графов можно ограничиться плоскостью \mathbb{R}^2 (такие графы называются **планарными**), а для каких-то это не удастся. При погружении в \mathbb{R}^3 топология на графе индуцируется из стандартной евклидовой топологии. При этом топологическая связность графа эквивалентна его связности.

Далее можно развить аналогичную топологии теорию. Можно говорить о гомотопии путей и циклов, о классе гомотопных циклов, проходящих через выделенную вершину графа. Это приводит нас к понятию фундаментальной группы графа и различным усилениям понятия связности.

Подробнее о топологических свойствах графов см. в [113].

Для каждой вершины a графа G можно определить степень $\deg(a)$ как количество инцидентных ей ребер (при этом петля учитывается дважды, т. к. она предоставляет два входа-выхода из вершины). Кроме того, можно ввести полустепень исхода $\deg^+(a)$ и полустепень входа $\deg^-(a)$, соответственно, как число исходящих и число входящих в эту вершину ребер (в случае орграфа). Ясно, что $\deg(a) = \deg^+(a) + \deg^-(a)$. Кроме того, ясно, что $\sum_a \deg(a) = 2\|E(G)\|$. | Упражнение 6.5.

Наконец, граф без циклов именуется **лесом**, а связный граф без циклов — **деревом**. Легко видеть, что лес — это такой граф, в котором все компоненты связности являются деревьями. Характеристики связности (как реберной, так и вершинной) дерева всегда равны 1. Лес можно также определить как подграф дерева. Например, если в дереве удалить одну вершину, то мы получим лес, содержащий столько деревьев, сколько было инцидентных этой вершине ребер.

Заметим, что лес и дерево определяются без учета ориентации на графе, и чаще всего их относят к неориентированным графам. Тем не менее, выбрав у дерева какую-либо вершину (называемую **корнем**), мы можем единственным способом ориентировать его ребра от корня, так что каждое неориентированное дерево порядка n порождает ровно n **корневых** деревьев, ориентированных от корня.

Соответственно, каждый лес с N деревьями порядков n_1, \dots, n_N порождает $n_1 \cdots n_N$ корневых лесов, ориентированных от их корней.

Следует отметить, что термин *ориентированное дерево (лес)* есть более широкое понятие, чем *корневое дерево (лес)*, поскольку он обозначает пересечение понятий орграф и дерево (лес), т. е. ориентированное дерево — это дерево, где все ребра каким-то способом ориентированы (не обязательно от корня).

Соберем для сравнения введенные понятия в таблицу (см. табл. 6.2).

Деревья (как математическая структура) играют огромную роль во многих областях знаний. Прежде всего, организация данных в виде дерева позво-

Таблица 6.2: Терминология.

	Общий термин	Ориентированный
Связная послед-ть	Маршрут	Путь
Замкнутость	Замкнутый маршрут	Контур
Без дублирующих ребер	Цепь	Орцепь
Замк. без дубл. ребер	Цикл	Орцикл
Без дублирующих вершин	Простая цепь/цикл	Простая орцепь/орцикл
Степень	К-во инцидентных ребер	Полустепень входа/исхода
Лес/дерево	Нет циклов	Ребра ориентированы

ляет осуществлять быстрый поиск нужного узла. Так, если дерево бинарное (т. е. у каждой вершины исходящая степень не более 2), то поиск нужного узла имеет порядок скорости работы $O(\log_2 n)$, где n — количество вершин в дереве. Кстати, обычно говорят не порядок дерева/леса, а объем дерева/леса.

В виде дерева обычно организуют различные каталоги, файловые директории и библиотеки (хотя потом и накручивают на них кросслинки как дополнительные ребра).

Классификация видов животных и растений — тоже дерево. И т.д.

Комментарий 25. О пользе деревьев.

Организация каталога с описанием товаров для интернет-магазина в виде дерева также очень полезна. Например, мы строим древо с уровнями

Категория→Подкатегория→Семейство→Серия→Модель→Артикул,

объединяя конкретные товары (например, автомобили определенной конфигурации) в модели, в пределах которой отличие может быть только в цвете или каких-то дополнительных опциях, далее, модели — в серии, где отличие происходит уже по более весомым параметрам, далее серии — в семейства (или линейки), объединяющие родственные, но уже сильно отличающиеся товары (скажем, у автомобилей это могут быть буквенные обозначения семейств, означающих габаритный класс или целевой сегмент рынка).

Далее мы можем построить контентное описание автомобилей по правилу «от общего к частному». Так, на уровне семейства мы описываем только отличительные признаки этого семейства, на уровне серии мы уточняем описание, пополняя его отличительными признаками этих серий, на уровне моделей мы почти полностью описываем автомобили, оставляя незаполненными только отличия внутри моделей (цвет и т.п.), наконец на уровне артикула мы завершаем описание, указав конкретные отличительные признаки конкретного товара. В итоге у нас получается, что каждая контентная сущность описывается в одном месте (что позволяет уменьшить число ошибок и сократить работу по модификации контента), а полное описание артикула складывается из описаний всех вышестоящих родительских узлов по правилу наследования вдоль дерева.

Иначе говоря, древовидное структурирование информации позволяет существенно упростить и ускорить (в $O(\ln n)$ раз) создание контента товаров, а также оптимизировать работу над ошибками.

Для неориентированного графа G всегда существует **остовный лес**, т. е. такой подграф графа G , который содержит все его вершины и при этом является лесом. Соответственно, для связного графа существует остовное дерево.

Если у нас имеется большой граф (с миллионами вершин и ребер), то в ряде случаев (например, для построения поискового кеша) его нужно упорядочить в виде дерева. Для этого подходят алгоритмы поиска остовного дерева:

- Алгоритм «поиска в глубину» — поиск остовного дерева связного графа, в результате работы которого все ребра делятся на два класса: ребра, принадлежащие остовному дереву, и ребра, соединяющие пары вершин, лежащие на корневом пути (т. е. ребра, сокращающие путь);
- Алгоритм «поиска в ширину» — поиск остовного дерева с одновременной минимизацией расстояния от корня до вершин. При этом часть ребер графа (не взятых в дерево) могут соединять вершины, лежащие на разных корневых путях (кросслинки). Данный алгоритм оптимизирует время поиска.
- Алгоритм Дейкстры — поиск кратчайших путей из одной заданной вершины во все остальные с учетом весов, заданных на ребрах. Например, позволяет вычислять кратчайшие (по длине) дороги из одного города в другой, если известны длины дорог между городами (Google Maps), или рассчитать минимальную стоимость перелета с пересадками.

Эти алгоритмы могут также дать ответ на вопрос о связности графа, более того, на выходе мы получим остовный лес, каждое дерево которого будет остовным деревом какой-либо компоненты связности исходного графа, оптимизированное тем или иным способом.

Другой вариант остовного подграфа для связного графа — это остовный простой(!) цикл, содержащий все ребра (а, следовательно, и все вершины) графа. Если в графе существует такой цикл, то и такой цикл, и такой граф называется **эйлеровым**. Известна следующая

Теорема 6.2 (Эйлера). *Конечный неориентированный граф G эйлеров тогда и только тогда, когда он связан и все вершины имеют четную степень.*

С этой теоремой связана Задача о кёнигсбергских мостах: можно ли пройти по всем мостам города, вернувшись в точку отправления, не проходя ни по какому мосту дважды? Ответ — можно, если каждому острову инцидентно четное количество мостов (под островом понимается часть суши, ограниченная водой и, возможно, границей города). Если говорить конкретно о карте Кёнигсберга XVIII века, то оказывается, что на ней нет эйлерова пути, т. к. все острова инцидентны трем, либо пяти мостам (см. рис. 6.2).

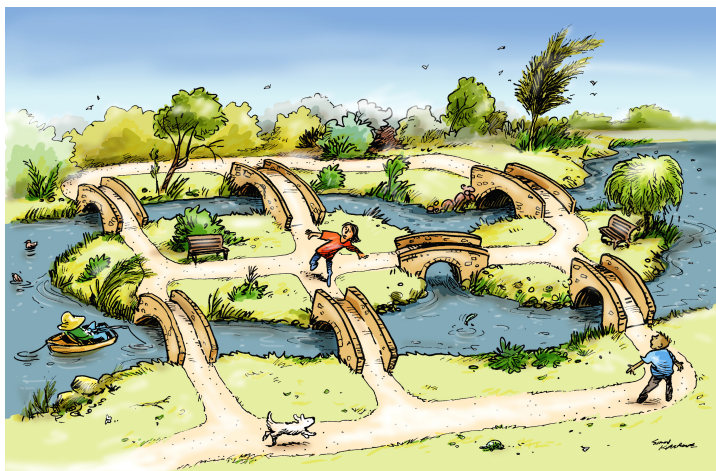


Рис. 6.2: Задача о мостах Кёнигсберга
(размещено с разрешения автора <https://simonkneebone.com/>)

Помимо эйлеровых циклов существуют также **гамильтоновы циклы** — простые циклы, включающие все вершины графа (но не обязательно все ребра) ровно по одному разу. Известно, что в общем случае задача поиска гамильтонова цикла сильно сложнее (по времени работы алгоритма), чем задача поиска эйлерова цикла, что является критичной проблемой в биоинформатике при сборке генома по известным различным фрагментам одинаковой длины.

Рассмотрим эту задачу чуть подробнее. Пусть у нас имеется конечный алфавит $A = \{a_0, a_1, a_2, \dots\}$ мощности n (например, различные пары нуклеотидов), и мы выписываем все слова длины k (так называемые k -риды или фрагменты ДНК). Ясно, что всего таких различных слов может быть n^k (наборы этих слов могут быть получены в результате секвенирования биополимеров.)

При этом, очевидно, многие части слов будут совпадать. Задача состоит в том, чтобы восстановить слово в этом же алфавите (геном), но уже длины n^k , зацикленное так, что все слова длины k будут отрезками данного слова (с учетом его цикличности). Иначе говоря, нас интересует последовательность $(a_{i_1}, a_{i_2}, \dots)$ такая, что все слова $a_{i_1} \dots a_{i_k}$, $a_{i_2} \dots a_{i_{k+1}}$, и т.д., $a_{i_{n^k}} \dots a_{i_{k-1+n^k}}$, получаемые сдвигом на 1, различны. Такая последовательность называется **последовательностью де Брёйна**, а если ее записать в виде цикла длины n^k , то такой цикл называется **циклом де Брёйна**. На рис. 6.3 показан цикл де Брёйна длины 8 для алфавита $A = \{0, 1\}$ и слов длины 3, а также граф де Брёйна, который помогает восстановить такой цикл.

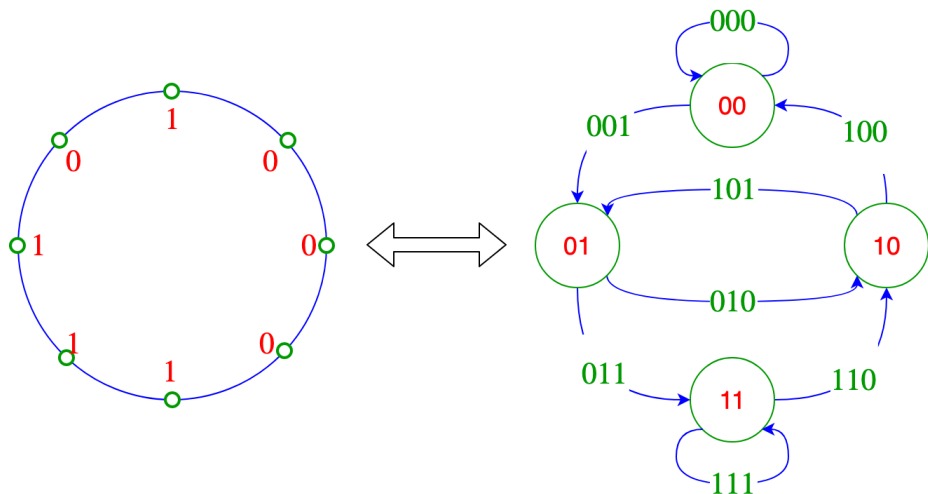


Рис. 6.3: Последовательность и граф де Брёйна для $n = 1, k = 3$.

Поясним. Сначала строится граф порядка n^{k-1} , вершины которого помечены всеми возможными словами длины $k - 1$, а всякое ребро связывает вершину a с вершиной b , если существует слово длины k такое, что a является началом этого слова, а b — окончанием. Само это слово длины k является меткой для данного ребра. Например,

$$00 \xrightarrow{000} 00, \quad 00 \xrightarrow{001} 01, \quad 01 \xrightarrow{010} 10, \quad 10 \xrightarrow{101} 01, \dots$$

Начать удобно со слова $a_0 a_0 \dots$, поместив его в очередь. Затем берем слово из очереди, приписывая ему в конце различные символы алфавита A , получаем одновременно новые n вершин и ребер. При этом новые вершины помещаем в очередь (если таковые появились), а текущую удаляем из нее.

Получив граф, ищем эйлеров цикл, т. е. такой цикл, который обходит граф по всем ребрам, не проходя никакое ребро дважды. В итоге мы получаем последовательность ребер, которая и дает нам искомые слова длины k , расположенные по циклу де Брёйна. Заметим, что одновременно гамильтонов цикл

для графа де Брёйна дает нам слова длины $k - 1$, расположенные в цикле де Брёйна (в нашем примере это слова 00-01-11-10 и цикл 0-0-1-1).

Одним из первых алгоритмов поиска эйлерова цикла является алгоритм Флёрри (1883), однако его скорость работы составляет порядка $O(\|E(G)\|^2)$. Современные алгоритмы, например, **алгоритм на основе циклов**, имеют линейный порядок относительно количества ребер, т. е. $O(\|E(G)\|)$.

6.2.1 Разнообразие деревьев

Итак, под термином *дерево*, как правило, понимается помеченный связный граф без циклов, а под термином *корневое дерево* — ориентированное от корня дерево. Очевидно, что дерево есть класс эквивалентных корневых деревьев, где отношение эквивалентности «забывает» о наличии корня или, что то же самое, об ориентации на корневых деревьях. Таким образом, дерево есть первый шаг факторизации на корневых деревьях, а корневые деревья следует рассматривать как основу для дальнейшего определения различных классов деревьев.

Для корневых деревьев вводятся некоторые специальные термины и характеристики, например:

- высота дерева — это максимальная длина пути, начинающегося в корне;
- высота вершины — это длина пути из корня в эту вершину;
- уровень или слой дерева — множество вершин одинаковой высоты;
- предок (родитель) вершины a — вершина, связанная с a и имеющая меньшую высоту (высоту на 1 меньше);
- потомок (дочка) вершины a — вершина, связанная с a и имеющая большую высоту (высоту на 1 больше);
- лист — вершина без потомков (лист инвариантен относительно выбора корня дерева, кроме случая, когда лист выбирается корнем);
- ветвь с корнем a — поддерево, содержащее вершину a и всех ее потомков;
- средняя степень ветвления дерева — $\sum_a \deg^+(a)/k$, где k — число нелистовых вершин, суммирование производится по нелистовым вершинам.

Все эти термины теряют смысл в некорневом дереве.

Для некорневых деревьев существует несколько эквивалентных критериев. А именно, справедлива

Теорема 6.3. Для любого простого графа G следующие утверждения эквивалентны:

- (1) G — дерево,
- (2) любые две вершины графа G соединены единственным простым путем,
- (3) в связном графе G все ребра являются мостами,
- (4) G — связный граф, в котором $\|E(G)\| + 1 = \|V(G)\|$.

Предлагаем читателю самостоятельно разобраться в доказательстве этой теоремы. Скажем только, что импликация (1) \Rightarrow (2) почти очевидна, т. к. наличие двух путей приводит к циклам, импликация (2) \Rightarrow (3) также почти очевидна, импликация (3) \Rightarrow (4) основана на том, после изъятия $n - 1$ моста останется n компонент, т. е. граф без ребер, а импликация (4) \Rightarrow (1) следует из того, что если G — не дерево, то в нем есть цикловые ребра, удаление которых не нарушает связность, а после удаления всех цикловых ребер мы придем к дереву, для которого в силу уже доказанного выполняется то же равенство (хотя ребер стало меньше).

*Упражнение 6.6.
Докажите теорему.*

Заметим также, что свойство графа быть деревом эквивалентно тому, что его характеристики связности равны 1: $\kappa(G) = 1 = \lambda(G)$.

Комментарий 26. Классификация по дереву (DecisionTree)

В этом сюжете мы снова вернемся к машинному обучению и рассмотрим на это раз метод классификации с помощью деревьев.

Итак, пусть у нас имеется два класса точек — красные и синие — на плоскости xOy (см. рис. 6.4). Наша задача состоит в том, чтобы с помощью последовательно задаваемых нехитрых математических вопросов, максимально точно разрезать плоскость на ячейки так, чтобы в каждой из них оказались точки только одного цвета (или, по крайней мере, вероятность другого цвета была бы минимальной, т. е. энтропия распределения цветов в ячейке должна быть минимальной).

Самый простой способ классификации — это задавать вопросы с двоичным ответом «да» или «нет», тем самым создавая ветвление в виде бинарного дерева. Например, начав с координаты x , мы можем спросить $x < 4.5$? Как видим на картинке 6.4, это не самый плохой вариант для первоначального деления. Как видим, условие $x = 4.5$ задает гиперплоскость (прямую), делящую нашу плоскость на 2 части, что напоминает нам классификацию по методу SVM. На самом деле, здесь действительно можно применять SVM и получать довольно изощренные вопросительные высказывания для получения бинарного ответа, но для экономии вычислительных ресурсов желательно использовать его упрощенный вариант с гиперплоскостью, ортогональной некоторой оси пространства (в нашем случае оси Ox).

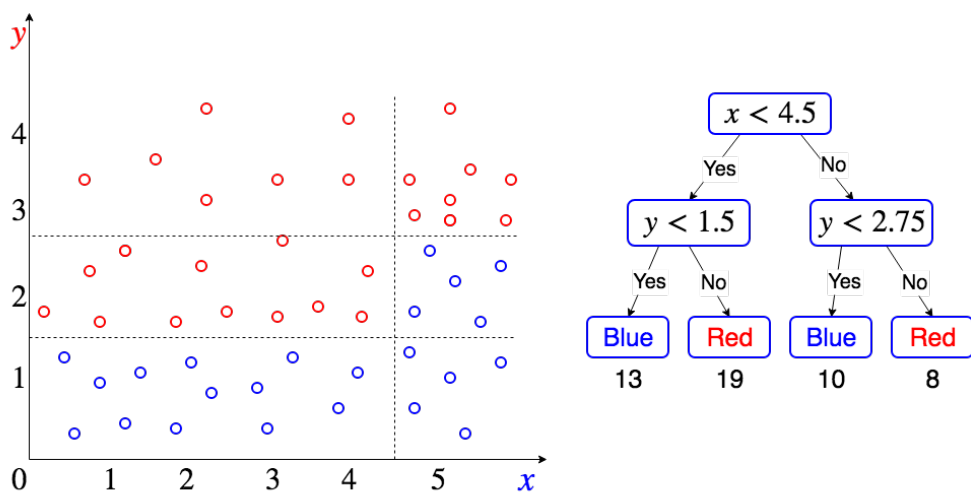


Рис. 6.4: Дерево решений

Проделав первое разделение, мы видим, что в левой и правой части все еще есть представители обоих классов (красного и синего). Каждую часть мы теперь рассматриваем как новую задачу классификации и снова производим деление по условию — своему для каждой части плоскости. В первом случае мы разделяем классы гиперплоскостью $y = 1.5$, во втором — гиперплоскостью $y = 2.75$. В результате у нас выстраивается бинарное дерево, вершины которого содержат разделяющие условия, ветви соответствуют вариантам ответа на такое разделение (см. рис. 6.4), а листья содержат ответ дерева, причем у каждого листа можно посчитать его «вес», т. е. количество объектов, соответствующих этому листу (на рисунке вес обозначен числами под листовыми узлами).

Построив такое дерево, мы можем применять его к новой точке для определения ее класса. Например, взяв точку $(3, 5)$, мы спустимся по дереву через условия $x < 4.5$ (yes) и $y < 1.5$ (no), и получим ответ, что эта точка должна быть красного цвета.

Как в любом методе, здесь существуют свои нюансы применимости и интерпретации результата. Так, построение дерева «до упора», т. е. пока в листьях этого дерева не окажется строго один класс точек (т. е. листья станут «чистыми»), может привести к переобучению. То есть мы получим очень большое и очень точное дерево, которое будет идеальным лишь для обучающей выборки, но на новых данных начнет часто ошибаться.

Поэтому в программных пакетах типа *scikit-learn* для Python подобные алгоритмы имеют большое число настроек. Например, можно ограничить высоту дерева, максимальное количество листьев, минимальный вес листьев.

Кроме того, можно для каждого узла дерева посчитать энтропию распреде-

ления классов, зная долю точек каждого класса в общем весе узла дерева. Так, в корне дерева из нашего примера мы имеем дело с полным объемом данных, и вероятности красного (p_r) и синего (p_b) цвета составляют

$$p_r = 27/50, \quad p_b = 23/50,$$

так что энтропия составит

$$E = -p_r \log_2 p_r - p_b \log_2 p_b = 0.995,$$

что почти достигает максимума для бинарного распределения (который равен 1 и достигается при равных вероятностях).

Аналогично считаем энтропию для дочерних узлов:

$$E_{left} = 0.974, \quad E_{right} = 0.991.$$

На рис. 6.5 в узлах дерева приведены значения энтропии, а числами под ними указано распределение классов в этих узлах (черным — вес узла, синим — количество в нем синих точек, красным — красных точек).

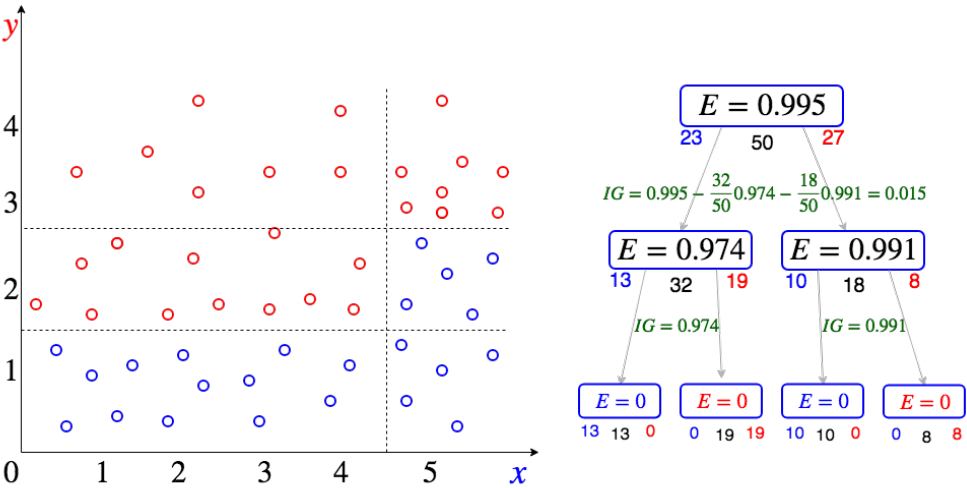


Рис. 6.5: Энтропия на узлах дерева

Как видим, на листьях дерева из примера энтропия обнуляется, поскольку они содержат «чистые» (однозначные) ответы.

На практике, конечно же, дерево может быть урезано (выбором параметров, о которых было сказано выше) на таких узлах, где ответ все еще будет неоднозначный, и потому энтропия будет отличаться от нуля. Тем не менее, при построении дерева система должна стремиться к тому, чтобы каждый переход от родительского узла к дочерним как можно быстрее уменьшал энтропию. Это требование также учитывается при построении деревьев решений стандартными алгоритмами `scikit-learn`.

Для того, чтобы отслеживать сокращение энтропии, вводится дополнительный функционал на дереве, который называется **Information Gain** (можно перевести как рост информации) и вычисляется по правилу:

$$IG = E - \sum_i w_i E_i,$$

где E — энтропия родительского узла, E_i — энтропия i -го дочернего узла, w_i — нормированный вес дочернего i -го узла, т. е. количество точек выборки, соответствующее этому дочернему узлу, деленное на количество точек выборки, соответствующее родительскому узлу (на рис. 6.5) эти количества обозначены черными числами.

В нашем примере мы видим, что значение IG между корнем дерева и его потомками невелико, всего 0.015, т. е. на первом уровне классификации энтропия упала несильно, а при переходе на следующий уровень IG резко возрастает, поскольку на втором уровне дерева мы получаем чистые (однозначные) ответы дерева решений.

Ранее мы приводили формулу (6.2) для числа связанных помеченных графов на n вершинах. Количество помеченных графов (без требования связности) на n вершинах равно $2^{n(n-1)/2}$. Если же снять нумерацию с вершин графа, т. е. факторизовать множество графов по изоморфизму, то разнообразие помеченных графов становится существенно уже, их количество асимптотически равно $2^{n(n-1)/2}/n!$.

Аналогичными вопросами можно задаться в отношении деревьев и лесов.

До сих пор мы рассматривали лишь два вида деревьев (корневые и некорневые). В обоих случаях они являлись помеченными графами, т. к. строились на пронумерованных вершинах. Однако видов деревьев существует очень много. Это разнообразие достигается с помощью факторизации класса однотипных деревьев по какому-либо признаку (например, снятие или частичное снятие нумерации вершин), либо добавлением некоторых условий на деревья (например, бинарные деревья или деревья с ограниченной степенью ветвления). Ясно, что количество деревьев объема n должно сильно зависеть от типа этих деревьев или, что то же самое, от глубины детализации фактормножества, построенного на корневых помеченных деревьях.

Известна **теорема Кэли** о том, что количество помеченных (некорневых) деревьев с вершинами $1, \dots, n$ равно n^{n-2} . На рис. 6.6 представлены все такие деревья для $n = 4$. Если выделить корневую вершину в каждом дереве, то получится n^{n-1} корневых помеченных деревьев с n вершинами (рисунок содержал бы вчетверо больше деревьев, поэтому он здесь не приводится).

Если снять пометки с вершин, то на 4 вершинах получим 2 класса изоморфных (некорневых) деревьев — простую цепь длины 3 и граф, который

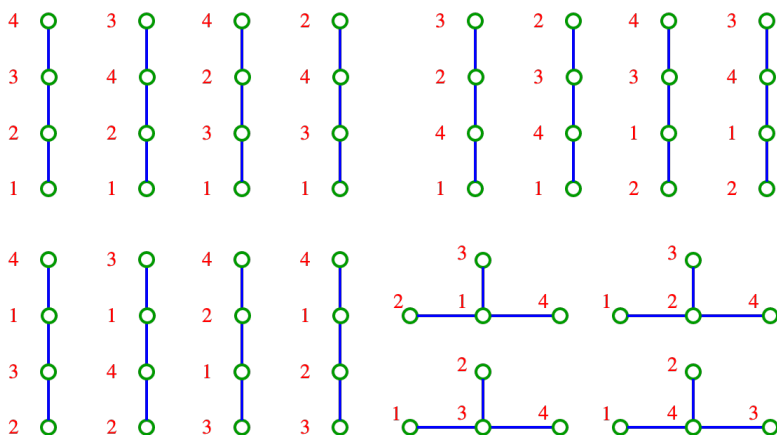


Рис. 6.6: Четырехвершинные помеченные (некорневые) деревья.

еще называется звездой или полным двудольным графом $K_{1,3}$ (1 вершина — центр, 3 вершины — лучи).

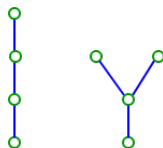


Рис. 6.7: Четырехвершинные непомеченные (некорневые) деревья.

Если теперь вернуть отметку корней на эти непомеченные деревья, то их количество возрастет не в четверо, как могло бы показаться, а всего вдвое, поскольку отсутствие нумерации на вершинах некоторые разметки корней на одном и том же непомеченном дереве отождествляет (для цепи из 4-х вершин корнем можно выбрать либо внутреннюю вершину, либо крайнюю, а для звезды $K_{1,3}$ — либо центр, либо луч).

На рис. 6.8 представлены корневые, но *непомеченные* деревья с 4 вершинами.

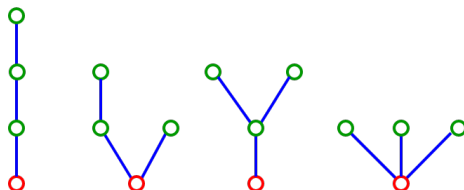


Рис. 6.8: Четырехвершинные непомеченные корневые деревья.

Как уже отмечалось выше, некоторые графы можно уложить в плоскость так, что все их ребра попарно не будут пересекаться. Очевидно, что все деревья являются планарными графами. Однако, укладка в плоскость позволяет

циклически упорядочивать (нумеровать) инцидентные ребра, обходя их общую вершину по кругу в положительном направлении (против часовой стрелки). Непомеченное дерево с таким упорядочением, индуцированным ориентацией плоскости, называется **плоским деревом**. В этом случае перестановка местами разных (неизоморфных) ветвей дерева дает, вообще говоря, другое плоское дерево.

При наличии корня в плоском дереве, у нас появляется возможность линейно упорядочить исходящие из одной вершины ребра, начиная обход от входящего в эту вершину ребра. Проблема возникает только с корневыми ребрами, т. к. в корень не входит никакое ребро. Поэтому для корневых ребер можно выбрать несколько упорядочений, а чтобы указать, какое именно упорядочение выбрано, добавляется фиктивное ребро, входящее в корень (как разделитель для исходящих ребер), а лишняя вершина (фиктивный корень) скрывается так, словно дерево посажено в землю. Соответственно, такие деревья называются **посаженными**. На рис. 6.9 приведены все плоские посаженные непомеченные деревья объема 4. Как видим, их стало чуть больше в сравнении с непомеченными корневыми деревьями того же объема, поскольку перестановка ветвей обеспечивает появление двух разных деревьев (второе и третье на данном рисунке).

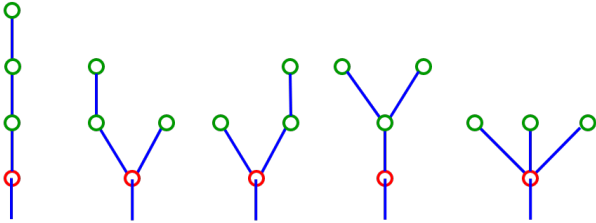


Рис. 6.9: Четырехвершинные плоские посаженные деревья.

Плоские посаженные деревья можно определенным способом пометить,⁶ оставляя ориентацию и порядок ребер с общей вершиной, но не переходя к сквозной нумерации вершин, как в обычном помеченном дереве. Именно, припишем корню однокомпонентный кортеж (0) . Затем, предполагая, что вершина v помечена кортежем $(0, d_1, \dots, d_k)$, а ее дочерние вершины имеют порядок $v_0 < v_1 < \dots < v_N$, то этим вершинам дадим, соответственно, метки $(0, d_1, \dots, d_k, 0)$, и т.д., $(0, d_1, \dots, d_k, N)$.

Такая нумерация напоминает определение действительных и сюрреальных чисел через последовательности с той лишь разницей, что в числах мы используем всего лишь две цифры — 0 и 1, — а здесь мы ничем не ограничиваем степень ветвления и количество чисел для нумерации вершин. Если каждой вершине приписать также индекс высоты в виде $x^{h(v)}$, где $h(v)$ есть

⁶Поэтому называть их непомеченными было бы неверно. Но и путать с произвольно помеченными тоже нельзя.

высота вершины v , то каждый корневой путь в плоском посаженном дереве будет отвечать многочлену с натуральными коэффициентами (с нулевым свободным членом, т. к. в корне стоит число 0).

Кроме того, если из дерева убрать корень, то получится лес из плоских посаженных деревьев, причем метки вершин на этих деревьях имеют ровно тот же самый вид, за исключением меток корней — они пронумерованы в соответствии со своим порядком в изначальном дереве. То есть удаление корня в плоском посаженном дереве приводит к лесу из плоских посаженных деревьев с пронумерованными деревьями.

Еще один часто встречающийся способ нумерации вершин состоит в следующем. Среди всех помеченных корневых деревьев выбираются только такие, у которых нумерация возрастает в направлении ориентации, т. е. корень всегда имеет номер 1, а дочки всякой вершины дерева имеют номера большие, чем сама эта вершина. Такие деревья называются **рекурсивными** [123]. Рекурсивное дерево строится начиная с корневой вершины с номером 1 и последовательным добавлением вершины с номером $j + 1$ к любой из вершин с номерами $1, \dots, j$. Легко видеть, что существует $(n - 1)!$ рекурсивных деревьев объема n . На рис. 6.10 представлены все 4-вершинные рекурсивные деревья.

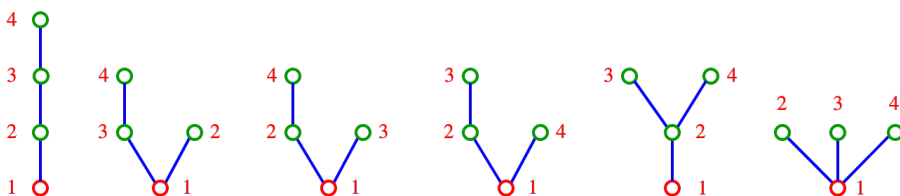


Рис. 6.10: Четырехвершинные рекурсивные деревья.

Рекурсивные деревья имеют применение, например, при моделировании эпидемий. Интересно также, что существует взаимно однозначное соответствие между рекурсивными деревьями объема n и циклами длины $n - 1$.

Таким образом, если нам дано некоторое множество деревьев (помеченные корневые заданного объема), то мы можем конструировать из него новые виды деревьев путем факторизации, т. е. отождествления несущественных признаков (нумерации, выделения корня, упорядочения и т. п.) или банального выделения подмножества (как в случае рекурсивных деревьев). Полученное фактор-множество будет моделировать какой-то новый вид деревьев, а каждый класс эквивалентности будет представителем такого вида. В случае корневых деревьев мы можем, удаляя корень, получать леса, состоящие из деревьев того же вида.

На рис. 6.11 представлены 6 деревьев объема $n = 4$ из шести разных классов. Класс T_0 — это помеченные корневые деревья, класс T_1 — помеченные

некорневые деревья (факторизация «забывает» корень), класс T_2 — рекурсивные деревья (ограничение на способ нумерации вершин), класс T_3 — плоские посаженные деревья (факторизация «забывает» нумерацию, но оставляет порядок ветвей), класс T_4 — корневые непомеченные деревья (факторизация «забывает» нумерацию), класс T_5 — некорневые непомеченные деревья (факторизация сохраняет только инцидентность).

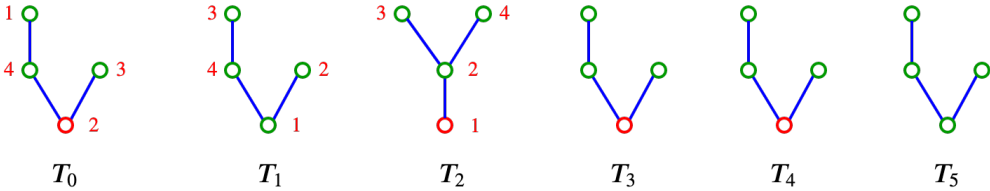


Рис. 6.11: Различные типы деревьев.

Здесь мы сталкиваемся с такой же ситуацией, как в разделе 1.1.5, где мы говорили о равенстве множеств. Уровень детализации может быть выбран разный за счет того, как мы хотим отождествлять изначально разные деревья. Вспомним также, что мы определяли деревья начальных множеств и мультимножеств, приравнивая отношение предок–потомок к отношению принадлежности на множествах. И если начальному мультимножеству взаимно однозначно соответствует корневое непомеченное дерево, то обычному начальному множеству соответствует корневое непомеченное дерево, в котором ветви, инцидентные общей вершине, попарно неизоморфны (см. 1.1.1, 1.1.6, 1.1.9 о гамма-деревьях и скобочных записях).

Заметим, что если на мультимножествах добавить упорядочение элементов, то им будут соответствовать плоские посаженные деревья, а также скобочные записи, удовлетворяющие аксиомам $(b1)$, $(b4)$, $(b5)$ из раздела 1.1.6. Количество таких записей с n открывающими скобками, как мы знаем, равно числу Каталана

$$C_{n-1} = \frac{1}{n} \binom{2n-2}{n-1}.$$

Соответственно, это же число есть количество всех плоских посаженных деревьев объема n .

Чтобы ответить на вопрос о числе непомеченных деревьев классов T_4 и T_5 , нам придется обратиться к производящим функциям.

Пусть a_n — количество корневых непомеченных деревьев объема n , а b_n — количество некорневых непомеченных деревьев объема n . Введем производящие функции этих чисел:

$$A(z) = \sum_{n=1}^{\infty} a_n z^n, \quad B(z) = \sum_{n=1}^{\infty} b_n z^n.$$

Известно, что производящая функция A удовлетворяет функциональному тождеству [120, 126]

$$A(z) = z \exp \left\{ \sum_{k=1}^{\infty} \frac{1}{k} A(z^k) \right\}.$$

Если некоторое время повозиться и вычислить $T'''(0)$, то мы получим число 96, так что $a_4 = 96/4! = 4$, что соответствует рис. 6.8.

Известно также, что

$$B(z) = A(Z) - \frac{1}{2}(A^2(z) - A(z^2)).$$

Очень часто в перечислительных задачах комбинаторики приходится вместо конкретных чисел в качестве точного ответа предъявлять их производящую функцию, а в качестве приближенного — асимптотику количества объектов или асимптотику доли этих объектов в некотором более широком классе. Знание производящей функции позволяет получать, например, рекурсии (которые мы уже видели на примере графов), что бывает полезно при компьютерном моделировании.

Соберем для наглядности оценки количества деревьев заданного объема пяти описанных видов в одну таблицу 6.3.

Таблица 6.3: Количество деревьев разных типов.

Тип	$n = 4$	Общий случай
Корневые помеченные	64	n^{n-1}
Некорневые помеченные	16	n^{n-2}
Рекурсивные	6	$(n-1)!$
Плоские посаженные	5	$\frac{1}{n} \binom{2n-2}{n-1}$
Корневые непомеченные	4	$A(z) = z \exp \left\{ \sum_{k=1}^{\infty} \frac{1}{k} A(z^k) \right\}$
Некорневые непомеченные	2	$B(z) = A(Z) - \frac{1}{2}(A^2(z) - A(z^2))$

Ранее мы отмечали, что лес — это граф, каждая компонента связности которого есть дерево. И это утверждение было дано для корневых и некорневых помеченных лесов. Мы можем пойти дальше и, обозначив за F_0 класс помеченных корневых лесов, с помощью факторизации или выделения подмножества переходить к F_1 — классу помеченных некорневых лесов, к F_2 — классу рекурсивных лесов, к F_3 — классу плоских посаженных лесов, к F_4 — классу корневых непомеченных лесов, к F_5 — классу некорневых непомеченных лесов. Однако, если мы хотим учитывать еще и порядок компонент

связности, удобнее всего в качестве F_0 брать сумму $\bigcup_N T_0^N$ прямых произведений класса деревьев с самим собой, и точно так же получать остальные классы лесов. Хотя вопрос о выборе формальной конструкции (теоретико-множественной модели) здесь уже не принципиален.

Куда интереснее вопрос о том, как вычислить количество лесов с N деревьями и n вершинами. И снова оказывается, что в случае помеченных лесов ответ получить проще, чем для случая непомеченных. Так, количество лесов объема n с N корневыми помеченными деревьями равно Nn^{n-N-1} , а количество плоских посаженных лесов объема n с N деревьями равно

$$\frac{N}{n} \binom{2n - N - 1}{n - N}.$$

Вывод этих формул можно найти, например, в [124]. С остальными классами ситуация хуже. Так, количество лесов, состоящих из N помеченных некорневых деревьев, равно

$$\frac{n!}{N!} \sum_{n_1 + \dots + n_N = n} \frac{n_1^{n_1-2} \dots n_N^{n_N-2}}{n_1! \dots n_N!}, \quad (6.4)$$

где суммирование ведется по всем наборам $n_1 + \dots + n_N = n$ с положительными n_k . Объяснить эту формулу довольно просто: набор из n вершин подвергается разбиениям на N подмножеств всеми возможными способами, которых $n!/n_1! \dots n_N!$, затем на каждом подмножестве вычисляется количество возможных деревьев, после чего все нормируется на $N!$ — число способов упорядочить деревья в лесе.

Аналогично, количество рекурсивных лесов с N занумерованными деревьями и n вершинами равно

$$\sum_{n_1 + \dots + n_N = n} \frac{n!}{n_1 \dots n_N},$$

где суммирование ведется по всем наборам $n_1 + \dots + n_N = n$ с положительными n_k .

Как видим, в случае F_1 и F_2 количество лесов уже не выражается простой арифметической формулой.

Для классов F_4 и F_5 , как и для деревьев того же класса, потребуется оперировать производящими функциями, чтобы получить количество лесов с N деревьями и n вершинами, т. е. выписать формулу в замкнутом виде не получится.

6.3 Вероятности на графах

Подобно тому, как любое число (как в узком арифметическом, так и в широком архетипическом смысле) может быть случайным, т. е. представлено некоторой измеримой функцией, определенной на вероятностном пространстве и принимающей значения в числовом множестве, так и граф может быть случайным. Именно, измеримая функция из вероятностного пространства $\langle \Omega, \mathcal{P} \rangle$ в некоторое множество графов, наделенное сигма-алгеброй, называется **случайным графом**.

Определим наше основное рабочее пространство следующим способом. Пусть G_1 есть множество всех связных простых орграфов, построенных на вершинах $1, \dots, n$, где n — любое целое положительное число. Рассмотрим множество

$$or\bar{\mathcal{G}} = \bigcup_N G_1^N,$$

т. е. включим в него наборы конечной длины, составленные из связных простых орграфов. Такая конструкция позволяет нам рассматривать граф с упорядоченными компонентами связности. Чтобы перейти к графам с неупорядоченными компонентами, нам достаточно факторизовать данное множество $or\bar{\mathcal{G}}$ по отношению эквивалентности, «забывающему» порядок компонент. Обозначим соответствующее фактор-множество $or\mathcal{G}$. Наконец, чтобы перейти к неориентированным графам, мы должны факторизовать $or\bar{\mathcal{G}}$ по отношению, забывающему ориентацию ребер, в результате чего получим множество $\bar{\mathcal{G}}$ всех неориентированных графов с упорядоченными компонентами. И далее, факторизация по отношению, «забывающему» порядок компонент, дает множество \mathcal{G} всех обычных неориентированных графов.

На этих множествах мы можем строить классы лесов F_0, \dots, F_5 как с упорядоченными деревьями, так и с неупорядоченными, корневыми и некорневыми, факторизуя по соответствующим признакам.

Кроме того, полезно выделить конечные подмножества $or\bar{\mathcal{G}}_n$ и $or\mathcal{G}_n$, включающие только орграфы с n вершинами, а также их фактор-множества по снятию ориентации $\bar{\mathcal{G}}_n$ и \mathcal{G}_n , включающие только неориентированные графы с n вершинами.

Заметим, что множество $or\bar{\mathcal{G}}$ счетно, а это означает, что мы можем рассматривать на нем сигма-алгебру, включающую все его подмножества, т. е. $\mathcal{P}(or\bar{\mathcal{G}})$, которая тривиально индуцируется во все фактор-множества, а значит, позволяет индуцировать в них и меру, заданную на исходных орграфах.

Если на $or\bar{\mathcal{G}}$ или на каком-то его фактор-множестве или его подмножестве задана вероятностная мера, то тогда функция id на этом подмножестве представляет собой случайный граф соответствующего типа.

Часто задать вероятность непосредственно сложнее для понимания, чем сначала определить какую-то меру, а затем нормировать ее, получая вероят-

ности на заданном множестве графов. Так, мы можем задать на множестве $or\mathcal{G}$ всех простых неориентированных графов с неупорядоченными компонентами единичную меру (т. е. мера каждого графа равна 1), после чего вероятность одного орграфа g в множестве $or\mathcal{G}_n$ может быть задана равномерным способом

$$P\{g\} = \frac{1}{\|or\mathcal{G}_n\|} = \frac{1}{2^{n(n-1)}}.$$

Для перехода к вероятностям на \mathcal{G}_n можно пойти разными путями. Во-первых, считать все графы из \mathcal{G}_n (т. е. классы эквивалентности над $or\mathcal{G}_n$ по снятию ориентации) равновероятными, и тогда вероятность одного графа будет равна $1/2^{n(n-1)/2}$. Это соответствует первому подходу, описанному на стр. 446.

Во-вторых, можно индуцировать вероятность через разбиения по формуле (5.1). В этом случае будем иметь

$$P\{[g]_{\sim}\} = \frac{\|\{g' \mid g' \sim g\}\|}{\|or\mathcal{G}_n\|} = \frac{3^k}{2^{n(n-1)}},$$

где $k = \|E(g)\|$ — количество ребер в графе g , а 3 отвечает за выбор количества ориентаций между двумя вершинами: $a \rightarrow b$, $b \rightarrow a$, $a \leftrightarrow b$. Несложно видеть, что число классов $[g]_{\sim}$ с условием $\|E(g)\| = k$ равно $\binom{n(n-1)/2}{k}$, так что сумма указанных вероятностей равна 1.

Наконец, можно не ограничиваться равномерной мерой или ее индуцированием в рассматриваемый класс, а задать меру каким-то способом, привязанным к изучаемым характеристикам графа, например, к числу его вершин, ребер, компонент связности, степеням вершин и т. п.

Сформулируем общую **задачу перечисления графов** в следующем виде. Пусть G — одно из множеств $or\mathcal{G}$, $or\mathcal{G}$, $\overline{\mathcal{G}}$, \mathcal{G} , и пусть задана функция

$$\varkappa : G \rightarrow \omega^\alpha,$$

$\alpha < \varepsilon_0$:) | где α — ненулевой ординал. Эту функцию будем называть **характеристикой графов** из G . Ординал α нам нужен лишь для того, чтобы иметь возможность задать одну или несколько счетных серий параметров графа в одном векторе. В простейшем случае $\alpha = 1$ и $\varkappa(g)$ представляет собой какой-то один числовой параметр графа g . Условимся считать, что $\varkappa(g) = (\varkappa_0(g), \varkappa_1(g), \dots)$, т. е. \varkappa_i есть i -ая координата вектора \varkappa .

Далее, пусть

$$W_G(k_0, k_1, \dots) \doteq \#\{g \in G \mid \varkappa(g) = (k_0, k_1, \dots)\},$$

| $\#X$ — то
же, что
 $\|X\|$:)

т. е. W_G — это количество графов из G , у которых фиксированы характеристики \varkappa_i .

Задача перечисления графов, таким образом, представляет собой поиск точных или приближенных (асимптотических или вероятностных) оценок для числа W при заданных значениях параметров $\{k_i\}_{i<\alpha}$.

В общем случае число W может оказаться бесконечным, поэтому обычно одним из параметров \varkappa_i является $\|V(g)\|$. Например, пусть $\varkappa(g) = (\|V(g)\|)$, тогда

$$W_{\mathcal{G}}(n) = 2^{n(n-1)/2}.$$

Для изучения чисел W_G удобно использовать их производящие функции, т. е. формальные степенные ряды

$$F_G(z_0, z_1, \dots) = \sum_{k_0, k_1, \dots \in \alpha} K(k_0, k_1, \dots) W_G(k_0, k_1, \dots) z_0^{k_0} z_1^{k_1} \dots,$$

где мы вновь используем соглашение о том, что $0^0 = 1$ для сочетания параметров $z_i = 0 = k_i$. Функция K , как обычно для производящих функций, обозначает ядро, которое выбирается в зависимости от задачи.

Указанную сумму можно интерпретировать как сумму мер всех графов из G (т. е. меру G), где мера одного графа g задается как

$$\mu(g) = K(\varkappa_0(g), \varkappa_1(g), \dots) z_0^{\varkappa_0(g)} z_1^{\varkappa_1(g)} \dots,$$

а переменные z_0, z_1, \dots являются параметрами этой меры (естественно, предполагается, что $z_i \in [0; +\infty)$).

Рассмотрим такую характеристику простых неориентированных графов:

$$\varkappa(g) = (\|V(g)\|, \|E(g)\|, C(g)),$$

где $C(g)$ — количество компонент связности g . Кроме того, выберем ядро $K(n, k, s) = 1/n!$. Тогда

$$W_{\mathcal{G}}(n, k, s) = \#\{g \in \mathcal{G} \mid \|V(g)\| = n, \|E(g)\| = k, C(g) = s\}, \quad (6.5)$$

$$\mu(g) = z_0^{\|V(g)\|} z_1^{\|E(g)\|} z_2^{C(g)} / \|V(g)\|!$$

и

$$F_{\mathcal{G}}(z_0, z_1, z_2) = \sum_{n, k, s} W_{\mathcal{G}}(n, k, s) z_0^n z_1^k z_2^s / n!. \quad (6.6)$$

Теперь с помощью этой меры определим вероятность на множестве \mathcal{G}_n как отношение меры одного графа к мере всех графов в \mathcal{G}_n , полагая при этом $z_2 = 1$ (т. е. мы считаем вероятность не зависящей от количества компонент связности):

$$P\{g\} = \frac{z_1^{\|E(g)\|} / n!}{\sum_{k=0}^{n(n-1)/2} \binom{n(n-1)/2}{k} z_1^k / n!} = \frac{z_1^{\|E(g)\|}}{(1 + z_1)^{n(n-1)/2}}, \quad (6.7)$$

где в знаменателе мы сгруппировали графы с одинаковым числом ребер k .

Интересно, что данная вероятностная модель имеет очень простую интерпретацию и является классической моделью случайного графа, которую мы коротко рассмотрим в следующем разделе.

6.3.1 Модель Эрдёша–Реньи

Итак, рассмотрим случайный неориентированный граф с вершинами $1, \dots, n$. Для того, чтобы предметно обсуждать его свойства, необходимо некоторым определенным образом задать его вероятности. Классической моделью здесь является так называемая модель Эрдёша–Реньи в интерпретации Э. Гильберта, а именно: случайный граф с независимыми неориентированными ребрами на n помеченных вершинах.⁷

Этот другой Гильберт :)

В этой модели случайный граф можно представить себе как схему Бернулли из $n(n-1)/2$ испытаний, в которой каждое из возможных (неориентированных) ребер графа возникает независимо от других ребер с одинаковой для всех ребер вероятностью $p \in (0; 1)$ (случай $p = 0$ пустого графа и $p = 1$ полного графа можно сразу же исключить).

Обозначим такой случайный граф $G(n, p)$. Легко видеть, что вероятностное распределение в данном случае определяется следующим образом:

$$P\{G(n, p) = g\} = p^{\|E(g)\|} q^{n(n-1)/2 - \|E(g)\|},$$

где $g \in \mathcal{G}_n$ и $q = 1 - p$.

Что любят формулы? — Чтобы на них смотрели!

А теперь посмотрим на формулу (6.7) и заметим, что если вместо положительного параметра z_1 подставить дробь p/q (которая тоже может быть любым положительным числом), то мы в точности получим вероятность $P\{G(n, p) = g\}$! Таким образом, наш длинный рассказ об индуцированных мерах на графах обретает классические очертания на вполне конкретной и хорошо изученной модели.

Случайный граф $G(n, p)$ был впервые описан в работе Э. Гильберта [119], часть результатов были получены Эрдёшем и Реньи в [117, 118], позже подробно изучался Б. Боллобасом [116], В. Ф. Колчиным [122] и другими исследователями.

Надо заметить, что граф как таковой фигурирует в вероятностной модели только в момент определения случайного графа. В дальнейшем мы просто строим некоторые числовые характеристики графа, которые после определения вероятностного распределения на графах превращаются в обычные случайные величины. Получив те или иные результаты об этих характеристиках,

⁷Справедливости ради нужно отметить, что модель именно в таком виде была предложена Э. Гильбертом в 1959 году, а в том же году Эрдёш и Реньи в совместной статье ввели понятие случайного графа $G(n, M)$ с n вершинами и M ребрами, предполагая все реализации такого графа равновероятными.

мы возвращаемся назад к графам и производим некоторую интерпретацию полученных фактов, применительно уже к самим графам.

Тем не менее, все не так просто, как может показаться. Дело в том, что далеко не всегда числовые характеристики графа можно выразить явным образом через его исходные числовые же параметры (количество вершин, ребер, компонент связности). Наше знание об устройстве графа проявляется, например, в том случае, если нам требуется получить характеристическую функцию изучаемой величины или ее производных величин.

Ниже мы изложим нестандартный подход к изучению графа $G(n, p)$ на основе введенных выше производящих функций, а затем дадим ряд хорошо известных его свойств без доказательства.

Для случайного графа $G(n, p)$ обозначим через ν_n количество компонент связности,⁸ тогда

$$P\{\nu_n = s\} = \sum_{k=0}^{\infty} A_{n,k,s} (p/q)^k q^{n(n-1)/2}, \quad (6.8)$$

где $A_{n,k,s} = W_{\mathcal{G}}(n, k, s)$ из формулы (6.5) — количество графов из \mathcal{G}_n , содержащих k ребер и s компонент связности. Для крайних значений индексов положим по определению $A_{0,0,0} = 1$ и $A_{0,k,s} = A_{n,k,0} = 0$. Кроме того, обозначим $B_{n,k} \Leftarrow A_{n,k,1}$ количество связных графов с k ребрами, $A_{n,k} = \sum_s A_{n,k,s}$ количество всех графов с k ребрами.

Одним из ключевых соотношений, которые необходимы для работы с графами, является равенство следующего вида:

$$A_{n,k,s} = \sum_{\substack{n_1+\dots+n_s=n \\ k_1+\dots+k_s=k}} \frac{n!}{s!} \frac{B_{n_1,k_1} \cdots B_{n_s,k_s}}{n_1! \cdots n_s!}, \quad (6.9)$$

где суммирование ведется по всем $n_i, k_i \geq 0$, удовлетворяющим указанным условиям. В последней формуле можно угадать формулу, похожую на (6.4) (нужно заменить $s!$ на $N!$). И в этом нет ничего удивительного, т. к. и лес, и произвольный граф складывается из своих компонент связности, а факториальные коэффициенты отвечают за наличие/отсутствие нумерации вершин в графе или его компонентах, а также нумерации самих компонент связности.

*Упражнение 6.7.
Нужно перейти к графам с занумерованными компонентами, т. е. к \mathcal{G}_n .*

Предлагаем читателю вывести соотношение (6.9) самостоятельно. При работе с другими числовыми характеристиками графов возникают похожие равенства, позволяющие выразить эти характеристики как свертку аналогичных характеристик для связного графа. Характеристики графа с таким

⁸Мы опускаем в обозначении непрерывный параметр p , т. к. в дальнейшем считаем его константой или функцией от n .

свойством называют *разложимыми* [122]. Например, количество всех графов $A_{n,s} = \sum_k A_{n,k,s}$ также удовлетворяет аналогичному равенству, количество лесов различного вида — тоже.

Основная цель разложения (6.9) состоит в том, чтобы связать производящие функции для графов и связных графов. Пусть

$$A(x, y) \rightleftharpoons \sum_{n,k} \frac{A_{n,k}}{n!} x^n k y^k = F_{\mathcal{G}}(x, y, 1), \quad B(x, y) \rightleftharpoons \sum_{n,k} \frac{B_{n,k}}{n!} x^n k y^k = [z]F_{\mathcal{G}}(x, y, z).$$

где $F_{\mathcal{G}}$ — производящая функция из формулы (6.6), определяющая вероятности, $[z]F_{\mathcal{G}}(x, y, z)$ — коэффициент перед первой степенью z в формальном ряде $F_{\mathcal{G}}(x, y, z)$.

Из (6.9) непосредственно вытекает тождество

$$A(x, y) = e^{B(x, y)}, \quad (6.10)$$

в котором сходимость ряда не обязательна, тождество лишь означает совпадение коэффициентов при степенях $x^n y^k$. Тем не менее, равенство формальных степенных рядов сохраняет такие операции, как сложение и умножение рядов, а также дифференцирование и интегрирование. Что позволяет сравнительно небольшими усилиями выводить тождества для коэффициентов.

Далее мы зафиксируем переменную $y = p/q$ (выше мы делали то же самое, полагая $z_1 = p/q$ для перехода к модели Эрдёша–Реньи) и введем упрощенные обозначения $A \rightleftharpoons A(x, p/q)$ и $B \rightleftharpoons B(x, p/q)$. В то же время, из (6.8) и (6.9) легко получить, что

$$\sum_{n=0}^{\infty} \frac{\mathbf{P}\{\nu_n = s\}}{q^{n(n-1)/2}} \frac{x^n}{n!} = \frac{1}{s!} B^s,$$

Из равенства (6.10) видим, что $A = e^B$, так что

$$A = \sum_{n=0}^{\infty} \frac{x^n}{q^{n(n-1)/2} n!}$$

Дифференцируя равенство $A = e^B$ по x , получим

$$A'x = AB'x,$$

откуда, пользуясь формулой свертки для производящих функций (см. свойство 5 в разделе 5.2), нетрудно найти, что

$$n = \sum_{k=0}^n \binom{n}{k} q^{k(n-k)} k \mathbf{P}\{\nu_k = 1\},$$

откуда получается известная рекурсивная формула для вероятности связности графа $G(n, p)$:

$$P\{\nu_n = 1\} = 1 - \sum_{k=1}^{n-1} \frac{k}{n} \binom{n}{k} q^{k(n-k)} P\{\nu_k = 1\}.$$

Из этой формулы, кстати, выводится рекурсивная формула (6.2) количества связных графов порядка n . Для этого достаточно сделать все графы равновероятными, полагая $p = q = 1/2$, и заметить, что $P\{\nu_n = 1\} = c_n/2^{\binom{n}{2}}$.

Упражнение 6.8. Выведите формулу (6.2).

Мы видим, что A и B есть производящие функции последовательностей с общим ядром $K = 1/(q^{n(n-1)}n!)$, причем A соответствует тождественной единице, а B преобразует последовательность вероятностей связности $\{P\{\nu_n = 1\}\}$. Более того, ряд $B^s/s!$ соответствует последовательности $\{P\{\nu_n = s\}\}$, а ряд AB^s соответствует последовательности факториальных моментов количества компонент связности:

$$E\nu_n^s = \sum_k k^s P\{\nu_n = k\}.$$

В частности, ряд $E = AB$ соответствует последовательности математических ожиданий $\{E\nu_n\}$.

В статье [121] показано несколько соотношений, связанных с такими рядами, а также следствия, которые можно из них извлечь, пользуясь разрешенными операциями над формальными рядами. Например:

$$\begin{aligned} (n-1)^s \cdot q^{(n-1)s} &\leq E(\nu_n - 1)^s \leq 2(n-1)^s q^{(n-1)(s+1)/2}, \\ 1 - 2nq^{n-1} &\leq p_n \leq \frac{1}{nq^n}, \\ 1 - \frac{1}{nq^n} &\leq p_{i_n} \leq nq^{n-1}, \end{aligned} \tag{6.11}$$

где $p_n = P\{\nu_n = 1\}$ — вероятность связности, а p_{i_n} есть вероятность того, что $G(n, p)$ имеет изолированную вершину.

Анализ производящих функций позволяет получать различные соотношения, дающие упрощенные асимптотические оценки. Так, в этой же работе получено, например, что

$$P\{\nu_n = k\} \rightarrow \frac{\alpha^{k-1}}{(k-1)!} e^{-\alpha} \tag{6.12}$$

при $n \rightarrow \infty$, если параметр $q = 1 - p$ зависит от n так, что $nq^n \rightarrow \alpha > 0$ при $n \rightarrow \infty$. Иначе говоря, количество компонент связности графа $G(n, p)$ при

указанных условиях слабо сходится к распределению Пуассона с параметром α . Это же соотношение остается верным и при $\alpha = 0$, когда распределение ν_n асимптотически сходится в одну точку $k = 1$ (граф асимптотически почти наверное связан).

Если же $p \rightarrow 0$, а $n = \text{const}$, то мы имеем следующий предельный переход:

$$P\{\nu_n = k\} = f_{k,n} p^{n-k} + O(p^{n-k+1}), \quad (6.13)$$

где $f_{k,n}$ — количество помеченных лесов объема n с k неупорядоченными некорневыми деревьями.

Связь с такими деревьями в модели Эрдёша–Реньи, на самом деле, более глубокая. Обозначим

$$\beta(x) = \sum_{k=1}^{\infty} \frac{k^{k-2}}{k!} x^k$$

факториальную производящую функцию количества некорневых помеченных деревьев.

Тогда, как показано в [121], при $q^n \rightarrow e^{-\alpha}$ ($\alpha \geq 0$) имеем соотношение для факториальных моментов:

$$E(\nu_n - 1)^s = \left(\frac{n}{\alpha} \beta(\alpha e^{-\alpha}) \right)^s (1 + o(1)), \quad (6.14)$$

т. е. и тут мы неявно используем количество деревьев класса T_1 .

Случайный граф $G(n, p)$ обладает также следующими свойствами.

Gnp1 Математическое ожидание количества изолированных вершин равно nq^{n-1} . Отсюда, в частности, следует правая оценка в (6.11).

Gnp2 Пусть $p = (c \ln n)/n$ и $c > 1$, тогда $p_n \rightarrow 1$ (граф *асимптотически почти наверное связан*). При таком p имеем $nq^n \rightarrow 0$ и попадаем в зону действия формулы (6.12).

Gnp3 Пусть $p = (c \ln n)/n$ и $c < 1$, тогда $p_n \rightarrow 0$ (граф *асимптотически почти наверное несвязен*). При таком p имеем $nq^n \rightarrow \infty$ и из формулы (6.14) видим, что моменты числа компонент связности тоже уходят на бесконечность, т. е. граф получается сильно раздробленным.

Gnp4 Пусть $p = (\gamma + \ln n)/n$ ($\gamma \in \mathbb{R}$), тогда $p_n \rightarrow e^{-e^{-\gamma}}$. Это соответствует формуле (6.12) при $\alpha = e^{-\gamma}$.

Gnp5 Пусть $np \rightarrow c > 1$, тогда максимальная компонента связности имеет порядок n , а вторая по величине компонента (и все прочие) имеет порядок $O(\ln n)$. Этот эффект в случайных графах и схемах размещения называется возникновением **гигантской компоненты** (т. е. такой компоненты, что ее порядок есть доля порядка самого графа, а остальные компоненты бесконечно малы в сравнении с ним).

Gnp6 Пусть $np = 1$, тогда максимальная компонента графа имеет размер порядка $n^{2/3}$.

Gnp7 Пусть $np \rightarrow c < 1$, тогда максимальная компонента (и все остальные) имеет порядок $O(\ln n)$.

Gnp8 Пусть $p = o(1/n)$, тогда все компоненты графа в пределе становятся деревьями. Выше мы отмечали это обстоятельство — см. формулу (6.13).

Из Gnp2–4 видим, что отношение $(\ln n)/n$ является критическим значением для вероятности возникновения ребра в графе $G(n, p)$. Если вероятность p строго больше этого значения, то граф почти наверняка будет связан, если меньше или равно — почти наверняка будет несвязен. Связность при этом хорошо коррелирует с асимптотикой величины nq^{n-1} (среднего числа изолированных вершин), поскольку при $c > 1$ имеем $nq^{n-1} = 1/n^{c-1+o(1)}$, при $c < 1$ имеем $nq^{n-1} = n^{c-1+o(1)}$, а при $c = 1$ имеем $nq^{n-1} = e^{-\gamma}(1 + o(1))$.

Из Gnp5–8 видим, что при преодолении вероятностью p порога $1/n$ сверху вниз происходит качественный скачок в распределении вершин графа по компонентам связности. При $p > 1/n$ в графе возникает гигантская компонента. И хотя граф при этом несвязен, одна его компонента пытается компенсировать этот недостаток, захватывая почти весь граф (так иногда происходит с крупными городами–агломерациями). При снижении p стремление к гигантизму уменьшается, компоненты становятся более равновеликими, хотя и теряют в весе. Ну а при совсем уж ничтожной вероятности ребра граф «высушивается» до состояния леса.

В вероятностной теории графов асимптотическое изменение структуры графа при изменении соотношений его параметров называют **эволюцией случайного графа**. Здесь параметр p можно рассматривать как временной параметр, а точки $1/n$ и $(\ln n)/n$ — точками бифуркации системы.

Естественно, все эти результаты имеют строгую математическую формулировку и обычно означают сходимость по вероятности рассматриваемых случайных величин. Для более полного погружения в тему случайных графов рекомендуем книги [115, 122, 124], а также краткий видеокурс д.ф.-м.н. А. Райгородского «Случайные графы» [ММУ9bKMxmkE].

6.3.2 Случайные деревья и леса

В этом разделе, как и в предыдущем, мы ограничимся только одним примером. Тем не менее, он дает представление об общей методике работы со случайными графами, построенными на деревьях заданного объема.

Просто генерируемые деревья

Рассмотрим класс T_3 плоских посаженных деревьев. Этот класс обладает уникальным свойством: каждое его дерево t может быть взаимно однозначно представлено набором деревьев (t_1, \dots, t_k) того же класса T_3 , где k — это полустепень исхода корня дерева t . Действительно, если вспомнить способ нумерации вершин плоского посаженного дерева, описанный на стр. 554, то легко видеть, что вершину с пометкой $(0, d_1, d_2, \dots)$ нужно отнести в дерево t_{d_1+1} , где она получит номер $(0, d_2, d_3, \dots)$. И обратно, вершину с меткой $(0, d_1, d_2, \dots)$ из дерева t_j нужно пометить как $(0, j-1, d_1, d_2, \dots)$ в новом дереве t , которое получается присоединением к корню (0) корней деревьев t_1, \dots, t_k . При $k=0$ набор (t_1, \dots, t_k) пуст.

Следуя общей схеме, определим характеристику \varkappa для таких деревьев:

$$\varkappa_k(t) \Rightarrow \#\{v \in V(t) \mid \deg^+(v) = k\}, \quad k = 0, 1, 2, \dots$$

т. е. количество вершин дерева t , имеющих ровно k дочерних. В частности, $\varkappa_0(t)$ — это количество листьев.

Выбирая единичное ядро, получаем формулу для меры одного дерева:

$$\mu(t) \Rightarrow z_0^{\varkappa_0(t)} z_1^{\varkappa_1(t)} \dots \quad (6.15)$$

Последовательность неотрицательных чисел $\{z_k\}$ будем называть *генерирующей меру последовательностью*.

Архетип порождающего элемента!

Отметим, что для конечного дерева данное произведение конечно, т. к. в дереве объема n не может быть полустепеней исхода больше $n-1$, поэтому это произведение обрывается максимум на переменной $z_{\|V(t)\|-1}$ (далее просто идут единицы).

Пусть дереву $t \in T_3$ указанным выше способом соответствует кортеж деревьев (t_1, \dots, t_k) , тогда, как легко видеть,

$$\mu(t) = z_0^{\varkappa_0(t_1)+\dots+\varkappa_0(t_k)} \dots z_k^{1+\varkappa_k(t_1)+\dots+\varkappa_k(t_k)} \dots = z_k \mu(t_1) \dots \mu(t_k).$$

Пусть также

$$t_n \Rightarrow \sum_{t \in T_3, \|V(t)\|=n} \mu(t),$$

т. е. суммарная мера всех деревьев объема n .

Введем следующие производящие функции:

$$Z(x) \Rightarrow \sum_{k=0}^{\infty} z_k x^k, \quad T(x) \Rightarrow \sum_{n=1}^{\infty} t_n x^n.$$

Пользуясь предыдущим равенством, получаем, что

$$\begin{aligned}
 T(x) &= \sum_{n=1}^{\infty} \sum_{t \in T_3, \|V(t)\|=n} \mu(t)x^n = \sum_{n=1}^{\infty} \sum_{k=0}^n \sum_{\substack{t \in T_3, \|V(t)\|=n \\ \deg(\text{root})=k}} \mu(t)x^n = \\
 &= x \sum_{k=0}^{\infty} z_k \sum_{n>k} \sum_{n_1+\dots+n_k=n-1} \sum_{\substack{t_1, \dots, t_k \in T_3 \\ \|V(t_1)\|=n_1, \dots, \|V(t_k)\|=n_k}} \mu(t_1)x^{n_1} \dots \mu(t_k)x^{n_k} = \\
 &= x \sum_{k=0}^{\infty} z_k T^k(x) = xZ(T(x))
 \end{aligned} \tag{6.16}$$

Итак, мы получили функциональное уравнение

$$T = xZ(T), \tag{6.17}$$

которое определяет просто генерируемые деревья. Точнее, класс деревьев, для которых существует последовательность $\{z_k\}$ с производящей функцией $Z(x)$, порождающая меру $\mu(t)$ по формуле (6.15), и такая, что производящая функция $T(x)$ последовательности мер множеств деревьев заданного объема удовлетворяет тождеству (6.17), называется классом **просто генерируемых деревьев** [124]. Как видим, плоские посаженные деревья образуют класс просто генерируемых деревьев.

В частности, при $z_k = 1$ для всех k мы получаем, что числа t_n отвечают за количество плоских посаженных деревьев объема n , т. е. совпадают с числами Каталана, соответственно, $C_{n-1} = \frac{1}{n} \binom{2n-2}{n-1}$. Отсюда можно получить функциональное уравнение для производящей функции чисел Каталана

$$C(x) = \sum_{n=0}^{\infty} C_n x^n = \sum_{n=1}^{\infty} \frac{1}{n} \binom{2n-2}{n-1} x^{n-1} = T(x)/x = Z(xC(x)),$$

где

$$Z(x) = \sum_{k=0}^{\infty} x^k = \frac{1}{1-x},$$

так что

$$C(x) = \frac{1}{1-xC(x)}, \text{ или } C(x) = xC(x)^2 + 1, \quad C(x) = \frac{1 - \sqrt{1-4x}}{2x}.$$

Это равенство выводится независимым способом в книге [2] в рамках другой комбинаторной задачи.

Варьируя выбор z_k , можно переходить к некоторым подклассам класса T_3 , которые будут также просто генерируемыми. Например, мы можем задать $z_0 = z_2 = 1$, а остальные $z_k = 0$. Иначе говоря, мы допускаем только

бинарные плоские посаженные деревья (мера остальных деревьев равна нулю). В этом случае $Z(x) = 1 + x^2$ и $T(x) = x(1 + T(x)^2)$. Решая это уравнение и раскладывая $T(x)$ по степеням x , можно найти количество бинарных плоских посаженных деревьев объема n .

Наконец, переход к вероятностной модели осуществляется подходящей нормировкой. Так, если мы рассматриваем множество деревьев

$$\mathcal{T}_n \rightleftharpoons \{t \in T_3 \mid \|V(t)\| = n\},$$

то вероятность на нем задается формулой

$$\mathbf{P}\{t\} = \mu(t)/t_n.$$

Множество плоских посаженных лесов с N пронумерованными деревьями определяется как прямое произведение T_3^N пространства деревьев на себя N раз. Точно так же задается и мера на этом пространстве:

$$\mu(f) \rightleftharpoons \mu(t_1) \cdots \mu(t_N), \quad f = (t_1, \dots, t_N) \in T_3^N.$$

Легко видеть, что эта мера может быть определена так же, как на деревьях:

$$\mu(f) = z_0^{\kappa_0(f)} z_1^{\kappa_1(f)} \dots$$

Аналогично деревьям, обозначим суммарную меру лесов с N деревьями и n вершинами

$$f_{N,n} \rightleftharpoons \sum_{f \in T_3^N, \|V(f)\|=n} \mu(f),$$

и введем производящую функцию этих мер:

$$F_N(x) \rightleftharpoons \sum_{n=1}^{\infty} f_{N,n} x^n = \sum_{n=1}^{\infty} \sum_{n_1 + \dots + n_N = n} t_{n_1} \cdots t_{n_N} x^n = T(x)^N.$$

Вероятностное пространство для таких деревьев строится нормировкой меры на $f_{N,n}$. Точнее, на множестве деревьев

$$\mathcal{F}_{N,n} \rightleftharpoons \{f \in T_3^N \mid \|V(f)\| = n\}$$

вероятность задается формулой

$$\mathbf{P}\{f\} = \mu(f)/f_{N,n}. \quad (6.18)$$

Ясно, что вероятности на деревьях есть частный случай данных вероятностей при $N = 1$.

Произведем теперь еще ряд манипуляций, которые нам пригодятся в дальнейшем. Последовательность параметров $\{z_k\}$, определяющая меру μ , выбрана достаточно произвольно (кроме условия $z_k \geq 0$ мы ничего не требовали). Ясно, что работать при минимальных ограничениях с этими коэффициентами неудобно — мы даже не можем гарантировать конечное решение уравнения (6.17).

Поэтому всюду далее будем предполагать, что ряд $Z(x)$ сходится при некотором $x = x_0 > 0$. После чего нормализуем коэффициенты следующим образом. Положим

$$p_k(\lambda) \Rightarrow \frac{z_k \lambda^k}{Z(\lambda)}, \quad \lambda \in (0; x_0). \quad (6.19)$$

Ясно, что $p_k(\lambda)$ представляют собой вероятности. Их можно интерпретировать как вероятности полустепеней исхода в случайном дереве/лесе из рассматриваемого семейства. Посмотрим, что произойдет с мерой μ , если мы заменим последовательность $\{z_k\}$ на $\{p_k(\lambda)\}$. Новую меру обозначим

$$\mu(f, \lambda) \Rightarrow p_0(\lambda)^{x_0(f)} p_1(\lambda)^{x_1(f)} \dots = z_0^{x_0(f)} z_1^{x_1(f)} \dots \frac{\lambda^{n-N}}{Z(\lambda)^n} = \mu(f) \frac{\lambda^{n-N}}{Z(\lambda)^n}. \quad (6.20)$$

Как видим, отличие новой меры от старой состоит лишь в коэффициенте, зависящем от параметров N, n . А этот коэффициент сократится при нормировке вероятностей на множестве $\mathcal{F}_{N,n}$!

Следовательно, заданные выше вероятности мы можем изучать и с помощью искусственно заданного распределения $\{p_k(\lambda)\}$ вместо достаточно произвольной генерирующей последовательности $\{z_k\}$. Для этого определим

$$f_{N,n}(\lambda) \Rightarrow \sum_{f \in \mathcal{F}_{N,n}} \mu(f, \lambda) = f_{N,n} \lambda^{n-N} / Z(\lambda)^n,$$

тогда

$$P\{f\} = \frac{\mu(f, \lambda)}{f_{N,n}(\lambda)}. \quad (6.21)$$

Такой переход к параметрическим вероятностям удобен, если нам потребуется выбрать параметр λ каким-то специальным способом для получения предельных теорем. Чуть ниже мы увидим это.

Отметим также, что производящие функции новых коэффициентов и мер будут по-прежнему связаны тождеством. Положим

$$P(x, \lambda) \Rightarrow \sum_{k=0}^{\infty} p_k(\lambda) x^k = Z(x\lambda) / Z(\lambda),$$

$$F_N(x, \lambda) \Rightarrow \sum_{n=1}^{\infty} \sum_{f \in \mathcal{F}_{N,n}} \mu(f, \lambda) x^k = \lambda^{-N} F_N(x\lambda / Z(\lambda)) = \lambda^{-N} T(x\lambda / Z(\lambda))^N.$$

Соответственно, уравнение

$$T(x, \lambda) \Rightarrow F_1(x, \lambda) = xP(T(x, \lambda), \lambda) \quad (6.22)$$

полностью соответствует уравнению (6.17), в котором исходные функции $T(x)$ и $Z(x)$ заменены на параметрические $T(x, \lambda)$ и $P(x, \lambda)$.

Таким образом, вероятностное пространство $\langle \mathcal{F}_{N,n}, \mathbf{P} \rangle$ с вероятностью, определенной по формуле (6.18) с генерирующей меру последовательностью $\{z_k\}$ совпадает с вероятностным пространством, в котором вероятность определена по формуле (6.21) с генерирующей меру последовательностью $\{p_k(\lambda)\}$, определенной формулой (6.19).

В частности, вводя на множестве $\mathcal{F}_{N,n}$ равномерное распределение (с помощью $z_k = 1$) и рассматривая, таким образом, случайный плоский посаженный

Архетип редукции! лес с N деревьями и n вершинами, где вероятность всех реализаций одинакова, мы можем без потерь перейти к модели, где генерирующие вероятности составляют геометрическое распределение

$$p_k(\lambda) = \lambda^k(1 - \lambda),$$

причем параметр $\lambda \in (0; 1)$ может быть выбран как угодно, например, в зависимости от чисел N, n . Пространство $\langle \mathcal{F}_{N,n}, \mathbf{P} \rangle$ от этого не изменится.

Запомним этот замечательный факт и перейдем к рассмотрению некоторых результатов об этом пространстве.

Леса Гальтона–Ватсона

Известно, что плоские посаженные деревья можно рассматривать как реализации **ветвящегося процесса Гальтона–Ватсона**, стартующего с 1 частицы. Процесс Гальтона–Ватсона — это такой (случайный) ветвящийся процесс с дискретным временем, в котором на каждом шаге каждая из имеющихся частиц независимо от других порождает k пронумерованных новых частиц с вероятностью p_k ($k = 0, 1, 2, \dots$), после чего исчезает, т.е. частицы предыдущего поколения заменяются порожденными частицами. Историю процесса можно записывать таким же кодом, каким мы размечали плоские посаженные деревья, поэтому каждая конкретная реализация такого процесса есть элемент T_3 .

Название процесса Гальтона–Ватсона происходит от имен исследователей-социологов, которые изучали генеалогические деревья. С известной долей допущения математический процесс Гальтона–Ватсона может описывать развитие генеалогического дерева (если наблюдать только за одними полом — мужским или женским). При этом считается, что распределение $\{p_k\}$ вероятностей числа потомков остается неизменным в течение всей истории. Другая

возможная интерпретация данного процесса — это ядерная реакция, при которой с течением времени частицы распадаются, образуя новые, а переход реакции к взрыву или затуханию зависит от распределения числа потомков.

Вероятность того, что процесс пройдет по траектории $t \in T_3$ в точности равна мере $\mu(t)$ из формулы (6.15), если в качестве генерирующей последовательности $\{z_k\}$ взять распределение $\{p_k\}$ числа потомков частицы в этом процессе. Таким образом, изучение случайного процесса Гальтона–Ватсона можно свести к изучению случайного плоского посаженного дерева.

Пусть $\langle \mathcal{F}_{N,n}, \mathbf{P} \rangle$ — вероятностное пространство, определенное как в предыдущем разделе, с генерирующей последовательностью

$$p_k(\lambda) = \frac{p_k \lambda^k}{F(\lambda)}, \quad (6.23)$$

где $F(\lambda) = \sum_k p_k \lambda^k$ (этот ряд сходится, т. к. p_k — вероятности). Тогда тождественную функцию $GW_{N,n} = \text{id}$ на этом пространстве назовем **случайным лесом Гальтона–Ватсона**.

Такой случайный лес будет моделировать процесс Гальтона–Ватсона с распределением потомков $\{p_k\}$, но с одним ограничением — точным знанием числа частиц, родившихся в нем за всю его историю. Иначе говоря, мы изучаем случайный процесс в условных вероятностях: условием является равенство количества частиц параметру n . При этом мы не фиксируем ни n , ни N , а рассматриваем модель при различном соотношении этих параметров, когда хотя бы один из них стремится к бесконечности. Тем самым, задача изучения леса Гальтона–Ватсона напоминает задачу о графе с независимыми ребрами $G(n, p)$ с двумя переменными параметрами.

Так же, как при изучении случайного графа $G(n, p)$ мы видели, что при переходе p через критическое значение $(\ln n)/n$ резко меняется структура графа, для леса Гальтона–Ватсона существуют свои пороговые значения, изменяющие расклад в соотношении объемов деревьев этого леса.

Для того, чтобы привести основные свойства случайного леса $GW_{N,n}$, нам потребуется еще ряд обозначений и ограничений.

Обозначений больше, чем теорем!

Во-первых, предполагается, что $p_0 > 0$, распределение $\{p_k\}$ невырожденное и его носитель имеет шаг 1.⁹ Если шаг распределения больше 1, то мы всегда можем свести задачу к такой, где шаг будет равен 1 (об этом см. [112]).

Во-вторых, требуется, чтобы второй момент этого распределения был конечен: $F''(1) < \infty$.

⁹Носитель распределения — это множество точек ненулевой вероятности, а шаг 1 означает, что носитель не содержится ни в какой решетке вида $d\mathbb{Z} + s$ при целом $d > 1$ и любом целом s .

Пусть $\eta_{(1)} \leq \dots \leq \eta_{(N)}$ — вариационный ряд случайных величин, равных объемам деревьев леса $GW_{N,n}$, т. е. $\eta_{(N)}$ — объем самого большого леса, $\eta_{(N-1)}$ — следующего по величине и т.д. (некоторые могут совпадать).

Все утверждения об этих случайных величинах асимптотические, поэтому слова «сосредоточен на», «равен» и т.п. следует читать с припиской «с вероятностью, стремящейся к 1», т. е. с ростом параметров случайного леса отклонение от данных соотношений ничтожно маловероятно. А слова «имеет порядок» означают, что случайная величина некоторым образом размазана вокруг данного порядка (например, бернуллиевская сумма успехов имеет порядок np , где n — число испытаний в схеме Бернулли, а p — вероятность успеха в одном испытании). К сожалению, точные формулировки заняли бы у нас несколько страниц убористого текста формул, что пагубно скажется на высказывании центральной идеи. И в этом случае мы бы уже с трудом различали «за деревьями лес». Поэтому следующие утверждения мы не называем теоремами:

GWF1 Пусть $N \rightarrow \infty$ и $n - N = \text{const}$. Тогда существует конечное множество натуральных чисел $y_1 \leq \dots \leq y_k$ таких, что несколько наибольших компонент $\eta_{(N)}, \eta_{(N-1)}, \dots$ сосредоточены на этих числах, а все остальные компоненты (асимптотически почти наверное) равны 1 (т. е. содержат только корень). Слово «несколько» здесь заменяет строгое ограничение на число l в номерах компонент $\eta_{(N-l)}$.

GWF2 Пусть $n - N \rightarrow \infty$ так, что $(n - N)/N \rightarrow 0$. Тогда существуют две последовательности $\{r_{N,n}\}$ и $\{s_{N,n}\}$ порядка не выше $\ln N$ такие, что при каждом фиксированном l величина $\eta_{(N-l)}$ сосредоточена на множестве $\{r_{N,n}, s_{N,n}\}$ с вероятностями, зависящими от l .

GWF3 Пусть $N, n \rightarrow \infty$ так, что $(n - N)/N \asymp 1$.¹⁰ Тогда существует последовательность $\{r_{N,n}\}$, стремящаяся к ∞ слабее N и такая, что при любом фиксированном l величина $\eta_{(N-l)}$ имеет порядок $r_{N,n}$.

GWF4 Пусть $N, n \rightarrow \infty$ так, что $(n - N)/N \rightarrow \infty$ и $n/N^2 \rightarrow 0$. Тогда при любом фиксированном l величина $\eta_{(N-l)}$ имеет порядок $(n/N)^2$.

GWF5 Пусть $N, n \rightarrow \infty$ так, что $n/N^2 \asymp 1$. Тогда при любом фиксированном l величина $\eta_{(N-l)}$ имеет порядок n . Точнее, старшие компоненты распределяются вокруг некоторых долей числа n .

GWF6 Пусть $N, n \rightarrow \infty$ так, что $n/N^2 \rightarrow \infty$. Тогда $\eta_{(N)}$ асимптотически равна n (отклонения от n бесконечно малы в сравнении с n), а следующие по величине компоненты $\eta_{(N-l)}$ при фиксированном l имеют порядок N^2 .

¹⁰ $a \asymp b$ означает $a = O(b)$ и $b = O(a)$.

Нужно отметить, что указанное в GWF1 множество $\{y_1 \leq \dots \leq y_k\}$ определяется арифметической структурой носителя распределения $\{p_k\}$. В этом множестве каждый следующий элемент нельзя представить как линейную комбинацию всех предыдущих. Распределение числа некорневых вершин по этим числам соответствует решению задачи о монетах.¹¹

Можно заметить, что эволюция леса $GW_{N,n}$ определяется соотношением между числом некорневых вершин $n - N$ и количеством деревьев. Сначала, когда $n - N = \text{const}$, самые большие деревья имеют объемы (без учета корневой вершины), доставляющие разложение числа $n - N$ по «монетам» $y_1 \leq \dots \leq y_k$ (т.е. коэффициенты этого разложения), а остальные деревья состоят из одного лишь корня.

Далее, когда число некорневых вершин начинает расти, но остается сильно слабее количества деревьев, старшие по объему деревья «забывают» об арифметической структуре носителя $\{p_k\}$, начинают расти до величин порядка $\ln N$ (более точно порядок можно определить только зная распределение числа потомков) и скапливаются на двух точках. При этом с ростом номера l происходит «переползание» распределения $\eta_{(N-l)}$ с верхней точки на нижнюю.

*Насколько
сильно
«сильно
слабее»
слабее, чем
«слабо
сильнее»?*

В третьей зоне количество некорневых вершин примерно соответствует количеству деревьев. В этом случае старшие деревья имеют объем все еще сильно меньше N , однако у них появляется разброс значений вокруг некоторого центрального $r_{N,n}$. Распределение старших компонент по числам $r_{N,n} \pm k$ нельзя назвать похожим на нормальное, но некоторая аналогия с предельным поведением биномиального распределения прослеживается — есть уходящий на бесконечность центр, вокруг которого скапливается облако вероятностей.

В первых трех зонах изменения параметров N, n можно назвать лес «широким», т.к. количество деревьев сильно превосходит объем любого дерева.

Четвертую зону можно назвать переходной, т.к. в ней количество вершин начинает превалировать над количеством деревьев, хоть и недостаточно сильно. Здесь лес перестает быть «широким», объем максимального дерева растет быстрее числа деревьев, но медленнее общего объема леса.

В пятой зоне начинают формироваться компоненты порядка n , т.е. сравнимые с объемом всего леса. Пока их несколько, поэтому нельзя говорить об эффекте гигантской компоненты в графе, но лес уже можно считать противоположным «широкому» — «компактным», т.к. он концентрируется на небольшом количестве деревьев.

Наконец, когда количество вершин дерева сильно превышает квадрат количества деревьев, появляется дерево, которое асимптотически содержит весь

¹¹Знаменитая Coin Problem: представить произвольную сумму монетами из заданного набора. Начиная с некоторого значения любую сумму можно разложить по заданным монетам, но варианты разложения имеют различную вероятность в соответствии с вероятностями монет, т.е. нашими p_k .

объем леса. Второе по величине, третье по величине, и т.д., деревья остаются в прежнем объеме порядка N^2 , поэтому возникает явление, которое называется гигантской компонентой. Переход n через N^2 (в смысле порядка) аналогичен переходу вероятности возникновения ребра в случайном графе $G(n, p)$ через $(\ln n)/n$ (в смысле строгой асимптотики).

Схематично эти выводы собраны на рис 6.12.

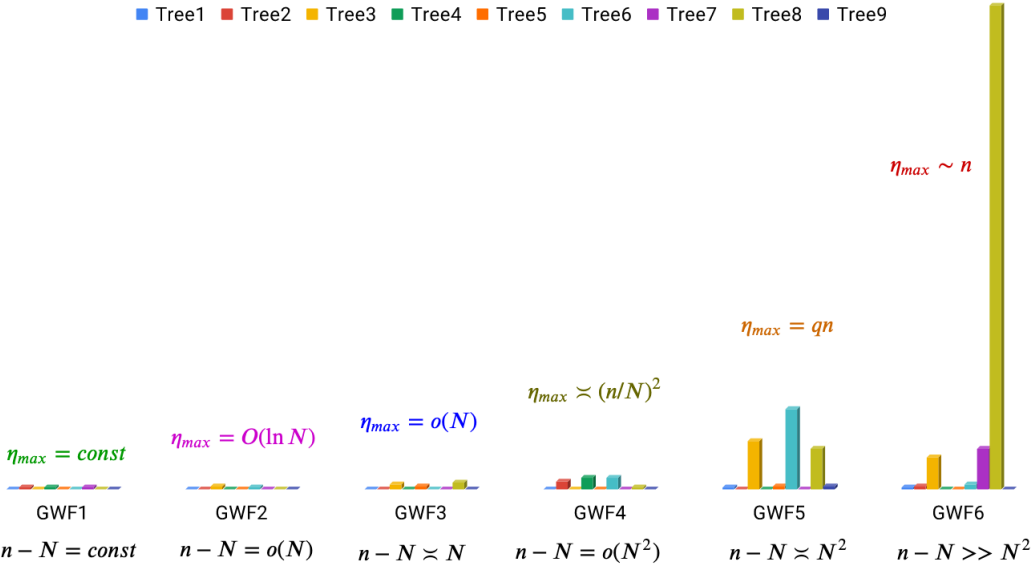


Рис. 6.12: Эволюция леса Гальтона–Ватсона (объемы деревьев).

Комментарий 27.

Лес Гальтона–Ватсона можно рассматривать как простейшую модель эволюции какого-либо вида животных (с человеком все сложнее, т. к. он сильнее управляет собственной эволюцией). Если мы зафиксируем в какой-то момент времени количество особей одного пола и начнем смотреть на их эволюцию на длительном промежутке времени (так, чтобы число поколений измерялось сотнями или тысячами), от мы попадаем в условия шестой зоны изменения параметров случайного леса. А это значит, что почти 100% особей того же пола данного вида по прошествии сотен или тысяч поколений будут составлять потомки какой-то одной из начальных особей. Если уж совсем «притягивать модель за уши», то можно сказать, что если когда-то и было несколько Адамов (несколько очагов цивилизации), то сейчас почти наверняка живут потомки какого-то одного Адама. В некотором смысле это оправдывает поверие, что все мы родственники в n -ом поколении, где n каждый может выбрать себе по вкусу (обычно выбирают 7, хотя такое маловероятно). Косвенно этот математический результат подтверждается исследованиями митохондриальных ДНК.

Ценность приведенных здесь теорем заключается как раз в том, что они чисто математические. То есть, если модель леса гальтона–Ватсона хорошо описывает какой-то физический или социальный процесс, то мы можем уверенно пользоваться данными результатами, не взирая на природу исследуемого процесса.

Стоит сказать несколько слов о *методе исследования*. Мы работаем на вероятностном пространстве $\langle \mathcal{F}_{N,n}, \mathbf{P} \rangle$, где вероятностная мера задается формулой (6.18). При этом вместо последовательности $\{z_k\}$ используется последовательность вероятностей $\{p_k\}$ с конечным вторым моментом, генерирующая процесс Гальтона–Ватсона.

Поскольку каждый лес из $\mathcal{F}_{N,n}$ задается кортежем из плоских посаженных деревьев, мы можем определить совместное распределение $\mathbf{P}\{\eta_1 = n_1, \dots, \eta_N = n_N\}$, где $\eta_i = \|V(t_i)\|$, $i = \overline{1, N}$, $f = (t_1, \dots, t_N)$. Эта вероятность выражается следующим образом

$$\mathbf{P}\{\eta_1 = n_1, \dots, \eta_N = n_N\} = \frac{\sum_{\substack{t_1, \dots, t_N \in T_3 \\ \|V(t_i)\| = n_i}} \mu(t_1) \cdots \mu(t_N)}{\sum_{\substack{t_1, \dots, t_N \in T_3 \\ \|V(t_1)\| + \dots + \|V(t_N)\| = n}} \mu(t_1) \cdots \mu(t_N)}.$$

Дальше нам бы хотелось видеть в числителе и знаменателе вероятности. Это значит, что нам нужна конечная мера множества T_3 . Эта мера совпадает с числом $T(1)$, которое является корнем уравнения $T = F(T)$ в силу (6.17). Такой корень всегда существует, т. к. $F(1) = 1$. Разделим числитель и знаменатель последней дроби на $T(1)^N$ и получим

$$\mathbf{P}\{\eta_1 = n_1, \dots, \eta_N = n_N\} = \frac{(t_{n_1}/T(1)) \cdots (t_{n_N}/T(1))}{\sum_{\substack{k_1, \dots, k_N \geq 1 \\ k_1 + \dots + k_N = n}} (t_{k_1}/T(1)) \cdots (t_{k_N}/T(1))}.$$

Определим теперь серию независимых неотрицательных одинаково распределенных случайных величин ξ_1, \dots, ξ_N со следующим распределением:

$$\mathbf{P}\{\xi_i = n\} = \frac{t_{n+1}}{T(1)}, \quad n = 0, 1, \dots,$$

Тогда

$$\mathbf{P}\{\eta_1 - 1 = n_1, \dots, \eta_N - 1 = n_N\} = \frac{\mathbf{P}\{\xi_1 = n_1\} \cdots \mathbf{P}\{\xi_N = n_N\}}{\mathbf{P}\{\xi_1 + \dots + \xi_N = n - N\}} \quad (6.24)$$

для всех неотрицательных разбиений $n - N = n_1 + \dots + n_N$.

Последнее равенство можно интерпретировать как случайное размещение $n - N$ частиц по N ячейкам, т. е. как вероятностную схему размещения. Если случайные размещения удовлетворяют равенству (6.24) с независимыми

одинаково распределенными ξ_k (не обязательно полученными через количественные характеристики деревьев), то такая схема размещения называется **обобщенной схемой размещения**. В [122] можно найти множество примеров использования такой конструкции для изучения дискретных объектов (графов, лесов и т.п.). В [112] доказано несколько достаточно общих теорем для обобщенной схемы размещения, позволяющих получать результаты для случайных графов (в том числе лесов Гальтона–Ватсона).

Основной бонус равенства (6.24) состоит в том, что изучение совместного распределения зависимых величин η_k сводится к изучению распределения независимых ξ_k и их суммы. Для суммы, как мы знаем, имеются предельные теоремы теории вероятностей,¹² а изучение распределения ξ_k сводится к генерирующей ее последовательности чисел t_n .

Тем не менее, получение предельных распределений для сумм независимых случайных величин сопряжено с определенными трудностями, если требуется изучать «хвосты» распределения, т. е. зону больших отклонений n от среднего значения $\xi_1 + \dots + \xi_N$. И вот тут наступает звездный час параметризации по формуле (6.23). Как мы уже видели в предыдущем разделе, такая параметризация не меняет вероятностное пространство случайных лесов, так что мы можем перейти от меры $\mu(t)$ к мере $\mu(t, \lambda)$, определенной в (6.20), ничего не теряя.

Упражнение 6.10. При этом вместо случайных величин ξ_k мы придем к величинам $\xi_k(\lambda)$ с распределением

Получите это равенство.

$$P\{\xi_i(\lambda) = n\} = \frac{t_{n+1}\lambda^{n+1}}{F(\lambda)^{n+1}T(\lambda/F(\lambda))}, \quad n = 0, 1, \dots,$$

То есть, мы приходим к обобщенной схеме размещения вида

$$P\{\eta_1 - 1 = n_1, \dots, \eta_N - 1 = n_N\} = \frac{P\{\xi_1(\lambda) = n_1\} \dots P\{\xi_N(\lambda) = n_N\}}{P\{\xi_1(\lambda) + \dots + \xi_N(\lambda) = n - N\}}, \quad (6.25)$$

где мы вольны выбирать λ как угодно в области сходимости ряда $F(\lambda)$.

Обычно для получения предельных теорем этот параметр выбирается так, чтобы выполнялось равенство

$$E(\xi_1(\lambda) + \dots + \xi_N(\lambda)) = n - N. \quad (6.26)$$

Заметим, что $T(x)$ на отрезке $[0; 1]$ монотонно растет и принимает значения от 0 до наименьшего положительного корня уравнения $\lambda = F(\lambda)$, которое

¹²Выше мы упоминали только центральную предельную теорему, которая относится к классу интегральных теорем, т. е. дает асимптотику функции распределения. Но существует также целый ряд локальных предельных теорем, дающих асимптотику для конкретных вероятностей или ограниченных интервалов, в которых данная сумма изменяется.

мы обозначим λ_0 .¹³ Это значит, что существует обратная функция $\mu(\lambda)$, определенная на $[0; \lambda_0]$ и такая, что $T(\mu(\lambda)) = \lambda$.

Но тогда в силу тождества (6.17) получаем $\lambda = T(\mu(\lambda)) = \mu(\lambda)F(T(\mu(\lambda)))$, т. е. $\mu(\lambda) = \lambda/F(\lambda)$, откуда $T(\lambda/F(\lambda)) = \lambda$.

Далее, легко видеть, что

$$E\xi_1(\lambda) = \sum_{n=0}^{\infty} (n+1) \frac{t_{n+1}\lambda^{n+1}}{F(\lambda)^{n+1}T(\lambda/F(\lambda))} - 1 = T'(\lambda/F(\lambda))/F(\lambda) - 1$$

Но поскольку $T(x) = xF(T(x))$, получаем $T'(x) = F(T(x)) + xF'(T(x))T'(x)$, откуда $T'(x) = F(T(x))/(1 - xF'(T(x)))$ и, подставляя $x = \lambda/F(\lambda)$, имеем

$$E\xi_1(\lambda) = \frac{\lambda F'(\lambda)/F(\lambda)}{1 - \lambda F'(\lambda)/F(\lambda)}.$$

Отсюда видно, что равенство (6.26) выполняется, если

$$\frac{\lambda F'(\lambda)}{F(\lambda)} = \frac{n - N}{n}. \quad (6.27)$$

Пользуясь неравенством Гёльдера, можно показать, что величина слева монотонно растет по λ и в точке λ_0 достигает значения $F'(\lambda_0)$. Таким образом, для тех распределений $\{p_k\}$, у которых $F'(\lambda_0) \geq 1$, мы можем выбрать параметр λ таким образом, чтобы выполнялось целевое равенство (6.26). В противном случае придется использовать более тяжелую артиллерию для получения предельных теорем для сумм случайных величин.

Например, если мы рассмотрим лес Гальтона–Ватсона, соответствующий равномерному распределению плоских посаженных лесов на множестве $\mathcal{F}_{N,n}$, т. е. возьмем в качестве генерирующей последовательности $\{z_k = 1\}$, то можем в качестве вероятностей числа потомков соответствующего процесса Гальтона–Ватсона взять

$$p_k = \alpha^k(1 - \alpha), \quad k = 0, 1, 2, \dots,$$

а затем перейти к параметрической схеме размещения вида (6.25), где

$$P\{\xi_i(\lambda) = k\} = (\alpha\lambda)^k(1 - \alpha\lambda), \quad k = 0, 1, 2, \dots$$

В этом случае $F(\lambda) = (1 - \alpha)/(1 - \alpha\lambda)$, и $\lambda_0 = 1$, а также

$$\frac{\lambda F'(\lambda)}{F(\lambda)} = \frac{\alpha\lambda}{1 - \alpha\lambda},$$

¹³Известно [114], что λ_0 является вероятностью вырождения соответствующего процесса Гальтона–Ватсона, стартовавшего с 1 частицы.

так что удовлетворить равенство (6.27) не составит труда, если основной параметр положить $\alpha = 1/2$ и вспомогательный параметр λ выбирать достаточно близким к 1.

В данном примере у нас получилась трехэтажная башня моделей одного и того же вероятностного пространства: 1) равномерное распределение с генерирующей последовательностью $\{z_k = 1\}$, 2) надстройка в виде $p_k = \alpha^k(1 - \alpha)$ для попадания в схему процесса Гальтона-Ватсона с нужным ограничением на второй момент, 3) параметризация в виде $p_k(\lambda) = (\alpha\lambda)^k(1 - \alpha\lambda)$ для удобства использования обобщенной схемы размещения. Если бы нам не нужно было делать отсылку к процессу Гальтона-Ватсона и вероятности его вырождения λ_0 , то мы бы ограничились шагом 2) и использовали бы α в качестве того параметра, который упрощает получение предельных теорем в обобщенной схеме размещения, т. к. уравнение $\alpha/(1 - \alpha) = (n - N)/N$ также легко разрешается.

В завершение нужно сказать, что мало получить оценку совместного распределения (η_1, \dots, η_N) , поскольку нас интересует вариационный ряд, полученный из этих величин. Для оценки членов вариационного ряда можно пользоваться формулой

$$P\{\eta_{(N-l)} \leq r\} = \sum_{k=0}^l \binom{N}{k} P_r^k (1 - P_r)^{N-k} \frac{P\{\xi_1^{(r)} + \dots + \xi_{N-k}^{(r)} + \bar{\xi}_1^{(r)} + \dots + \bar{\xi}_k^{(r)} = n - N\}}{P\{\xi_1 + \dots + \xi_N = n - N\}}, \quad (6.28)$$

где

$$\begin{aligned} P_r &= P\{\xi_1 > r\}, \\ P\{\xi_i^{(r)} = k\} &= P\{\xi_i = k \mid \xi_i \leq r\}, \\ P\{\bar{\xi}_i^{(r)} = k\} &= P\{\xi_i = k \mid \xi_i > r\}. \end{aligned}$$

То есть, изучение членов вариационного ряда также сводится к суммам случайных величин, определенным способом усеченных.

Случайные подстановки

Приведем еще один пример из вероятностной теории графов, который объединяет в себе как случайные графы, так и леса, и обобщенную схему размещения. Речь идет о случайных подстановках и рекурсивных лесах.

Случайная подстановка — это элемент группы S_n , которому естественным образом соответствует граф, каждая компонента связности которого есть ориентированный цикл. Например, подстановке $(123)(45)$ соответствует

граф на вершинах $\{1, 2, 3, 4, 5\}$, в котором дуги и ребра связаны следующим образом:

$$1 \rightarrow 2 \rightarrow 3 \rightarrow 1, \quad 4 \rightarrow 5 \rightarrow 4.$$

Легко видеть, что различных циклов длины n существует $(n-1)!$. Точно так же, существует $(n-1)!$ различных рекурсивных деревьев.

Рассмотрим множество $\mathcal{S}_{N,n}$ всех подстановок на n точках с N занумерованными(!) циклами. Введем на этом множестве равномерное распределение. Тогда тождественная функция $RS_{N,n} = \text{id}$ на этом множестве будет **случайной подстановкой**. Легко видеть, что количество подстановок в $\mathcal{S}_{N,n}$, у которых длины циклов равны k_1, \dots, k_N , равно

$$\frac{n!}{k_1! \dots k_N!} (k_1 - 1)! \dots (k_N - 1)! = \frac{n!}{k_1 \dots k_N},$$

где $n = k_1 + \dots + k_N$, $k_i > 0$.

Следовательно, если обозначить η_1, \dots, η_N случайные величины, равные длинам циклов в случайной подстановке $RS_{N,n}$, то получим

$$\mathbf{P}\{\eta_1 = n_1, \dots, \eta_N = n_N\} = \frac{(n_1)^{-1} \dots (n_N)^{-1}}{\sum_{k_1 + \dots + k_N = n} (k_1)^{-1} \dots (k_N)^{-1}}. \quad (6.29)$$

Ясно, что это распределение удовлетворяет формуле (6.25), если в качестве независимых одинаково распределенных случайных величин ξ_i выбрать такие:

$$\mathbf{P}\{\xi_i = k\} = \frac{\lambda^{k+1}}{k \ln(1 - \lambda)^{-1}}, \quad k = 0, 1, 2, \dots$$

Точно так же, можно взять множество рекурсивных лесов объема n с N занумерованными деревьями, ввести на нем равномерное распределение, и окажется, что для изучения совместного распределения объемов деревьев случайного рекурсивного леса можно пользоваться схемой (6.29). Поэтому ниже все результаты об этой схеме являются общими для случайной подстановки и случайного рекурсивного леса.

Как и раньше, обозначим $\eta_{(1)} \leq \dots \leq \eta_{(N)}$ вариационный ряд компонент рассматриваемой схемы. Для изучения предельного поведения величин $\eta_{(N-l)}$ используется тождество (6.28) и предельные теоремы для сумм независимых случайных величин.

Известны следующие утверждения, точность которых, как и в случае лесов Гальтона–Ватсона, переведена в падающий режим, а за более подробными описаниями и доказательствами мы отсылаем к работе [112].

RS1 Пусть $N \rightarrow \infty$ и $n - N = \text{const}$. Тогда $\mathbf{P}\{\eta_{(N-l)} = 2\} \rightarrow 1$ для старших $n - N$ компонент, остальные компоненты вырождены (асимптотически почти наврное).

- RS2 Пусть $n - N \rightarrow \infty$ так, что $n - N = o(N)$. Тогда существует последовательность $\{r_{N,n}\}$ порядка $o(\ln N)$ такая, что при каждом фиксированном l величина $\eta_{(N-l)}$ сосредоточена на множестве $\{r_{N,n}, r_{N,n} + 1\}$ с вероятностями, зависящими от l .
- RS3 Пусть $N, n \rightarrow \infty$ так, что $(n - N)/N \asymp 1$. Тогда существует последовательность $\{r_{N,n}\}$ порядка $\ln N$ такая, что при любом фиксированном l величина $\eta_{(N-l)}$ имеет порядок $r_{N,n}$.
- RS4 Пусть $N, n \rightarrow \infty$ так, что $(n - N)/N \rightarrow \infty$ и $(\ln n)/N \rightarrow 0$. Тогда при любом фиксированном l величина $\eta_{(N-l)}$ имеет порядок $o(n)$.
- RS5 Пусть $N, n \rightarrow \infty$ так, что $(\ln n)/N \asymp 1$. Тогда при любом фиксированном l величина $\eta_{(N-l)}$ имеет порядок n . Точнее, старшие компоненты распределяются вокруг некоторых долей числа n .
- RS6 Пусть $N, n \rightarrow \infty$ так, что $(\ln n)/N \rightarrow \infty$. Тогда $\eta_{(N)}$ асимптотически равна n (отклонения от n бесконечно малы в сравнении с n), а следующие по величине компоненты $\eta_{(N-l)}$ при фиксированном l имеют порядок (\cdot) .

Как видим, здесь наблюдается практически такая же картина, как в случае лесов Гальтона–Ватсона, только «контрольным» отношением является $(\ln n)/N$, что имеет свою логику, поскольку схема (6.29) логарифмическая.¹⁴

Лес является «широким» в первых трех зонах (пока число вершин по порядку не превосходит число деревьев/циклов), а затем максимальное дерево/цикл начинают свой рост, в пятой зоне достигает порядка всего графа, а шестой зоне мы имеем подавляющую гигантскую компоненту, т. е. здесь случайная подстановка практически полностью состоит из одного цикла.

На этом мы завершаем наш экскурс в теорию графов и ее вероятностную составляющую. Вместе с главой мы приходим к финалу книги.

¹⁴ для лесов Гальтона–Ватсона установлено, что коэффициенты, определяющие схему, являются степенями вида $k^{-\delta}$, где $\delta > 1$, а это существенно отличается от логарифмической схемы хотя бы в силу того, что ряд из таких величин сходится.

Если теперь коротко окинуть взором все то, что мы отстроили, то картина будет примерно следующая.

Перед нами — планета, имеющая свой неповторимый рельеф и географию, природа которой задается несколькими первородными структурами — множествами. Над этими структурами трудятся функции и формулы, производящие с помощью рекурсий все новые и новые виды форм и структур, населяющих эту планету. Между тем, в мире царит алгоритмическая гармония, чистота мысли и красота теорем. Над первородным ландшафтом постепенно выстроились и расцвели целые города и страны математических наук со своими заводами-теориями, башнями различных факторизаций и алгебраических структур, магистралями быстрых алгоритмов и провинциальными дорогами NP-полных задач, междисциплинарными коммуникациями формальных языков и строгими компьютерными конвейерами сложных вычислений, ветвистыми садами и парками из случайных лесов и убегающими в небо трансфинитными универсумами.

Эти математические науки огромной живой сетью оплели весь этот чудесный мир и то и дело обосновываются на сияющих снежных вершинах великих открытий, пускаются в рискованное плавание по океану сложных современных проблем не только математики, но и науки в целом, проникают в параллельные миры естественных и социальных наук.

На протяжении книги мы старались наблюдать на конкретных примерах за теми архетипами, которые мы обозначили в самом начале. Мы начали с освоения письменности — рассмотрели простейшую грамматику, порассуждали о скобочной записи множеств, ввели обозначения натуральных чисел, научились записывать высказывания формулами. Далее мы перешли к оформлению мыслительного процесса с помощью освоенной письменности — определили свойства изучаемых объектов через аксиоматику, ввели много полезных определений вроде функций и отношений, зафиксировали первые архетипы, напрямую связанные с основными идеями построения математики с нуля.

При этом мы все время старались держать в голове принятую в начале книги идею о том, что нам хочется объяснить математику машине. Это означает, что в достаточно бедный язык нужно было умудриться уложить все разнообразие форм и структур, о которых говорилось чуть выше.

Отчасти это удалось, т.к. мы успешно объяснили, как в «голой» теории множеств, даже в ее начальном сегменте, можно задать натуральные числа и их продолжения — целые числа, ординалы, рациональные числа, действительные и сюрреальные числа.

Отчасти — не удалось, поскольку мы уперлись в проблему отделения понятия от объекта, ибо множество ω — хорошая модель натуральных чисел, но она обладает рядом ненужных свойств, пользоваться которыми в арифметике ни в коем случае нельзя. Это, с одной стороны, дает нам повод усомниться в силе оснований математики и возможности научить компьютер абстрактным понятиям, а с другой стороны заставляет лучше понять, что математика — это не просто язык, но еще и сложная система понятий и свойств, которые могут жить сами по себе, независимо от породившего их формализма. Более того, эти понятия способны проникать в разные теории, в том числе за пределами математики, и давать там новые ростки знаний, невидимые ранее с прежнего уровня восприятия.

Поработав с числами и их обобщениями, мы окунулись в общую алгебру и получили возможность посмотреть на математический мир под другим углом, а заодно увидеть «работу» высших абстракций в конкретных задачах вроде теорем Ферма, движений сферы и разрешения уравнений в радикалах. Попутно мы выявили ряд новых архетипов под общим названием числовых архетипов математики.

Понимание того, что в математике все переплетено так, словно она выросла из одного священного источника, заставляет нас обратиться к рассмотрению самой тонкой и самой уязвимой для критики материи — метаматематике или математической логике. Уязвимая она по той причине, что занимается анализом математических текстов, и, тем самым, кажется отстоящей от реальности еще дальше, чем общая алгебра. Тем не менее, матлогика дает ответы на вопросы о состоятельности математических теорий, об их границах и взаимосвязях. А кроме того, находит прямое применение в компьютерном мире, где исследование формализмов жизненно необходимо для создания емких, удобных и выразительных языков взаимодействия с компьютером.

Таким образом, мы прошли полный круг от основ к основаниям через алгебру, но при этом поднялись на новый уровень восприятия математики. После чего уместно было снова окунуться в мир чисел и исследовать уже аналитические исчисления, вроде интегрального и вариационного, подкрепляя это исследование общими знаниями о пространствах, мерах и операторах. На этом пути мы снова встречали ранее обнаруженные архетипы (базового множества, функции, факторизации, редукции и т.п.), а также добавили новый — архетип связности–непрерывности, который является сутью топологии и анализа континуума.

С самого начала книги мы привлекали архетип графа для иллюстрации тех или иных конечных объектов, прежде всего, формул, термов и «начальных» множеств. Даже многие вспомогательные иллюстрации книги можно отнести к графам. И это неслучайно, поскольку граф является наиболее простой и в то же время достаточно богатой иллюстрацией всех хитросплетений знаний, методов, выводов и алгоритмов, которыми насыщена математика (да

что математика — вся жизнь!)

Поэтому сияющей вершиной и кульминацией книги является глава, посвященная графам. В этой главе мы вновь увидели знакомые архетипы, а также смогли на примерах убедиться в глубокой связи всех математических дисциплин, когда рассматривали различные свойства графов и случайных графов.

Остановимся еще раз на некоторых архетипах (полный перечень которых находится в приложении А). Как водится, самое простое часто оказывается самым сложным. В подтверждение можно вспомнить об архетипе равенства. Казалось бы, все понимают, что такое $a = b$, однако при детальном рассмотрении оказывается, что определить равенство не так-то просто. В нашей книге мы посвятили этому архетипу немало страниц, показывая, что равенство в математике — это прежде всего свойство языка, оно не задано кем-то свыше и не может быть универсальным понятием. Так, равенство отрезков (пар точек) в геометрии Тарского, по сути дела, есть конгруэнтность отрезков в обычной геометрии, т. е. формальная геометрия Тарского беднее геометрии с точками и отрезками. В общем же случае равенство задается вместе с языком как такое свойство объектов теории, что формализм этой теории не в состоянии отличить равные объекты (см. аксиомы равенства).

Архетип множества пронизывает всю математику, какой бы формальной или прикладной она ни была. Без этого понятия нет ни Алгебры, ни теории графов, ни Анализа. Да, можно спорить о пользе аксиоматик, о том, что множества можно смоделировать, например, графами или ограничиваться только начальной теорией множеств, но без архетипа множества математика просто не существует. Объясняется это очень просто: множество есть область истинности утверждения. Соответственно, если теория описывает формулами какие-то свойства объектов, то она уже «под капотом» имеет архетип множества. А кроме того, это понятие позволяет наглядно выстраивать очень многие формальные конструкции, т. е. служит удобным строительным материалом.

Архетип графа примечателен не столько конкретной реализацией в математической структуре с тем же названием, сколько самой идеей наглядно демонстрировать связи и структуры. Графами можно описать алгоритмы, отношения, термы, начальные множества, уравнения, дать определения математических понятий (см. картинку определения отношений и функций), изобразить логические диаграммы и симметрии, и т. д. Не исключено, что на голом понятии графа вообще можно выстроить весь формализм теории множеств и любой другой математической теории.

Примерно половина книги посвящена числам, но подчеркнем еще раз, что под числами мы понимаем такие сущности, которые позволяют построить исчисление. То есть такую теорию, решение задач в которой можно практически полностью отдать на откуп компьютеру — просто бери и делай. Числа всегда связаны с операциями сложения и умножения (друг на друга и/или

на число), но не обязаны быть линейно упорядоченными. Числа нужны для того, чтобы сложные математические структуры можно было оцифровать и свести к привычным операциям и соотношениям.

Архетип связности-непрерывности, подмеченный нами в связи с изучением топологии, также глубоко пронизывает математические идеи. По-видимому, как и в случае чисел, это отражает наше стремление к овеществлению математических объектов, чтобы их можно было представлять как некоторые протяженные физические сущности, которые можно деформировать, связывать, растягивать, разрезать, в общем, совершать с ними такие преобразования, которые укладываются в рамки физической интуиции.

Помимо архетипов, связывающих различные математические объекты и структуры, мы также выявили несколько **архетипических методов**. Например, редукцию, которая предполагает возможность свести сложную задачу к простой (функция Жуковского, ортогональный базис и т.д.) с приемлемой потерей количества информации, выполнить некоторые ранее разработанные преобразования, а затем вернуться в сложную задачу, имея о ней уже новое представление и, возможно, даже решение.

Наиболее сложным архетипическим методом является трансцендентный объективизм (да, название получилось так себе). Суть метода заключается в том, чтобы опровергнуть детерминистское утверждение об изучаемых объектах методом от противного. Так мы опровергаем, что все множества образуют множество, что простых чисел существует лишь конечный набор, что континуум счетен и что всякое истинное утверждение доказуемо.

Наконец, стоит отметить и самые простейшие архетипы, которыми мы пользуемся так же, как родным языком: часто, непринужденно и незаметно для себя. Речь идет об архетипах базового множества, порождающего элемента и неподвижной точки.

Конечно, мы совершили лишь краткую экскурсию по планете математики, хотя и пресыщенную порой тяжеловесными теоремами. Но это была экскурсия для искушенных туристов! Кому-то она могла показаться слишком трудной, кому-то слишком занудной, но для кого-то, как, например, для автора, стала комнатой с многими дверями, за которыми таится чудесный бесконечный мир математики.

Указатель архетипов

Равенство. Свойство математического языка, отделяющее описываемые этим языком объекты. Является самым мелкомасштабным отношением эквивалентности в данном языке.	32
Множество. Совокупность элементов произвольной природы, рассматриваемая как единая актуальная сущность («коробка с карандашами»).	26
Система множеств. Множества сами становятся элементами и образуют новую совокупность, обладающую определенными свойствами («палета с коробками»).	69
Универсум. Упаковывание всех изучаемых объектов в башню универсумов, конструктивно покрывающую все объекты некоторым рекурсивным способом («фура с палетами»).	99
Факторизация. Сведение исходной совокупности объектов к упрощенной путем отождествления объектов, различия между которыми нам неинтересны. Тесно связан с архетипом изоморфизма.	78
Функция. Совокупность правил, предписаний, преобразований, позволяющая получать из одних объектов или высказываний другие.	27
Изоморфизм. Возможность видеть общую структуру в множествах и системах различной природы. Абстрагирование от природы сущностей. Изучение понятий и свойств вместо конкретных объектов.	78
Подобие. По аналогии с хорошо разработанным понятием вводится его аналог или обобщение так, что и свойства этого нового понятия весьма похожи на свойства исходного понятия. Этот архетип подобен изоморфизму и равенству.	445
Граф. Совокупность элементов произвольной природы, некоторые из которых связаны однонаправленными или обоюдными связями—стрелками.	531
Дерево. Структура каталога, формулы, алгоритма, «начального» множества, а также деревья решений, остовные деревья и т.д.	543
Рекурсия. Построение бесконечных объектов путем неограниченного (само)применения процедуры к построенным на предыдущих шагах объектам.	27

Индукция. Рекурсивный метод рассуждений.	27
Порождающий элемент. В паре с какой-то функцией, формулой и/или рекурсией задает множество с определенными свойствами.	69
Числа. Система объектов, которые можно складывать, вычитать, умножать и делить, а также сравнивать (с теми или иными ограничениями), причем на интуитивном уровне операции должны быть подобны операциям с объемами и кратностями: сложение одинаковых объемов увеличивает их кратность, умножение на кратность означает повторение сложения заданное число раз.	101
Ноль. Он же — пустое множество, он же — нейтральный элемент, он же — начало отсчета (имя ему — легион). Количество свойств нуля равно $1/0$.	332
Башни операций. Многократное сложение — умножение, многократное умножение — степень, многократная степень — башня степеней, многократные башни степеней — тройные степени, многократная факторизация — башни структур, и т.д.	332
Упорядочение. Приписывание объектам сравнимых по величине чисел — шкалирование.	332
Числа над числами. Использование одних чисел (более общих) для изучения других, и обратно: конструирование «новых» чисел из «старых».	332
Неограниченное расширение. Каким бы огромным ни был объект, всегда есть непротиворечивый способ рассмотреть его как атом в более обширной вселенной.	149
Безграничное деление. Какими бы близкими ни казались объекты, всегда можно найти непротиворечивый способ вставить между ними целую вселенную.	149
Недостижимость. Существуют столь мощные объекты, которые не могут быть достигнуты менее мощными средствами, чем они сами.	383
Базовое множество (отождествление структуры с ее носителем). Восприятие базового множества в совокупности со структурой, заданной на нем, но формально не являющейся его частью и не зависящей от него.	197
Инвариант. Свойство, которое сохраняется при разрешенных манипуляциях (функциях, рекурсиях, изоморфизмах и т.п.) над объектами в пределах какого-то множества.	213
Вариативность представления. Один и тот же объект (понятие) может изоморфно вкладываться в структуры, природа построения которых не имеет ничего общего. Возникают своего рода «червоточины», позволяющие переходить между математическими мирами и осуществлять трансцендентное восприятие.	237

Двойственность. Выделение парных (смежных) сущностей, которые могут обмениваться ролями, в результате чего возникает двойная интерпретация задачи.	290
Связность-непрерывность. Основной топологический архетип. Он же в топологическом смысле характеризует и графы.	415
Компактность. Конечность математической структуры в смысле ее пространственных или образующих свойств.	431
Неподвижная точка.	428
Архетипические методы:	
Трансцендентное восприятие. Погружение изучаемой области в более обширную с целью изучать и решать задачи исходной области внешними по отношению к ней методами (взгляд со стороны).	191
Метод бесконечного спуска. Показываем, что при нашем предположении можно бесконечно снижать числовой параметр системы, при том, что этот параметр является ординалом. А это невозможно.	244
Трансцендентный объективизм. В предположении некоторой общности заданного определения к нему строится контрпример (например, диагональным методом), опровергающий регулярный характер этого определения.	95
Редукция. Сведение задачи (объекта, формулы) к более простой, ранее хорошо изученной.	227
Подходящий базис. Среди эквивалентных базисов выбирается тот, в котором решение задачи становится проще. Разновидность редукции.	345
Создание исчислений. Этап развития математики, когда накопленные знания упаковываются в стройную теорию-исчисление, которое можно преподносить как готовый численно-логический инструмент.	389

Схемы и таблицы

Таблица В.1: Карта книги в виде тегов. Стрелки обозначают переход от простого к сложному, от идеи к теореме, или ничего не значат.

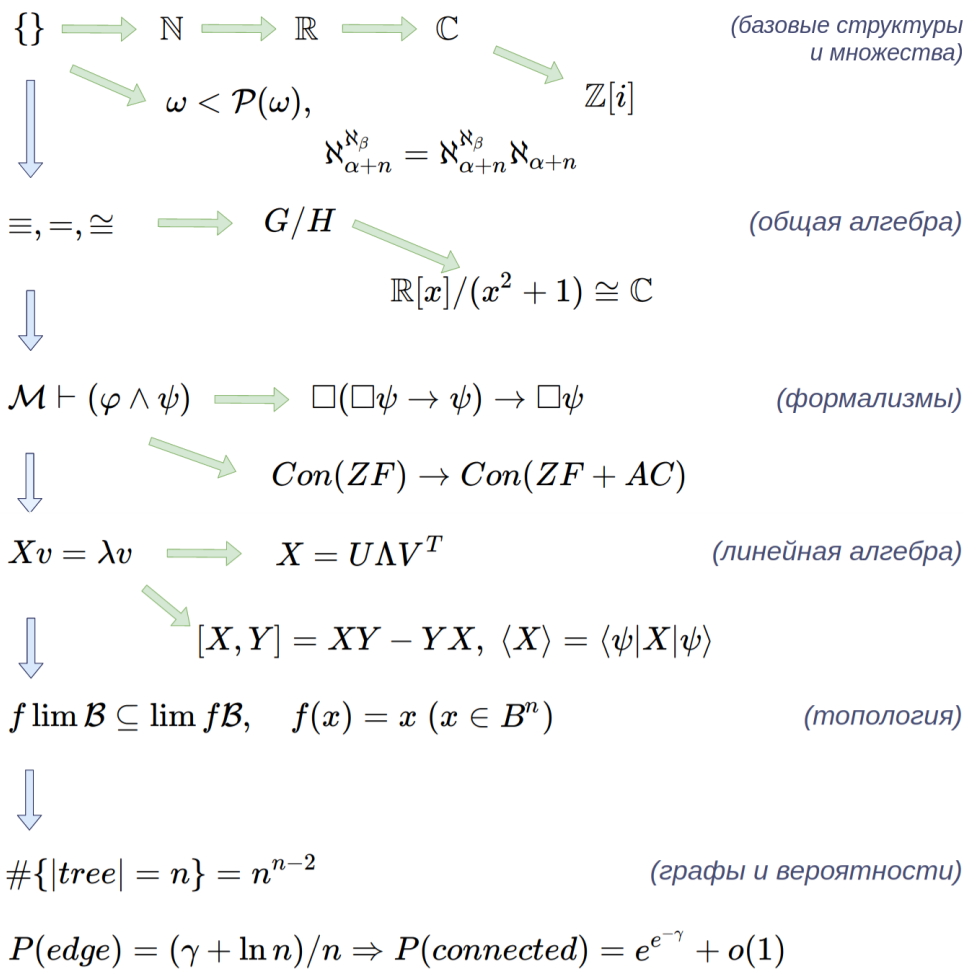


Таблица В.2: Сравнение арифметики в числовых структурах.

	Натуральные числа	Ординалы	Кардиналы
Ноль	$n + 0 = n = 0 + n$	$\alpha + 0 = \alpha = 0 + \alpha$	$\varkappa + 0 = \varkappa = 0 + \varkappa$
Единица	$n \cdot 1 = n = 1 \cdot n$	$\alpha \cdot 1 = \alpha = 1 \cdot \alpha$	$\varkappa \cdot 1 = \varkappa = 1 \cdot \varkappa$
Делители нуля	$nm = 0 \rightarrow n = 0 \vee m = 0$	$\alpha\beta = 0 \rightarrow \alpha = 0 \vee \beta = 0$	$\varkappa\mu = 0 \rightarrow \varkappa = 0 \vee \mu = 0$
Особенности вычисления			$\varkappa + \mu = \max(\varkappa, \mu)$ $\varkappa \cdot \mu = \max(\varkappa, \mu)$, если $(k \geq \omega) \vee (\mu \geq \omega)$
Обратимость	нет	нет	нет
Коммутативность	$n + m = m + n$ $nm = mn$	$\omega + 1 \neq 1 + \omega$ $\omega \cdot 2 \neq 2 \cdot \omega$	$\varkappa + \mu = \mu + \varkappa$ $\varkappa \cdot \mu = \mu \cdot \varkappa$
Ассоциативность	$n + (m + k) = (n + m) + k$ $n(mk) = (nm)k$	$\alpha + (\beta + \gamma) = (\alpha + \beta) + \gamma$ $\alpha \cdot (\beta \cdot \gamma) = (\alpha \cdot \beta) \cdot \gamma$	$\varkappa + (\mu + \nu) = (\varkappa + \mu) + \nu$ $\varkappa \cdot (\mu \cdot \nu) = (\varkappa \cdot \mu) \cdot \nu$
Дистрибутивность слева	$n(m + k) = nm + nk$	$\alpha \cdot (\beta + \gamma) = \alpha \cdot \beta + \alpha \cdot \gamma$	$\varkappa \cdot (\mu + \nu) = \varkappa \cdot \mu + \varkappa \cdot \nu$
Дистрибутивность справа	$(m + k)n = mn + kn$	$(\omega + 1) \cdot 2 \neq \omega \cdot 2 + 2$	$(\mu + \nu) \cdot \varkappa = \mu \cdot \varkappa + \nu \cdot \varkappa$
Монотонность слева	$n < m \rightarrow n + k < m + k$ $n < m \rightarrow nk < mk \ (k > 0)$	$1 + \omega = 2 + \omega$ $1 \cdot \omega = 2 \cdot \omega$	$1 < \omega$, но $1 + \omega = \omega + \omega$; $1 < \omega$, но $1 \cdot \omega = \omega \cdot \omega$
Монотонность справа	$n < m \rightarrow k + n < k + m$ $n < m \rightarrow kn < km \ (k > 0)$	$\alpha < \beta \rightarrow \gamma + \alpha < \gamma + \beta$ $\alpha < \beta \rightarrow \gamma\alpha < \gamma\beta \ (\gamma > 0)$	$1 < \omega$, но $\omega + 1 = \omega + \omega$; $1 < \omega$, но $\omega \cdot 1 = \omega \cdot \omega$
Сокращение слева	$k + n = k + m \rightarrow n = m$ $kn = km \rightarrow n = m \ (k > 0)$	$\gamma + \alpha = \gamma + \beta \rightarrow \alpha = \beta$ $\gamma\alpha = \gamma\beta \rightarrow \alpha = \beta \ (\gamma > 0)$	$\omega + 1 = \omega + 2$; $\omega \cdot 1 = \omega \cdot 2$
Сокращение справа	$n + k = m + k \rightarrow n = m$ $nk = mk \rightarrow n = m \ (k > 0)$	$1 + \omega = 2 + \omega$ $1 \cdot \omega = 2 \cdot \omega$	$1 + \omega = 2 + \omega$ $1 \cdot \omega = 2 \cdot \omega$
Вычитание слева	если $n < m$, то $\exists!$ $k = -n + m$: $n + k = m$	если $\alpha < \beta$, то $\exists!$ $\gamma = -\alpha + \beta$: $\alpha + \gamma = \beta$	если $\varkappa < \mu$ и $\mu \geq \omega$, то $-\varkappa + \mu = \mu$
Вычитание справа	если $n < m$, то $\exists!$ $k = m - n$: $k + n = m$	$\omega - 10$ не существует	если $\varkappa < \mu$ и $\mu \geq \omega$, то $\mu - \varkappa = \mu$
Деление слева с остатком	если $m > n > 0$, то $m = nk + s \ (k = \lfloor n^{-1}m \rfloor)$	если $\beta > \alpha > 0$, то $\beta = \alpha\gamma + \delta \ (\gamma = \lfloor \alpha^{-1}\beta \rfloor)$	если $\varkappa < \mu$ и $\mu \geq \omega$, то $\varkappa^{-1} \cdot \mu = \mu$
Деление справа с остатком	если $m > n > 0$, то $m = kn + s \ (k = \lfloor m/n \rfloor)$	ω^ω не делится на ω справа (не существует $\omega^{\omega^{-1}}$)	если $\varkappa < \mu$ и $\mu \geq \omega$, то $\mu/\varkappa = \mu$

Таблица В.3: Продолжение таблицы В.2

	Вычеты по модулю p	Квадратные матрицы	Числа Гаусса
Ноль	$n + 0 = n = 0 + n$	нулевая матрица	$z + 0 = z = 0 + z$
Единица	$n \cdot 1 = n = 1 \cdot n$	E_n	$z \cdot 1 = z = 1 \cdot z$
Делители нуля	нет, если p — простое	ЛЗ строки/столбцы	$zw = 0 \rightarrow z = 0 \vee w = 0$
Особенности вычисления	$k + m \bmod p$ $k \cdot m \bmod p$	$c_{mn} = \sum a_{mk} b_{kn}$	4 делителя единицы и ассоциированные числа
Обратимость	$\frac{-n}{n^{-1}} = \frac{p-n}{n^{p-2}} \bmod p$ $n + m = m + n$	A^{-1} , если $\det A \neq 0$	нет
Коммутативность	$nm = mn$	существуют $A \cdot B \neq B \cdot A$	$z + w = w + z$ $zw = wz$
Ассоциативность	$n + (m + k) = (n + m) + k$ $n(mk) = (nm)k$	$A + (B + C) = (A + B) + C$ $A(BC) = (AB)C$	$z + (w + q) = (z + w) + q$
Дистрибутивность слева	$n(m + k) = nm + nk$	$A(B + C) = AB + AC$	$z(w + q) = zw + zq$
Дистрибутивность справа	$(m + k)n = mn + kn$	$(B + C)A = BA + CA$	$(z + w)q = zq + wq$
Монотонность слева	нет, т.к. цикл	нет естеств. упорядочения	нет упорядочения
Монотонность справа	нет, т.к. цикл	нет естеств. упорядочения	нет упорядочения
Сокращение слева	$k + n = k + m \rightarrow n = m$ $kn = km \rightarrow n = m$ ($k \perp p$)	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} =$	$q + z = q + w \rightarrow z = w$ $qz = qw \rightarrow z = w$ ($q \neq 0$)
Сокращение справа	$n + k = m + k \rightarrow n = m$ $nk = mk \rightarrow n = m$ ($k \perp p$)	$= \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}$	$z + q = w + q \rightarrow z = w$ $zq = wq \rightarrow z = w$ ($q \neq 0$)
Вычитание слева	если $n < m$, то $\exists!$ $k = -n + m$: $n + k = m$	$\forall A, B \exists! C = -B + A$	$\forall z, w \exists! q = -w + z$
Вычитание справа	если $n < m$, то $\exists!$ $k = m - n$: $k + n = m$	$\forall A, B \exists! C = A - B$	$\forall z, w \exists! q = z - w$
Деление слева с остатком	если $m > n > 0$, то $\exists! k \exists! s$: $m = nk + s$	неоднозначно $M = NK + S$	4 варианта остатка
Деление справа с остатком	если $m > n > 0$, то $\exists! k \exists! s$: $m = kn + s$	неоднозначно $M = KN + S$	4 варианта остатка

Таблица В.4: Таблица умножения симметрической группы S_4

Знакопеременная группа A_4											
Четверная группа Клейна											
e	(12)(34)	(13)(24)	(14)(23)	(123)	(132)	(124)	(142)	(134)	(143)	(234)	(243)
(12)(34)	e	(14)(23)	(13)(24)	(243)	(143)	(234)	(134)	(142)	(132)	(124)	(123)
(13)(24)	(14)(23)	e	(12)(34)	(142)	(234)	(143)	(123)	(243)	(124)	(132)	(134)
(14)(23)	(13)(24)	(12)(34)	e	(134)	(124)	(132)	(243)	(123)	(234)	(143)	(142)
(123)	(134)	(243)	(142)	(132)	e	(13)(24)	(143)	(234)	(14)(23)	(12)(34)	(124)
(132)	(234)	(124)	(143)	e	(123)	(243)	(14)(23)	(12)(34)	(142)	(134)	(13)(24)
(124)	(143)	(132)	(234)	(14)(23)	(134)	(142)	e	(13)(24)	(243)	(123)	(12)(34)
(142)	(243)	(134)	(123)	(234)	(13)(24)	e	(124)	(132)	(12)(34)	(14)(23)	(143)
(134)	(123)	(142)	(243)	(124)	(14)(23)	(12)(34)	(234)	(143)	e	(13)(24)	(132)
(143)	(124)	(234)	(132)	(12)(34)	(243)	(123)	(13)(24)	e	(134)	(142)	(14)(23)
(234)	(132)	(143)	(124)	(13)(24)	(142)	(134)	(12)(34)	(14)(23)	(123)	(243)	e
(243)	(142)	(123)	(134)	(143)	(12)(34)	(14)(23)	(132)	(124)	(13)(24)	e	(234)
(12)	(34)	(1324)	(1423)	(23)	(13)	(24)	(14)	(1342)	(1432)	(1234)	(1243)
(13)	(1234)	(24)	(1432)	(12)	(23)	(1243)	(1423)	(34)	(14)	(1342)	(1324)
(14)	(1243)	(1342)	(23)	(1234)	(1324)	(12)	(24)	(13)	(34)	(1423)	(1432)
(23)	(1342)	(1243)	(14)	(13)	(12)	(1324)	(1432)	(1234)	(1423)	(34)	(24)
(24)	(1432)	(13)	(1234)	(1423)	(1342)	(14)	(12)	(1324)	(1243)	(23)	(34)
(34)	(12)	(1423)	(1324)	(1243)	(1432)	(1234)	(1342)	(14)	(13)	(24)	(23)
(1234)	(13)	(1432)	(24)	(1324)	(14)	(1342)	(34)	(1423)	(23)	(1243)	(12)
(1243)	(14)	(23)	(1342)	(1432)	(34)	(1423)	(13)	(24)	(1324)	(12)	(1234)
(1324)	(1423)	(12)	(34)	(14)	(1234)	(1432)	(23)	(1243)	(24)	(13)	(1342)
(1342)	(23)	(14)	(1243)	(24)	(1423)	(34)	(1234)	(1432)	(12)	(1324)	(13)
(1423)	(1324)	(34)	(12)	(1342)	(24)	(13)	(1243)	(23)	(1234)	(1432)	(14)
(1432)	(24)	(1234)	(13)	(34)	(1243)	(23)	(1324)	(12)	(1342)	(14)	(1423)

*Здесь особым фоном выделены элементы, образующие группу, изоморфную \mathbb{Z}_3 , поскольку их 3-я степень равна e.

Таблица В.5: Продолжение таблицы В.4

(12)	(13)	(14)	(23)	(24)	(34)	(1234)	(1243)	(1324)	(1342)	(1423)	(1432)
(34)	(1432)	(1342)	(1243)	(1234)	(12)	(24)	(23)	(1423)	(14)	(1324)	(13)
(1423)	(24)	(1243)	(1342)	(13)	(1324)	(1432)	(14)	(34)	(23)	(12)	(1234)
(1324)	(1234)	(23)	(14)	(1432)	(1423)	(13)	(1342)	(12)	(1243)	(34)	(24)
(13)	(23)	(1423)	(12)	(1243)	(1234)	(1342)	(1324)	(24)	(34)	(1432)	(14)
(23)	(12)	(1432)	(13)	(1324)	(1342)	(34)	(24)	(1243)	(1234)	(14)	(1423)
(14)	(1324)	(24)	(1234)	(12)	(1243)	(1423)	(1432)	(1342)	(13)	(23)	(34)
(24)	(1342)	(12)	(1423)	(14)	(1432)	(23)	(34)	(13)	(1324)	(1234)	(1243)
(1234)	(14)	(34)	(1324)	(1342)	(13)	(1243)	(12)	(1432)	(1423)	(24)	(23)
(1243)	(34)	(13)	(1432)	(1423)	(14)	(12)	(1234)	(23)	(24)	(1342)	(1324)
(1342)	(1423)	(1234)	(24)	(34)	(23)	(1324)	(13)	(14)	(1432)	(1243)	(12)
(1432)	(1243)	(1324)	(34)	(23)	(24)	(14)	(1423)	(1234)	(12)	(13)	(1342)
е	(132)	(142)	(123)	(124)	(12)(34)	(234)	(243)	(13)(24)	(134)	(14)(23)	(143)
(123)	е	(143)	(132)	(13)(24)	(134)	(12)(34)	(124)	(243)	(234)	(142)	(14)(23)
(124)	(134)	е	(14)(23)	(142)	(143)	(123)	(12)(34)	(132)	(13)(24)	(234)	(243)
(132)	(123)	(14)(23)	е	(243)	(234)	(134)	(13)(24)	(124)	(12)(34)	(143)	(142)
(142)	(13)(24)	(124)	(234)	е	(243)	(14)(23)	(143)	(134)	(132)	(123)	(12)(34)
(12)(34)	(143)	(134)	(243)	(234)	е	(124)	(123)	(14)(23)	(142)	(13)(24)	(132)
(134)	(14)(23)	(234)	(124)	(12)(34)	(123)	(13)(24)	(132)	(142)	(143)	(243)	е
(143)	(243)	(13)(24)	(12)(34)	(123)	(124)	(142)	(14)(23)	(234)	е	(132)	(134)
(14)(23)	(124)	(243)	(134)	(132)	(13)(24)	(143)	(142)	(12)(34)	(123)	е	(234)
(234)	(142)	(12)(34)	(13)(24)	(134)	(132)	(243)	е	(143)	(14)(23)	(124)	(123)
(13)(24)	(234)	(123)	(142)	(143)	(14)(23)	(132)	(134)	е	(243)	(12)(34)	(124)
(243)	(12)(34)	(132)	(143)	(14)(23)	(142)	е	(234)	(123)	(124)	(134)	(13)(24)
Желтым фоном выделена таблица подгруппы 8 порядка. Данная подгруппа некоммутативна.											

Листинги программ

Листинг С.1: Вычисление последовательности Коллатца на языке Python.

```
while True:                # Работаем пока вводится целое число
    try:
        x = int(input("Введите стартовое целое число >0: "))
    except Exception:
        print("Введенно что-то не то! Выходим...")
        break
    if x<=0:
        print("Число неположительное, пропускаем")
        continue
    s = 0                    # количество шагов до 1
    pattern = '('            # плюсы и минусы
    while x>1:               # мы уверены, что не заиклимся :)
        pattern = pattern + '-' if x%2 == 0 else pattern + '+'
        x = x//2 if x%2 == 0 else 3*x+1
        s +=1
        print(x,end=" ") if x>1 else print(x)
    print("Количество итераций: ",s)
    print("Схема:", pattern+')')
```

Листинг С.2: Вычисление последовательности Гудстейна на языке Python.

```
def gudstein(x: int, n: int, m: int):
    """ Раскладывает число x по супероснованию n,
        затем подменяет его на m и вычисляет результат """
    if x < n:                # сразу отдаем x, т.к. он меньше основания
        return x
    seq = []                # массив коэффициентов перед степенями n
    i = 0                   # счетчик степеней основания
    while x > 0:
        if x % n > 0:
            seq.append((x % n, i)) # запоминаем остаток и степень
            x = x // n              # переходим к неполному частному
            i += 1
    s = 0
    for k, i in seq:         # заменяем n на m, рекурсия по степеням
        s += (k * (m ** gudstein(i, n, m)) if i > 0 else k)
    return s

while True:                # Работаем пока вводится целое число
    try:
        x = int(input("Введите стартовое целое число >0: "))
    except Exception:
        print("Введено что-то не то! Выходим...")
        break
    if x <= 0:
        print("Число неположительное, пропускаем")
        continue
    n = 2                   # стартуем с основания 2
    while x > 1 and x < 10 ** 1000 and n < 101: # ограничители
        x = gudstein(x, n, n + 1) - 1          # рекурсивный шаг
        print(x, " , шаг: ", n, sep=" ")
        n = n + 1
    if x > 1:
        if x >= 10 ** 1000:
            print("Число цифр стало больше 1000")
        else:
            print("Число итераций достигло ста")
```

```

import re
def subst_prod(s2: str, s1: str):
    """ Вычисляет произведение двух перестановок, записанных в
        скобочном виде.
        Элементы могут обозначаться цифрами или латинскими буквами.
        Например, (12)(ab)*(12a)=(2ba) (умножаем справа налево) """
    assert (re.match("^\([a-z0-9]+\)\)*e?$", s1) is not None
        and re.match("^\([a-z0-9]+\)\)*e?$", s2) is not None), \
        "Ошибка в записи подстановки!"
    if s1=='e':
        # умножение на единицу тривиально
        return s2
    if s2=='e':
        return s1
    Alphabet = sorted(set(s1+s2))[2:] # массив всех символов
    resplit = "\\(|\\)|\\" # регулярка для разбивки
    S1=re.split(resplit, s1)[1:-1] # разбиваем по скобкам на циклы
    S2=re.split(resplit, s2)[1:-1] # (12)(345) -> ['12', '345']
    assert len(''.join(S1)) == len(set(''.join(S1))), \
        "Ошибка в записи подстановки!"
    assert len(''.join(S2)) == len(set(''.join(S2))), \
        "Ошибка в записи подстановки!"
    S = [Alphabet[0]] # начальное состояние результата S=['1']
    Alphabet = Alphabet[1:] # удаление первого элемента алфавита
    while len(Alphabet)>0:
        Y = Next(Next(S[-1][-1], S1), S2) # следующий символ
        if Y not in S[-1]:
            S[-1] = S[-1]+Y # новый элемент в конец цикла
            Alphabet.remove(Y) # и удаляем его из алфавита
        elif S[-1] == Y:
            # цикл из одного элемента
            S[-1] = Alphabet[0] # вместо него начинаем новый
            Alphabet = Alphabet[1:] # и удаляем символ из алфавита
        else:
            # иначе строим следующий цикл
            S.append(Alphabet[0]) # помещаем туда очередной символ
            Alphabet = Alphabet[1:] # и удаляем его из алфавита
    if len(S[-1]) == 1:
        # если в конце остался
        # единичный цикл
        S = S[0:-1] # то удаляем его
    return ('('+'+'.join(S)+')').replace('()', 'e')
    # результат ['12', '345'] -> (12)(345), () -> e
def Next(Symbol, Subst):
    """ Возвращает символ, в который переходит Symbol под действием
        подстановки Subst """
    C=0 # счетчик циклов
    while C<len(Subst) and Symbol not in Subst[C]: # ищем цикл
        C+=1
    if C == len(Subst): # если символ не найден в подстановке
        return Symbol # возвращаем его же
    i=Subst[C].index(Symbol) # иначе: ищем позицию символа
    return Subst[C][i+1] if i<len(Subst[C])-1 else Subst[C][0]
    # берем следующий по циклу

```

```

while True:
    A = str(input("Введите первую подстановку (0 — чтобы выйти): "))
    if A=='0':
        break
    B = str(input("Введите вторую подстановку: "))
    print(subst_prod(A,B))

```

Листинг С.4: Самопечатающий код на Python.

```

q=chr(39)
code1 = [] # начальная часть кода
code2 = [] # алгоритм вывода
code1.append('q=chr(39)')
code1.append('code1 = [] # начальная часть кода')
code1.append('code2 = [] # алгоритм вывода')
code2.append('for s in code1: # вывод начальной части кода')
code2.append('    print(s)')
code2.append('for s in code1: # самовывод массива строк с кодом')
code2.append('    print("code1.append(",q,s,q,"",sep="")')
code2.append('for s in code2:')
code2.append('    print("code2.append(",q,s,q,"",sep="")')
code2.append('for s in code2: # вывод алгоритма вывода')
code2.append('    print(s)')
for s in code1: # вывод начальной части кода
    print(s)
for s in code1: # самовывод массива строк с кодом
    print("code1.append(",q,s,q,"",sep="")
for s in code2:
    print("code2.append(",q,s,q,"",sep="")
for s in code2: # вывод алгоритма вывода
    print(s)

```

```

""" Выполняет подбор слагаемых из заданного списка под
    определенную сумму. Список слагаемых берется из 1 колонки
    файла sum.csv. Ответом будет тот же файл, где
    во второй колонке написано Yes у выбранных слагаемых
"""
import csv
data = list(csv.reader(open("sum.csv"), delimiter=',')) # data
N = len(data)
k = 1 # коэффициент для сведения к целым числам
d = 0 # допуск отклонения вверх от нужной суммы
S = 127 # целевая сумма в рублях
Delta = int(round(d*k)) # допуск с учетом коэффициента
Ves = int(round(S*k))+Delta # целевая сумма алгоритма
Numbers = [0]*N # заготовка под вектор ответов
for i in range(N): # заполняем массив для алгоритма
    Numbers[i] = int(round(float(data[i][0].replace(",","."))*k))
T=[]
for i in range(N+1):
    T.append([0]*(Ves+1))
for i in range(N+1): # начальное состояние матрицы
    T[i][0]=1
for i in range(1,N+1): # алгоритм подбора
    for j in range(1,Ves+1):
        if j >= Numbers[i-1]:
            if T[i-1][j] > T[i-1][j-Numbers[i-1]]:
                T[i][j] = T[i-1][j]
            else:
                T[i][j] = T[i-1][j-Numbers[i-1]]
        else:
            T[i][j] = T[i-1][j]
dmin = Ves # поиск минимума
summ=0
for j in range(Ves,0,-1):
    if T[N][j] == 1:
        if dmin > abs(Ves-Delta-j):
            dmin = abs(Ves-Delta-j)
            summ=j
for i in range(N,0,-1): # вывод результатов работы в столбец 2
    if T[i][summ] == T[i-1][summ]:
        if len(data[i-1]) >= 2:
            data[i-1][1] = "No"
        else:
            data[i-1].append("No")
    else:
        if len(data[i-1]) >= 2:
            data[i-1][1] = "Yes"
        else:
            data[i-1].append("Yes")
        summ = summ - Numbers[i-1]
csv.writer(open("sum.csv", "w", newline='')).writerows(data)

```


СПИСОК ЛИТЕРАТУРЫ

Общематематические книги

- [1] Арнольд В. И. Гюйгенс и Барроу, Ньютон и Гук. — М.: Наука, 1989.
- [2] Грэхем Р., Кнут Д., Паташник О. Конкретная математика. — М.: Мир, 1998.
- [3] Клайн М. Математика. Утрата определенности. — М.: Мир, 1984.
- [4] Математическая составляющая / Редакторы-составители Н. Н. Андреев, С. П. Коновалов, Н. М. Панюнин; Художник-оформитель Р. А. Кокшаров. — М.: Фонд «Математические этюды», 2015.
- [5] Курант Р., Роббинс Г. Что такое математика? — Изд. 7-е., стереот. — М.: МЦНМО, 2015.
- [6] Проблемы Гильберта / под ред. П. С. Александрова — М., Наука, 1969.
- [7] Рид К. Гильберт. — М.: Наука, 1977.

Логика и Теория множеств

- [8] Беклемишев Л. Д. Введение в математическую логику. Конспект лекций. — М.: МГУ, 2008.
- [9] Беклемишев Л. Д. Теоремы Гёделя о неполноте и границы их применимости. // Успехи Математических Наук. — 2010. — Т.65, N5.
- [10] Бурбаки Н. Архитектура математики // Очерки по истории математики. — М.: ИИЛ, 1963. — С.245–259.
- [11] Верещагин Н. К., Шень А. Начала теории множеств. — М.: МЦНМО, 2012.
- [12] Верещагин Н. К., Шень А. Лекции по математической логике и теории алгоритмов. Часть 2. Языки и исчисления. — М.: МЦНМО, 2012.
- [13] Верещагин Н. К., Шень А. Лекции по математической логике и теории алгоритмов. Часть 3. Вычислимые функции. — М.: МЦНМО, 2012.

- [14] Гёдель К. Совместимость аксиомы выбора и обобщенной континуум-гипотезы с аксиомами теории множеств // Успехи мат. наук. — 1948. — Т.8, вып.1. — С.96–149.
- [15] Гильберт Д., Бернайс П. Основания математики. — в 2-х томах. — М.: Наука, 1979–1982.
- [16] Гудстейн Р. Л. Математическая логика. — М.: URSS, 2010.
- [17] Ершов Ю. Л. Σ -определимость и теорема Гёделя о неполноте: Учебное пособие. — Новосибирск: Научная книга, 1995.
- [18] Йех Т. Теория множеств и метод форсинга. — М. Мир, 1973.
- [19] Киселев А. А. Недостижимость и субнедостижимость, Часть I. — Минск.: Бел. гос. ун-т, 2011.
- [20] Киселев А. А. Недостижимость и субнедостижимость, Часть II. — Минск.: Бел. гос. ун-т, 2011.
- [21] Клини С. К. Математическая логика. — М.: Мир, 1973.
- [22] Кнут Д. Э. Сюрреальные числа / Перевод Н. Шихова. — М.: «Бином. Лаборатория знаний», 2014.
- [23] Колмогоров А. Н., Драгалин А. Г. Математическая логика. Дополнительные главы: Учеб. пособие. — М.: Изд-во Моск. ун-та, 1984.
- [24] Коэн П. Дж. Теория множеств и континуум-гипотеза. — М.: Мир, 1969.
- [25] Куратовский К., Мостовский А. Теория множеств. — М.: Мир, 1970.
- [26] Петровский А. Б. Пространства множеств и мультимножеств. — М.: Едиториал УРСС, 2003.
- [27] Пономарев И. Н. Введение в математическую логику и роды структур: Учебное пособие. — М.: МФТИ, 2007.
- [28] Справочная книга по математической логике в 4-х частях; под ред. Дж. Барвайса; Ч.2 Теория множеств. — М., Наука, 1982.
- [29] Alling, Norman L. Foundations of Analysis over Surreal Number Fields. // Mathematics Studies. — North-Holland, 1987. — 141.
- [30] Conway J. H. On numbers and games, second edition. — A. K. Peters, 2001.
- [31] Dushnik B. Miller E. W. Partially ordered sets // American Journal of Mathematics. — 1941. — Vol.63, N3. — P.600–610.

- [32] Ehrlich Philip. *The Absolute Arithmetic Continuum and the Unification of All Numbers Great and Small*. // The Bulletin of Symbolic Logic. — 2012 — V.18, N1. — P.1–45.
- [33] Gentzen G. Die Widerspruchsfreiheit der reinen Zahlentheorie // Mathematische Annalen. — 1936. — N112. — P.493–565.
- [34] Gonshor Harry. *An Introduction to the Theory of Surreal Numbers*. // London Mathematical Society, Lecture Note Series 110. — Cambridge University Press, 1986.
- [35] Henle James M. *An Outline of Set Theory*. — New York etc.: Springer-Verlag, 1986. Русская версия: Хенл Дж. М. Введение в теорию множеств: Пер. с англ. — М.: Радио и связь, 1993.
- [36] Jech T. J. *The Axiom of Choice*. — Amsterdam etc.: North-Holland, 1973.
- [37] Kirby L., Paris J. *Accessible independence results for Peano arithmetic* // Bulletin London Mathematical Society. — 1982. — V.14: P.285–293.
- [38] Levy A. *Basic Set Theory*. — Berlin etc.: Springer-Verlag, 1979.
- [39] Sierpiński, Waław, Cardinal and ordinal numbers. // Polska Akademia Nauk Monografie Matematyczne. — Warsaw: 1958 — N34. — Państwowe Wydawnictwo Naukowe, MR 0095787.
- [40] Schwichtenberg H., Wainer S. S. *Proofs and Computations Perspectives in Logic*. — Cambridge University Press, 2012.
- [41] Tøndering Claus *Surreal Numbers — An Introduction*, 2019.

Computer Science

- [42] Барендрегт Х. Ламбда-исчисление. Его синтаксис и семантика. — М.: Мир, 1985.
- [43] Воеводин В. В., Воеводин Вл. В. *Параллельные вычисления*. СПб.: БХВ-Петербург, 2002.
- [44] Воронцов К. В. *Лекции по методу опорных векторов* [Электронный ресурс]. — 2007.
- [45] Ершов Ю. Л. *Определимость и вычислимость. Сибирская школа алгебры и логики*. — Новосибирск: Научная книга, 1996; English transl., Ershov Yu. L. *Definability and computability, Siberian School of Algebra and Logic*. — New York: Consultants Bureau, 1996.

- [46] Кнут Д. Искусство программирования. Том 2. Получисленные алгоритмы. — В 4-х томах. Пер. с англ. — 3-е изд. — М.: Вильямс, 2007.
- [47] Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. — О’Reilly, 2017.
- [48] Орельен Ж. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow. Концепции, инструменты и техники для создания интеллектуальных систем. — О’Reilly, 2017.
- [49] Dongarra J. J., Duff L. S., Sorensen D. C., Vorst H. A. V. Numerical Linear Algebra for High-Performance Computers (Software, Environments, Tools) // Soc. for Industrial & Applied Math. — 1999.

Алгебра и Теория чисел

- [50] Айерленд К., Роузен М.. Классическое введение в современную теорию чисел. — М.: Мир, 1987.
- [51] Атлас представлений конечных групп [Электронный ресурс] — Режим доступа: <http://brauer.maths.qmul.ac.uk/Atlas/v3/>, свободный.
- [52] Артин Э. Теория Галуа. Пер. с англ. А. В. Самохина. — М.: МЦМНО, 2004.
- [53] Бурбаки Н. Группы и алгебры Ли. Главы I—III. — М.: Мир, 1976.
- [54] Бурбаки Н. Группы и алгебры Ли. Глава IX. — М.: Мир, 1986.
- [55] Ван дер Варден Б. Л. Алгебра. — М.: Наука, 1976.
- [56] Винберг Э. Б. Курс алгебры — М.: Факториал Пресс, 2001.
- [57] Городенцев А. Л. Алгебра: Учебник для студентов-математиков. — М.: факультет математики ВШЭ, 2011.
- [58] Корн Г., Корн Т. Алгебра матриц и матричное исчисление // Справочник по математике. — 4-е издание. — М: Наука, 1978.
- [59] Кострикин А. И. Введение в алгебру. — М. ФИЗМАТЛИТ, 2004.
- [60] Курош А. Г. Общая алгебра. — М.: Наука, 1974.
- [61] Ленг С. Алгебра. — М.: Наука, 1971.
- [62] Понтрягин Л. С. Обобщения чисел. — М.:Едиториал УРСС, 2018.

- [63] Постников М. М. Теория Галуа. — М.: Факториал Пресс, 2003.
- [64] Прасолов В. В. Многочлены. 4-е изд., испр. — М.: МЦНМО, 2014.
- [65] Хованский А. Г. Топологическая теория Галуа. Разрешимость и неразрешимость уравнений в конечном виде. — М.: МЦНМО, 2008.
- [66] Чашкин А. В., Жуков Д. А. **Элементы конечной алгебры**. — М.: Изд. МГТУ им.Баумана, 2016.
- [67] **Baffling ABC maths proof now has impenetrable 300-page ‘summary’** [Electronic Resource]
- [68] Gorenstein D., Lyons R., Solomon R. The classification of the finite simple groups. — Providence, R.I.: American Mathematical Society, 1994. — Vol.40.1. — (Mathematical Surveys and Monographs).
- [69] Gorenstein D., Lyons R., Solomon R. The classification of the finite simple groups. Number 2. Part I, chapter G: General group theory. — Providence, R.I.: American Mathematical Society, 1996. — Vol.40.2. — (Mathematical Surveys and Monographs).
- [70] Gorenstein D., Lyons R., Solomon R. The classification of the finite simple groups. Number 3. Part I, chapter A: Almost simple \mathcal{K} -groups. — Providence, R.I.: American Mathematical Society, 1998. — Vol.40.3. — (Mathematical Surveys and Monographs).
- [71] Gorenstein D., Lyons R., Solomon R. The classification of the finite simple groups. Number 4. Part II, chapters 1–4: Uniqueness theorems. — Providence, R.I.: American Mathematical Society, 1999. — Vol.40.4. — (Mathematical Surveys and Monographs).
- [72] Gorenstein D., Lyons R., Solomon R. The classification of the finite simple groups. Number 5. Part III, chapters 1–6: The generic case, stages 1–3a. — Providence, R.I.: American Mathematical Society, 2002. — Vol.40.5. — (Mathematical Surveys and Monographs).
- [73] Gorenstein D., Lyons R., Solomon R. The classification of the finite simple groups. Number 6. Part IV: The special odd case. — Providence, R.I.: American Mathematical Society, 2005. — Vol.40.6. — (Mathematical Surveys and Monographs).
- [74] Gorenstein D., Lyons R., Solomon R. The classification of the finite simple groups. Number 7. Part III, chapters 7–11: The generic case, stages 3b and 4a. — Providence, R.I.: American Mathematical Society, 2018. — Vol.40.7.

- [75] Reid M. *Galois Theory*. — University of Warwick, Coventry, 2014.
- [76] Solomon R. *A brief history of the classification of the finite simple groups* // American Mathematical Society. Bulletin. New Series. — 2001. — Т.38, вып.3. — С.315–352.

Анализ, Геометрия, Топология

- [77] Боровков А. А. Теория вероятностей: Учеб. пособие для вузов. — М.: Наука, 1986.
- [78] Гарасько Г. И., Кокарев С. С., Тришин В. Н., Балан В., Бринзей Н., Сипаров С. В., Чернов В. М., Панчелюга В. А. *Основы финслеровой геометрии и ее приложения в физике* // Материалы Международной школы-семинара для старшекурсников, аспирантов физико-математических факультетов и молодых ученых. — М.: МГТУ им.Н.Э.Баумана, 2010.
- [79] Гордон Е. И., Кусраев А. Г., Кутателадзе С. С. Инфинитезимальный анализ. — Новосибирск: Институт математики, 2006.
- [80] Домрин А. В., Сергеев А. Г. *Лекции по комплексному анализу. В 2 частях*. — М.: МИАН, 2004.
- [81] Колмогоров А. Н., Фомин С. В. Элементы теории функций и функционального анализа. — М.: Физматлит, 2009.
- [82] Крамер Г. Математические методы статистики. — М.: Мир, 1975.
- [83] Ландау Л. Д., Лифшиц Е. М. Квантовая механика (нерелятивистская теория). — Издание 4-е. — М.: Наука, 1989.
- [84] Масси У., Столлингс Дж. Алгебраическая топология. Введение. — М.: Мир, 1977. *English version*: Massey W. Algebraic Topology: An Introduction. 1967; Stallings J. Group Theory and Three-Dimensional Manifolds. 1971.
- [85] Павлов Д. Г., Кокарев С. С. Алгебра, геометрия и физика двойных чисел. // Сб. трудов РНОЦ "Логос". — Вып.9. — 2014. — С.7–124.
- [86] Плахов А. Ю. *Рассеяние в билиардах и задачи ньютоновской аэродинамики* // Успехи математических наук. — 2009. — Т.64. Вып.5 (389). — С. 97–166.
- [87] Прасолов В. В. *Геометрия Лобачевского*. — М.: МЦНМО, 2016.

- [88] Прасолов В. В., Тихомиров В. М. **Геометрия**. — М.: МЦНМО, 2007.
- [89] Сергеев А. Г. **Введение в некоммутативную геометрию**. [Электронный ресурс] — 2016.
- [90] Сипаров С. В. **Введение в анизотропную геометродинамику**. Нью Джерси — Лондон — Сингапур, World Scientific, 2011.
- [91] Соболев В. И. Лекции по дополнительным главам математического анализа. — М.: Наука, 1968
- [92] Сосов Е. Н. **Геометрия Лобачевского и ее применение в специальной теории относительности**: учеб.-метод. пособие. — Казань: Казан. ун-т, 2016.
- [93] Успенский В. А. Что такое нестандартный анализ. — М.: Наука, 1987.
- [94] Фейнман Р. Лейтон Р. Сэндс М. Фейнмановские лекции по физике. Выпуск 8,9. Квантовая механика. Учебное пособие. — М.: Либроком, 2014.
- [95] Фоменко А. Т. Фукс Д. Б. Курс гомотопической топологии: Учеб. пособие для вузов. — М.: Наука, 1989.
- [96] Хелстром К. Квантовая теория проверки гипотез и оценивания. — М.: Мир, 1979.
- [97] Чеботарев А. М. **Введение в теорию вероятностей и математическую статистику для физиков**. — М., МФТИ, 2008.
- [98] Шашкин Ю. А. Неподвижные точки. — М.: Наука, 1989.
- [99] Эльсгольц Л. Э. Вариационное исчисление: Учебник. — М.: Издательство ЛКИ, 2019.
- [100] Alexandroff A. D. Additive set-functions in abstract spaces I // Матем. сборник 1940. — V.8(50), N 2. P.307-348.
- [101] Alexandroff A. D. Additive set-functions in abstract spaces II // Матем. сборник 1941. — V.9(51), N 3. P.563-628.
- [102] Alexandroff A. D. Additive set-functions in abstract spaces III // Матем. сборник 1943. — V.13(55), N 2. P.169-293.
- [103] Bejancu A., Farran H. R. Geometry of Pseudo-Finsler Submanifolds, — Kluwer, Dordrecht, 2000.
- [104] Connes. A. **Noncommutative Geometry**. [Electronic Resource]

- [105] Engelking R. General Topology. — Warszawa.: PWN, 1977. *Русское издание:* Энгелькинг Р. Общая топология. — М.: Мир, 1986.
- [106] Kanovei V., Reeken M. Nonstandard Analysis, Axiomatically. — Berlin: Springer-Verl., 2004.
- [107] Makarios T. J. M. *A further simplification of Tarski's axioms of geometry.* [Electronic Resource] — 2013.
- [108] Nelson E. *Books on Edward Nelson's Home Page on the Princeton University site* [Electronic Resource]
- [109] Nica E.. *The Mazur-Ulam Theorem.* — Göttingen: Mathematisches Institut, Georg-August-Universität, 2013.
- [110] Polyanin A. D., Manzhirov A. V., Handbook of Integral Equations. — CRC Press, Boca Raton, 1998.

Графы и ветвящиеся процессы

- [111] Казимиров Н. И. *Возникновение гигантской компоненты в случайной подстановке с известным числом циклов.* // Дискрет. матем. — 2003. — Т.15, Вып.3. — С.145–159.
- [112] Казимиров Н. И. *Леса Гальтона–Ватсона и случайные подстановки:* дис. ... канд.физ.-мат.наук.: 01.01.09; Институт прикладных математических исследований Карельского научного центра РАН. — Петрозаводск, 2003. — 127 с. (*Автореферат доступен по ссылке*)
- [113] Ландо С. К. *Графы и топология.* [Электронный ресурс] — 2018.
- [114] Севастьянов Б. А. Ветвящиеся процессы. — М.: Наука, 1971.
- [115] Харари Ф. Теория графов. — М.: Мир, 1973.
- [116] Bollobás, B. Random Graphs. — Cambridge University Press, 2001.
- [117] Erdős, P. and Rényi, A. *On Random Graphs.* // Publicationes Mathematicae. — 1959. — **6**. — P.290-297.
- [118] Erdős, P. and Rényi, A. On the Evolution of Random Graphs. // Publ. Math. Inst. Hungar. Acad. Sci. — 1960. — **5**. — P.17-61.
- [119] Gilbert E. N. *Random Graphs.* // *Annals of Mathematical Statistics.* — 30 (4). — P.1141–1144.

- [120] Harary F., Palmer E. M. **Graphical Enumeration** — New York, Academic Press, 1973.
- [121] Kazimirow N. **On some estimates for Erdős-Rényi random graph**. [Electronic Resource] — 2015.
- [122] Kolchin V. F. **Random Graphs**. — New York: Cambridge University Press, 1998.
- [123] Meir A., Moon J. W. **On the altitude of nodes in random trees** // Can. J. Math. — 1978. — 30, N5. — P.997–1015.
- [124] Pavlov Yu. L. **Random Forests**. — Utrecht: VSP, 2000.
- [125] Pitman J. **Enumerations of trees and forests related to branching processes and random walks**. // In. D. Aldous and J. Propp, editors, **Microsurveys in Discrete Probability**. №41 in DIMACS Ser. Discrete Math. Theoret. Comp. — Sci. Providence RI, Amer. Math. Soc., 1998. P.163–180.
- [126] Polya G. **Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen**. // Acta Math. — 1937. — 68. — P.145–254.

НАУЧНО-ПОПУЛЯРНОЕ ИЗДАНИЕ
POPULAR SCIENCE EDITION

НИКОЛАЙ ИГОРЕВИЧ КАЗИМИРОВ

АРХЕТИПЫ МАТЕМАТИКИ

общие методы, приемы, конструкции, идеи математики и ее оснований

N. I. KAZIMIROV

ARCHETYPES OF MATHEMATICS

general methods, techniques, constructions, ideas of mathematics and its foundations

ИЗДАТЕЛЬСТВО «ЮСТИЦИНФОРМ»

юридическая, экономическая и деловая литература;
журналы «Право и экономика», «Вестник арбитражной практики»,
«Журнал предпринимательского и корпоративного права»

«YUSTITSINFORM» PUBLISHING HOUSE

legal, economic and business literature
magazines «Law and Economics», «Bulletin of arbitration practice»,
«Journal of Entrepreneurship and Corporate Law»

Главный редактор
В. А. Вайпан
Chief editor V. A. Vaypan

Генеральный директор
В. В. Прошин
CEO V. V. Proshin

Оригинал-макет подготовлен с использованием макро-пакета $\mathcal{A}\mathcal{M}\mathcal{S}\text{-}\mathcal{L}\mathcal{A}\mathcal{T}\mathcal{E}\mathcal{X}$

This publication was typeset by $\mathcal{A}\mathcal{M}\mathcal{S}\text{-}\mathcal{L}\mathcal{A}\mathcal{T}\mathcal{E}\mathcal{X}$

Подписано в печать 05.12.2019.

Формат 70x100/16. Бумага офсетная. Печ. л. 38.25.

Тираж 30 шт.

Signed in print 05.12.2019.

Format 70x100/16. Offset paper. Print. 38.25. 30 copies.

Юстицинформ

119607, г. Москва, ул. Лобачевского, 94, оф. 7.

Тел.: (495) 232-12-42

Yustitsinform

119607, Moscow, Lobachevskogo street, 94, office 7.

Phone number: (495) 232-12-42

<http://www.jusinf.ru>

E-mail: info@jusinf.ru

Н. И. Казимиров

Архетипы математики

общие методы, приемы,
конструкции, идеи
математики и ее оснований

В книге на примере изучения логики, теории множеств, алгебры, геометрии, анализа, теории графов рассматриваются некоторые общие характерные приемы и методы конструирования математических объектов и теорий, так называемые **архетипы**. Даются их наименования, описание и ссылки на них по ходу изложения материала. Сам материал представлен как на наивном, так и на современном строгом уровне, ряд теорем приводится с доказательством, подробно рассматривается построение конечных множеств и логики первого порядка с целью донести до широкого круга читателей возможность компьютерного моделирования математических доказательств. Вместе с тем, делается акцент на необходимости участия человека в этом процессе. В книге также дается ряд примеров из теории чисел, в частности, доказательство Великой Теоремы Ферма для случая $n=3, 4$, подробно рассмотрена теорема Гудстейна. Кроме того, проводится детальное описание движений в геометрии, всесторонне рассматривается аксиома выбора и ее влияние на математику, приводится ряд примеров из теории случайных графов с анализом их свойств.

Официальный сайт книги: <https://mathem.at>
