С.Ленг

МАТЕМАТИЧЕСКИЕ БЕСЕДЫ ДЛЯ СТУДЕНТОВ

Книга крупнейшего алгебраиста современности С.Ленга написана неформальным языком и носит характер бесед со студентами, записанных во время многочисленных посещений Ленга университетов Европы и Америки по их приглашениям. В ней в увлекательной форме рассказывается о классических и современных (в том числе и нерешенных) проблемах математики. Каждая тема завершается списком литературы, с помощью которого читатель сможет самостоятельно ознакомиться с задачей.

Будет полезна преподавателям, студентам, интересующимся математикой.

Содержание

Предисловие

Много лет я проводил беседы со студентами по избранным вопросам математики, которые можно изложить на уровне, понятном для студентов, изучивших курс математического анализа. Я проводил подобные беседы в классах Хирцебруха в Бонне, по приглашению Шломьюка в университете Монреаля, несколько раз в Высшей Нормальной Школе в Париже лет десять назад, позднее в Университете им. Гумбольта в Берлине по приглашению Крамера и регулярно в течение 1980-х по приглашению Вустгольца в ЕТН в Цюрихе. Я благодарен У. Штамбаху, который перевел Цюрихские лекции на немецкий язык для публикации в журнале «Элементы Математики» с 1992 по 1997 г. Последний раз я выступал с такими беседами в ассоциации студентов-математиков в Беркли летом 1998 г. и в Университете Техаса в Остине в ноябре 1998 г.

Теперь я предлагаю вам сборник бесед со студентами в виде книги. Подобные беседы могут проводиться преподавателями, но, что еще лучше, они могут проводиться студентами на семинарах, организованных самими студентами. Такие беседы проводились, например, в Йейльском университете.

В Соединенных Штатах необходимо ликвидировать математическую безграмотность значительной группы студентов и также необходимо улучшить математические знания школьных учителей и научить их эффективным способам помощи студентам при изучении математики на всех уровнях. Необходимо подчеркнуть различие между обучением в начальной школе и обучением в старших классах. Ситуация зависит от многих факторов. Кто-то из людей, причастных к преподаванию в начальной школе, описал мне, какой урон математическому образованию и математическим знаниям уже нанесен к тому времени, когда школьники приходят в старшие классы. Разумеется, я говорю об общем положении дел, частные случаи необходимо оценивать по их собственным меркам.

Под давлением некоторых правительственных и социальных сил и из-за вышеупомянутых необходимостей имели место некоторые перегибы, что привело к понижению стандартов и к тому, что жаргон стал необходимым условием получения финансирования для различных образовательных проектов. Более десятилетия назад возникло выражение, что математика должна быть «насосом, а не фильтром». Несколько лет назад в программе, предназначенной для стимулирования математически одаренных школьников старших классов, Дэвид Рорлич собрал материал по p-адическим L-функциям, который является понятным школьникам выпускных классов. В то время я задавал вопрос: p-адические L-функции — это насос или фильтр? Mutatis mutandis, тот же вопрос можно задать и о темах, выбранных для этой книги. Гипотеза Римана — это насос или фильтр? Закон полупараллелограмма — это насос или фильтр? Изучение гармонических функций — это насос или фильтр?

Выражение «насос, а не фильтр» — это заголовок раздела (стр. 6) публикации NAS (Национальной академии наук) «Everybody Counts» (Лишних людей нет). Эта фраза изначально принадлежала Роберту М. Уайту, который в то время был президентом Национальной академии инженерных наук. Он сказал: «Математика должна быть насосом, а не фильтром в трубопроводе американского образования». Группа учителей из Масачусетса однажды попросили меня присоединиться к проекту, связанному с математическим образованием. Пояснительная записка к проекту была озаглавлена «Связанная геометрия — усиление насоса» и была написана соответствующим жаргоном. Я спросил организаторов, обязаны ли они писать таким образом, чтобы получить финансирование, и ответ по телефону был «да». Я отказался присоединиться. Это не означает, что я не заинтересован в математическом образовании на всех уровнях. Смотрите мою книгу «Математика: встречи со школьниками старших классов», а также книгу «Геометрия: курс для старшеклассников» (совместно с Джином Мюрроу). Что касается книги «Геометрия», я однажды получил письмо от учителя (а также руководителя факультета) в школе для девочек. Она написала: «Прилагаются письма, написанные некоторыми ученицами девятого класса, изучавшими курс геометрии. Им хотелось бы встретиться с автором их учебника и, представившись вам, они приглашают вас посетить занятия, если вы окажетесь в ... местности.

Первый год я преподаю геометрию по вашему учебнику, и, как подтверждают некоторые девушки, он довольно сложный. Я нахожу ваш курс освежающим. Вы убедили меня перейти к преподаванию доказательств в «параграфном» стиле. Ваши конструкции хорошо адаптируются к программе «Геометрический чертежный стол», которую мы часто используем. Язык книги строгий, но ясный и исчерпывающий. Учебник превосходно подходит для нашего высокого уровня занятий геометрией».

К ее письму были приложены письма от девушек из этого класса.

Некоторые политики и, как результат, Национальный Фонд Науки явно выражают точку зрения, что денег достаточно лишь для финансирования исследований, которые имеют прямые приложения, а не исследований «по причине любопытства». Я возражаю. Такая риторика смешивает вместе несколько вопросов, которые заслуживают того, чтобы этот узел был распутан. Среди немедленных приложений — необходимость иметь компетентные кадры преподавателей математики на всех уровнях. Разумеется, многие учителя компетентны, но также широко распространена некомпетентность, в дополнение к наличию нескольких систематически неблагополучных округов. Также существует возможность применения математики в инженерном деле, биологии, экологии, экономике, что относится к несколько другой области. Куда идут деньги, является по закону политическим решением. Но противодействие исследованиям «из любопытства», как будто такие исследования бесполезны (по каким-нибудь подсчетам), не желательно. По историческим обзорам любопытство привело к замечательным открытиям всех типов, и некоторые наиболее развитые общества в странах от Китая и Индии до Северной Африки и Европы (Соединенные Штаты часто становятся исключением) считают научное любопытство культурной ценностью. В большинстве своем открытия нашли так называемые приложения, иногда сразу, иногда десятилетиями или столетиями позже. В любом случае размышления о математике являются удовольствием для многих людей, и существуют свидетельства, что мы естественным образом запрограммированы любить математику, пока удовольствие не будет разрушено некомпетентным преподаванием или другими социальными факторами. Любой пятилетний ребенок, с которым я встречался, любил складывать или вычитать числа. Что происходит дальше, это нечто другое. Это нечто другое может также включать учительницу и девушек на ее уроке геометрии, упомянутом выше, с их любовью к геометрии и с их энтузиазмом. Повторяю, частные случаи должны оцениваться каждый по своей собственной мерке. Сколько детей были достаточно удачливы, чтобы попасть в такой класс?

Чиновники в правительстве, в университетах внедряют жаргон, включающий насосы и фильтры или подобные выражения, подтверждающие, что образование и наука подчинены сегодняшней социальной и политической ортодоксии. Последнее добавление к этому жаргону — это терминология «вертикально интегрированных» исследований и преподавания и VIGRE[1] гранты, «VI» расшифровывается как «вертикально интегрированный». А беседы в этой книге вертикально интегрированны? Или горизонтально? Или под каким-нибудь другим углом? Кто над кем смеется? Я бы хотел, чтобы все это было шуткой: в действительности это относится к финансированию.

Так что не обращайте внимание на насосы, фильтры и вертикальную интеграцию. Я надеюсь, что студенты или даже некоторые школьники старших классов получат удовольствие от этих бесед или, возможно, от некоторых из них. Разные студенты с разным уровнем математических знаний интересуются разными вопросами, в разное время и для разных целей. Даже среди бесед, предложенных в этой книге, студенты с разным уровнем математических знаний найдут некоторые беседы легкими, а другие более сложными или непонятными. Поэтому выбирайте. Вам придется поработать головой, и это удержит вас дома.

Нью Хавен 1998 *Серж*

---

[1]VIRGE — финансирование (грант) вертикально интегрированного проекта, связанного с исследованиями или с образованием. — *Прим. ред.*

Простые числа

Я не знаю другого раздела математики, отличного от теории чисел, в котором можно провести математическую беседу по крупнейшим нерешенным задачам, понятную людям, практически не имеющим математического образования. Школьники старших классов, знакомые с основами математического анализа, в состоянии понять большую часть этой беседы о простых числах. Эта глава написана по материалам бесед, которые я провел с несколькими группами слушателей.

Здесь и в некоторых других главах я частично сохранил стиль оригинальной беседы, когда студенты свободно задавали вопросы или прерывали рассказчика. Моя книга «Математика: встречи со школьниками старших классов» (издательство Шпрингер) была записана прямо с магнитофонных лент и поэтому полностью выдержана в разговорном стиле. Халмош когда-то охарактеризовал такой стиль как «вульгарный» и запретил публикацию отрывков в журнале «Математический ежемесячник» (Math Monthly). Десятилетие спустя, в 1990-х, эта беседа была снова предложена для публикации в «Математический ежемесячник» и была отвергнута редактором (на этот раз Роджером Хорном) из-за разговорного стиля. Но мне нравится разговорный стиль, и я обнаружил, что он весьма эффективен. Решите это для себя.

Простые числа

Кто знает, что такое простое число?

[*Разные студенты из аудитории дают определение.*]

Хорошо, давайте я запишу определение. Простое число p это целое число $\geqslant 2$, такое, что делители p — либо 1, либо само p. По соглашению мы берем $p \geqslant 2$. Принято считать, что 1 не является простым числом. Так какова же последовательность простых чисел? Они следующие:

$$2,\ 3,\ 5,\ 7,\ 11,\ 13,\ 17,\ 19,\ 23,\ \ldots,\ \text{и т. д.}$$

Что означает «и т. д.»? Будет ли последовательность простых чисел продолжаться бесконечно или она остановится? Другими словами: существует ли бесконечное количество простых чисел или существует только конечное количество простых чисел? Ответ на этот вопрос известен со времен Евклида.

Теорема. *Простых чисел бесконечно много.*

Вы знаете, как ее доказать?

[*Некоторые студенты сказали, что доказательство им неизвестно. Один из студентов привел доказательство.*]

Да, это доказательство, и я должен его записать. Пусть $2, 3, \ldots, p$ — простые числа, последовательно перечисленные до некоторого простого числа p. Теперь нам нужно показать, что существует еще одно простое число, отличающееся от $2, 3, \ldots, p$. Если мы сможем показать это, тем самым мы покажем, что существует бесконечно много простых чисел. Поэтому обозначим

$$N = (2 \cdot 3 \cdot 5 \cdots p) + 1.$$

Если N — простое, тогда N больше, чем p, и мы нашли еще одно простое число. Если N не простое число, то что происходит? В этом случае N делится на какое-то простое число. Действительно, N раскладывается на множители как произведение двух меньших положительных целых чисел. Если одно из них простое, то мы уже нашли простое число q, делящее N. Если нет, мы можем раскладывать на множители эти целые числа. Таким образом, вы продолжаете разложение на множители и вы не сможете продолжить этот процесс до бесконечности, т. к. множители становятся меньше. Поэтому на каком-то шаге вы получите простой делитель q числа N. Теперь наше утверждение состоит в том, что простой делитель q числа N должен отличаться от чисел $2, 3, \ldots, p$. Почему это так?

[*Некоторые студенты в аудитории дают ответ.*]

Да, если вы разделите N на любое из простых чисел $2, 3, \ldots, p$, то получите остаток 1, а q — простое число, которое делит N без остатка и поэтому q не равно никакому из простых чисел $2, 3, \ldots, p$. Таким образом мы доказали, что существует бесконечно много простых чисел.

Это доказательство составлено во времена Евклида.

Подсчеты

Мы хотим сосчитать, сколько существует простых чисел. Так как их бесконечно много, нам следует уточнить, что мы имеем в виду. Пусть имеется положительное число x. Обозначим

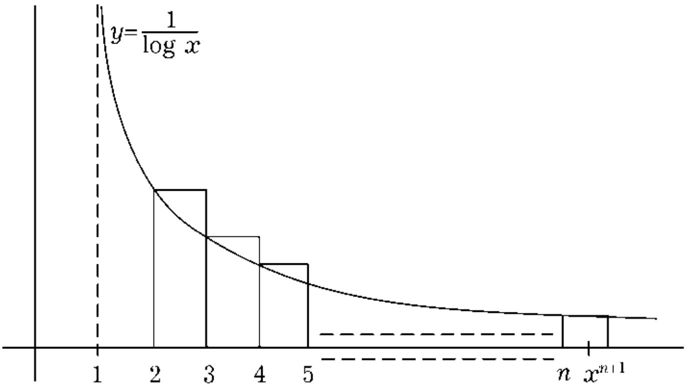
$$\pi(x) = \text{ количество простых чисел, которые } \leqslant x.$$

Мы хотим оценить число $\pi(x)$. Такая оценка зависит от вероятности. Общая философия, впервые появившаяся как предположение Гаусса, состоит в том, что вероятность того, что n — простое число, равна $\frac{1}{\log n}$. Гаусс написал это в письме к астроному Энке в 1849 г., и также написал, что он обнаружил эту формулу в 1792–1793 годах, когда ему было всего 15 или 16 лет. Гипотеза Гаусса — одна из основных в теории простых чисел. Что она означает? Если вам дано целое число n, то оно либо простое, либо не простое. Нам следует объяснить, что мы имеем в виду, когда говорим о вероятности $\frac{1}{\log n}$. Сам Гаусс объясняет это. Если вы возьмете общее количество простых чисел меньших, либо равных x, $\pi(x)$, тогда $\pi(x)$ должна быть приближенно суммой вероятностей

$$\pi(x) = \sum_{2 \leqslant n \leqslant x} \frac{1}{\log n} + E_1(x),$$

где $E_1(x)$ — некоторая ошибка, которая мала по отношению к величине суммы. Чем меньше величина ошибки, тем лучше аппроксимация $\pi(x)$ этой суммой. Чтобы учесть все детали, мы должны оценить величину ошибки, чтобы сделать ее как можно меньше по сравнению с $\pi(x)$. Мы скоро получим такую оценку, но сначала я хотел бы показать, что при использовании математического анализа сумма может быть записана другим образом. Пропустите следующие рассуждения, если вы не знаете математического анализа.

Вместо того, чтобы записывать сумму $\frac{1}{\log n}$, попытаемся использовать какое-нибудь другое обозначение. Чему приближенно равна такая сумма? Такая сумма должна напомнить вам о чем-то, что вы уже видели: об интеграле. Вам должен быть знаком следующий рисунок:



Вам нужно осознать, что эта сумма является римановой суммой для интеграла, поэтому сумма приближенно равна интегралу функции $\frac{1}{\log n}$, и мы можем записать

$$\pi(x) = \int_2^x \frac{1}{\log t}\,dt + E_2(x),$$

где $E_2(x)$ — другое выражение для ошибки аппроксимации, но $E_2(x)$ не слишком отличается от $E_1(x)$. Действительно, используя математический анализ, вы можете получить оценку

$$\left|\int_2^x \frac{1}{\log t}\,dt - \sum_{2\leqslant n\leqslant x} \frac{1}{\log n}\right| \leqslant 2.$$

Таким образом, мы можем оценить $\pi(x)$, используя интеграл или сумму. В некоторых случаях удобнее использовать интеграл, потому что интегрируя по частям, получим

$$\int_2^x \frac{1}{\log t}\,dt = \frac{x}{\log x} + \int_2^x \frac{1}{(\log t)^2}\,dt - \frac{2}{\log 2}.$$

Итак, в правой части вы интегрируете то, что намного меньше, чем $\frac{1}{\log t}$, а именно, вы интегрируете

$$\frac{1}{\log t}\frac{1}{\log t},$$

что стремится к 0 быстрее, чем $\frac{1}{\log t}$ при $t \to \infty$. Это означает, что наше выражение для $\pi(x)$ может быть записано

$$\pi(x) = \frac{x}{\log x} + E_3(x)$$

с ошибкой $E_3(x)$ такой, что

$$\lim_{x\to\infty} \frac{E_3(x)}{x/\log x} = 0.$$

Значит, $\frac{x}{\log x}$ является аппроксимацией первого порядка функции $\pi(x)$. Но $\frac{x}{\log x}$ не такая уж хорошая аппроксимация $\pi(x)$, поскольку вместо вычисления всей суммы вы просто взяли общее количество целых чисел от 2 до x, которое, в сущности, и есть x, и умножили на наибольшую вероятность, которая равна $\frac{1}{\log x}$. Вы сделали то, что не так уж хорошо. Вычисление суммы вероятностей или интеграла должно быть лучше, поскольку поделив x на $\log x$, вы усреднили слишком сильно. Но выражение $\frac{x}{\log x}$ имеет одно преимущество: это приятная формула в замкнутом виде с простыми функциями x и $\log x$. Эта формула дает грубый ответ на наш первый вопрос. Ответ был доказан в конце XIX столетия.

Теорема о простых числах (Адамар, де ла Валле–Пуссен, 1896).

$$\pi(x) \sim \frac{x}{\log x} \quad \textit{при} \quad x \to \infty.$$

Знак подобия тильда $\sim$ здесь имеет точный смысл. Если f и g — функции вещественного переменного x, то записываем

$$f(x) \sim g(x) \quad \text{при} \quad x \to \infty,$$

имея в виду, что

$$\lim_{x\to\infty} f(x)/g(x) = 1.$$

В этом случае мы говорим, что $f(x)$ *асимптотически эквивалентна* $g(x)$ при $x \to \infty$. Таким образом, $\pi(x)$ асимптотически эквивалентна $\frac{x}{\log x}$.

Теорема о простых числах дает первую аппроксимацию того, сколько всего простых чисел. Доказательство, даже сегодня, не такое простое. Если вы изучили курс комплексного анализа (вычеты, интегрирование по контуру), тогда вы сможете понять доказательство, и это может занять два часа, что не так уж плохо. Большие теоремы комплексного анализа (теорема Римана об отображении, к примеру) требуют примерно столько же времени, чтобы записать доказательство. На сегодняшний день самое короткое доказательство, в сущности, доказательство Ньюмана [New 80] (см. последнюю главу книги [La 93]).

Гипотеза Римана

Мы пришли к основному вопросу: какова ошибка аппроксимации? С точки зрения вероятности, предположим, что у вас случайное распределение. Когда вы вычисляете сумму, аппроксимирующую $\pi(x)$,

$$S(x) = \sum_{2\leqslant n\leqslant x} \frac{1}{\log n}, \quad \text{т. е.} \quad \pi(x) = S(x) + E_1(x),$$

или интеграл

$$\pi(x) = \int_2^x \frac{1}{\log t}\,dt + E(x) = \mathrm{Li}(x) + E(x),$$

где $\mathrm{Li}(x)$ — обозначение для интеграла

$$\mathrm{Li}(x) = \int_2^x \frac{dt}{\log t}$$

и $E(x)$ — ошибка аппроксимации, то какова должна быть оценка для величины ошибки $|E(x)|$? Гипотеза в вероятностной модели: существует константа C такая, что

$$|E(x)| \leqslant C(x^{1/2}\log x)$$

при всех достаточно больших x. Т. к. основное слагаемое равно $\frac{x}{\log x}$ и ошибка ограничена по абсолютной величине значением $C(x^{1/2}\log x)$, то ошибка достаточно мала по сравнению с $\frac{x}{\log x}$. Достаточно легко увидеть, что если константа C существует, тогда она мала.

Я благодарю Хуга Монтгомери за то, что он указал мне, что существует несколько публикаций по теории чисел, посвященных этой константе. Он особенно ссылался на [Scho 76], где было доказано, что при $x > 2657$ в предположении существования вышеупомянутой константы C выполнено неравенство

$$|E(x)| = |\pi(x) - \mathrm{Li}(x)| \leqslant \frac{1}{8\pi}x^{1/2}\log x.$$

Значения при $x \leqslant 2657$ могут быть легко вычислены и сведены в таблицу, и вы должны сделать это, чтобы почувствовать, насколько хороша аппроксимация $\pi(x)$ интегралом $\mathrm{Li}(x)$. Оценка $|E(x)| \leqslant C(x^{1/2}\log x)$ — это *гипотеза Римана*, которая по всеобщему соглашению является величайшей из математических гипотез. Таким образом, задача оценки количества простых чисел сводится к тому, чтобы придать точный смысл вероятностной модели, как можно точнее описывая ошибку аппроксимации.

В вероятностных терминах можно сказать, что гипотеза Римана утверждает, что последовательность простых чисел ведет себя как случайная последовательность в вероятностной модели, в которой вероятность того, что n — простое, равна $1/\log n$.

Также существуют другие феномены в теории чисел (и в других областях), которые могут быть выражены аналогичным образом. Задаем вероятностную модель и по закону больших чисел получаем величину ошибки для случайной последовательности. Тогда возможна гипотеза, что в данном частном случае данная последовательность, такая, как последовательность простых чисел, ведет себя как случайная последовательность этой модели. (См. [LaT 76] для примеров таких вероятностных моделей в других областях теории чисел.)

Мы сформулировали гипотезу Римана, не зная ничего, кроме определения простого числа и $\log x$.

Комментарий из аудитории. Обычно гипотеза Римана формулируется по-другому.

Серж Ленг. Вы правы, но обычная формулировка требует более глубокого знания математики для понимания и более сложна, т. к. использует комплексный анализ. Т. к. вы упомянули ее, я сделаю несколько комментариев по этому поводу. Эйлер ввел определение ζ-функции:

$$\zeta(s) = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1},$$

где произведение берется по всем простым числам и s — вещественное, большее 1. Риман показал, что существует аналитическая функция комплексного переменного, которая принимает значения Эйлера при вещественных $s > 1$ и однозначно определена по этим значениям. Оригинальная гипотеза Римана утверждает, что все нули $\zeta(s)$ при $0 < \mathrm{Re}(s) < 1$ лежат на прямой $\frac{1}{2} + it$, т. е. на прямой $\mathrm{Re}(s) = \frac{1}{2}$. Если вы сформулируете гипотезу Римана этим способом, то потребуется примерно два часа (после изучения курса комплексного анализа), чтобы показать, что эта формулировка эквивалентна той, которую мы дали в терминах оценки для ошибки аппроксимации $E(x)$.

Упоминание о комплексном анализе дает нам возможность прокомментировать пути математического прогресса. Адамар разработал определенные части комплексного анализа, движимый именно желанием доказать теорему о простых числах, заключающуюся в асимптотическом отношении $\pi(x) \sim \dfrac{x}{\log x}$. Таким образом, задача из чистой теории чисел привела к подъему в других теориях, которые используются в большинстве других разделов математики, а также в физике, инженерном деле и в других науках.

Условия на простые числа

Теперь мы оставим задачу оценки ошибки аппроксимации и обсудим задачу подсчета простых чисел при наложении дополнительных условий. Давайте начнем с некоторых примеров. В последовательности простых чисел имеются пары $(3,5)$, $(5,7)$, $(11,13)$, $(17,19)$, для которых

$$
\begin{aligned}
5 &= 3+2;\\
7 &= 5+2;\\
13 &= 11+2;\\
19 &= 17+2.
\end{aligned}
$$

Если имеется пара $(p, p+2)$ такая, что p и $p+2$ — простые числа, будем называть эти простые числа *двойными простыми* числами (twin primes). Какие из них следуют после $(17,19)$?

[*Несколько студентов произносят ответ*:]

$$(29, 31), \quad (41, 43), \quad (59, 61), \quad (71, 73), \quad \ldots$$

И каков же наш следующий вопрос?

Вопрос. Бесконечно ли много двойных простых чисел?

Кто скажет «Да»? [*Несколько рук поднято.*]

Кто скажет «Нет»? [*Меньше рук поднято.*]

[*Некоторые студенты начинают доказывать свои утверждения.*]

Я не спрашивал вас, как это доказать, я спросил, думаете ли вы, что существует бесконечно много двойных простых чисел или нет. Это другой вопрос. В начале, когда речь шла о самих простых числах, я спрашивал вас, думаете ли вы, что простых чисел бесконечно много. Тогда все вы считали, что ответ «Да», но вы не знали, как это доказать. Поэтому здесь два вопроса. Один из них: «Считаете ли вы, что таких пар простых чисел бесконечно много?», а другой «Можете ли вы это доказать?» Сейчас мы отвечаем на первый вопрос. Как вы считаете, двойных простых чисел бесконечно много или нет? Считает ли кто-нибудь, что их не бесконечно много, что их только конечное число?

[*Аудитория хранит молчание.*]

Что, слишком опасно высказываться? Или вы размышляете? Хорошо, я скажу вам ответ. Предположительно ответ «да», но доказательство неизвестно. Таково положение дел на сегодня: неизвестно, действительно ли бесконечно много двойных простых чисел. Это одна из знаменитых математических задач. Если вы решите ее, то попадете на страницы учебников истории.

Давайте рассмотрим еще одну задачу. Возьмем к примеру $n^2 + 1$ и спросим, простое ли это число? К примеру:

$$\begin{aligned} 2^2 + 1 &= 5 \text{ — простое;}\\ 4^2 + 1 &= 17 \text{ — простое;}\\ 6^2 + 1 &= 37 \text{ — простое;}\\ 8^2 + 1 &= 65 \text{ — не простое;}\\ 10^2 + 1 &= 101 \text{ — простое.} \end{aligned}$$

За исключением числа 65, все эти числа простые. Так какой же следующий вопрос я задам?

[*Некоторые студенты отвечают.*]

Да, следующий вопрос таков: бесконечно ли много простых чисел типа $n^2 + 1$?

Кто из вас думает, что их бесконечно много? [*Несколько студентов поднимают руки.*]

Кто из вас думает, что их конечное число? [*Меньше студентов поднимают руки.*]

Кто из вас благоразумно промолчит? [*Большинство хранит молчание.*]

Я скажу вам ответ: предположительно также «да», но доказательство неизвестно. Это еще одна нерешенная задача.

Какие аргументы можно привести, чтобы показать правдоподобие предположения о бесконечно большом количестве таких чисел? Я приведу вам аргументы, которые не только позволят вам предположить, что их бесконечно много, но и позволят предположить, насколько много. Для этого мы разберем следующие вопросы:

Как много двойных простых чисел p, $p + 2$ при $p \leqslant x$?

Как много простых чисел p типа $n^2 + 1$ при $p \leqslant x$?

Займемся вопросом о двойных простых числах. Предположим, что вы принимаете, что вероятность того, что число n — простое, равна $\frac{1}{\log n}$. Можете ли вы предположить, какова вероятность того, что n, $n+2$ двойные простые числа?

[*Студент произносит правильный ответ после некоторого колебания.*]

Да, вероятность действительно должна равняться произведению

$$\frac{1}{\log n} \cdot \frac{1}{\log(n+2)}.$$

Но так как значение $\log(n+2)$ очень близко к значению $\log n$ при больших n, мы можем вместо этого записать $\frac{1}{(\log n)^2}$, и поэтому, если мы будем обозначать через $\pi_{tw}(x)$ количество двойных простых среди целых чисел 2, 3, 4, 5, ... , x, тогда, предположительно, мы должны получить выражение

$$\pi_{tw}(x) \approx \frac{x}{(\log x)^2}.$$

Но когда я записываю выражение в правой части, я предполагаю, что вероятности независимые. А они независимые? Ответ — НЕТ. Поэтому мы должны внимательнее записывать формулу. Правильный ответ следующий: количество двойных простых $\pi_{tw}(x)$ удовлетворяет асимптотической формуле

$$\pi_{tw}(x) \sim C_{tw} \frac{x}{(\log x)^2} \quad \text{или лучше} \quad \sim C_{tw} \int\limits_2^x \frac{1}{(\log t)^2} dt,$$

где C_{tw} — положительная константа двойных простых чисел. Эта константа отражает вероятностные зависимости. Теперь задача заключается в вычислении этой константы. Нам нужна явная формула для константы, которая достаточно сложна и получена математиками Харди и Литтлвудом [HaL 23].

[*Чтобы не прерывать основной поток идей более техническим вопросом, обсуждение этой константы будет продолжено позже.*]

Теперь мы перейдем к вопросу, какова асимптотическая оценка количества $\pi_Q(x)$ простых чисел $p \leqslant x$, таких, что p

может быть записано в виде k^2+1 с некоторым положительным целым k? Используя те же самые рассуждения, мы можем описать вероятность того, что число n, представляемое в виде k^2+1, является простым числом. Эта вероятность должна приближенно равняться

$$\frac{1}{\log n}\cdot\frac{1}{\sqrt{n}}.$$

Заметим, что $\log\sqrt{x}=\frac{1}{2}\log x$ и поэтому с точностью до постоянного множителя, не имеет значения, записываем ли мы $\log n$ или $\log\sqrt{n}$. Затем мы можем записать сумму

$$\sum_{2\leqslant n\leqslant x}\frac{1}{\log n}\cdot\frac{1}{\sqrt{n}}\quad\text{или интеграл}\quad\int\limits_2^x\frac{1}{(\log t)t^{1/2}}\,dt.$$

Интегрируя по частям и предполагая, что должен быть некоторый постоянный множитель C_Q, чтобы учесть более скрытые вероятностные зависимости, мы получим гипотезу, что

$$\pi_Q(x)\sim C_Q\frac{\sqrt{x}}{\log x}.$$

Харди–Литтлвуд также определили значение константы C_Q (предположительно).

Большое обобщение: гипотеза Бейтмана–Хорна

Теперь мы займемся большим обобщением, которое систематически предсказывает асимптотическое поведение при подсчете простых чисел в очень общем случае, включающем частные случаи, упомянутые выше. Пусть $f(T)$ — полином с целыми коэффициентами. Нас интересует, являются ли значения $f(n)$ простыми числами для бесконечно большого количества положительных целых чисел n. Запишем полином через его коэффициенты.

$$f(T)=a_dT^d+a_{d-1}T^{d-1}+\ldots+a_0,$$

причем $a_0, \ldots, a_d$ — целые числа, а a_d называется старшим коэффициентом. Целое число d называется степенью полинома, если $a_d \neq 0$. Мы предполагаем, что выполнено также $d \geqslant 1$, так что полином не является константой. Каковы необходимые условия того, что среди значений f бесконечное количество простых чисел? Очевидно, мы должны предположить, что:

1. $f(T)$ неприводим.

2. Старший коэффициент a_d положительный.

3. Коэффициенты $f(T)$ не имеют общего простого множителя.

Этого достаточно?

[*Некоторые студенты сказали «да», некоторые благоразумно промолчали. Однажды в университете Калифорнии в Беркли двое профессоров из аудитории сказали «да». Затем я указал на студента и спросил: «А что вы скажете?» Студент ответил: «Разумеется».*]

На самом деле ответ «НЕТ»! Как обнаружил Буняковский еще в 1854, этих условий недостаточно. К примеру, у полинома

$$f(T) = T^3 - T - 3$$

все значения делятся на 3, т.е. $f(n)$ делится на 3 для всех целых чисел n. Элементарная теорема теории чисел, — малая теорема Ферма, — утверждает, что, даже в более общем случае, для любого простого числа p верно соотношение

$$n^p \equiv n \bmod p \quad \text{для всех целых } n.$$

Другими словами $n^p - n$ делится на p при всех целых n, но при этом коэффициенты полинома $T^3 - T - 3$ или $T^p - T - p$ взаимно простые. [*Указывая на студента, который сказал «Разумеется»*]: Так что не все так просто, не так ли? Не делайте то, что вы сделали, сказав «разумеется», основываясь на том, что сказал кто-то другой. Используйте ваши собственные мозги. Если вы не знаете или хотите подумать о задаче, так и скажите. Вы ведь не будете делать так снова?

Студент: Нет, не буду.

Серж Ленг: Хорошо. Давайте продолжим. Буняковский обнаружил, что нужно добавить. В 1854 г. он сделал следующее предположение [Bou 1854]. Пусть f — полином с целыми коэффициентами, степень полинома больше либо равна 1, и выполнены следующие три условия:

— старший коэффициент положительный;

— полином неприводим;

— никакое простое число не делит всех значений $f(n)$, когда n пробегает по множеству положительных целых чисел.

Тогда существует бесконечно много целых чисел n, таких, что $f(n)$ — простое. Это предположение было обобщено на случай нескольких полиномов Шинцелем [SchS 58]. Более того, Бейтман и Хорн получили количественную форму для этой гипотезы, точную формулировку которой мы сейчас приведем.

Пусть $f_1, \ldots, f_r$ — полиномы одной переменной с целыми коэффициентами и положительным старшим коэффициентом. Пусть $d_1, \ldots, d_r$ их степени. Пусть f — произведение, $f = f_1 \cdots f_r$. Предполагаем, что:

— каждый полином неприводим над полем рациональных чисел;

— полиномы попарно взаимно просты, т.е. $(f_i, f_j) = 1$ при $i \neq j$;

— никакое простое число не делит всех значений $f(n)$, когда n пробегает по множеству положительных целых чисел.

Будем использовать векторные обозначения, т.е. $(f) = (f_1, \ldots, f_r)$. Пусть $\pi_{(f)}(x)$ — количество положительных целых чисел $n \leqslant x$ таких, что $f_1(n), \ldots, f_r(n)$ — все простые числа. (Мы отбрасываем конечное число значений n, для которых некоторые $f_i(n)$ отрицательны). Заметьте, что Бейтман и Хорн дают альтернативный способ подсчета. В предыдущих примерах мы считали количество простых чисел $p \leqslant x$, удовлетворяющих разным условиям. Здесь мы считаем количество положительных целых $n \leqslant x$ таких, что $f_i(n)$ — простое число для всех $i = 1, \ldots, r$. Гипотеза Бейтмана–Хорна состоит в асимптотическом соотношении для $\pi_{(f)}(x)$ [BaH 62], а именно

$$\pi_{(f)}(x) \sim (d_1, \ldots, d_r)^{-1} C(f) \int_2^x \frac{1}{(\log t)^r}\, dt,$$

где $C(f)$ — бесконечное произведение по всем простым числам:

$$C(f) = \prod_p \left\{ \left(1 - \frac{1}{p}\right)^{-r} \left(1 - \frac{N_f(p)}{p}\right) \right\},$$

и $N_f(p)$ — количество целых n, $1 \leqslant n \leqslant p$, таких, что $f(n)$ делится на p. Если вам известен язык и понятия теории сравнений, то мы можем также сказать, что $N_f(p)$ — число решений по модулю p сравнения

$$f(n) \equiv 0 \mod p.$$

[Если $N_f(p) = p$ при некотором простом числе p, то $C(f) = 0$ и

$$\pi_{(f)}(x) \leqslant d_1 + \ldots + d_r \quad \text{при всех } x.$$

Давайте будем исключать этот тривиальный случай]. Бейтман и Хорн приводят экспериментальное подтверждение своей гипотезы, включающее компьютерные расчеты.

Я считаю важной задачей получение оценки для ошибки аппроксимации для гипотезы Бейтмана–Хорна аналогично оценке, полученной для гипотезы Римана. Это может привести к глобальному пересмотру контекста для римановых явных формул.

Все ранее приведенные примеры являются частными случаями гипотезы Бейтмана–Хорна.

Положим $r = 1$ и $f(T) = T$. Тогда гипотеза совпадает с теоремой о простых числах.

Пусть $r = 1$ и $f(T) = T^2 + 1$. Тогда гипотеза в действительности дает асимптотики, обсуждавшиеся для количества простых чисел вида $n^2 + 1$. Поскольку Бейтман–Хорн подсчитывают $n \leqslant x$, а не $p \leqslant x$, потребуется несколько строк, чтобы показать, что их асимптотика совпадает с ответом, полученным Харди–Литтлвудом. Меня однажды заинтересовала связь между двумя вариантами, и я провел вычисления, которые заняли у меня около одной страницы.

Пусть $r = 2$ и пусть $f_1(T) = T$, $f_2(T) = T + 2$. Теперь из гипотезы Бейтмана–Хорна получается гипотеза о подсчете двойных простых чисел. Также потребовалась примерно одна страница вычислений, чтобы показать, что в этом случае константа Харди–Литтлвуда та же самая, что и константа Бейтмана–Хорна.

Наконец, возвращаясь к одному полиному, пусть a, b — целые числа, $a > 0$ и a, b взаимно просты. Пусть $f(T) = aT + b$. Тогда гипотеза Бейтмана–Хорна приводит к следующему факту: в арифметической последовательности бесконечно много простых чисел, также из гипотезы находится плотность этих простых чисел, что в действительности является знаменитой теоремой Дирихле.

Выражение для константы $C(f)$ перед логарифмическим интегралом представляет собой пример очень общего феномена, часто встречающегося в теории диофантовых уравнений. Плотность решений выражается через произведение p-плотностей для каждого простого p и еще одного вещественного множителя, который в гипотезе Бейтмана–Хорна равен $\frac{1}{(d_1 \cdots d_r)}$. Множитель, который содержит простое число p описывает поведение по модулю p, т.е. поведение по отношению к делимости на p и, в частности, этот множитель содержит число решений сравнения по модулю p. В [HaL 23] и [LaT 76] вы найдете дальнейшие примеры этого феномена.

Приложение

Комментарии о константах

Гипотеза Бейтмана–Хорна дает систематический ответ для констант, возникающих перед логарифмическим интегралом. В некоторых частных случаях Харди–Литтлвуд вычислили константу [HaL 23], но использовали альтернативный способ подсчета, а именно простые числа $p \leqslant x$ вместо $n \leqslant x$. Мы сейчас вернемся к обсуждению констант Харди–Литтлвуда. Они выражаются как бесконечные произведения.

Мы начнем с константы двойных простых чисел C_{tw}. История этой константы весьма интересна. Сильвестер в 1871 и Брун в 1915 предположили неправильную константу, поскольку не приняли во внимание некоторые скрытые вероятностные зависимости между простыми числами. Харди и Литтлвуд в 1923 [HaL 23] отметили, что Сильвестер и Брун предположили неправильную формулу, и сами предположили правильное значение константы C_{tw}. «Правильное» в данном случае означает в частности, что вы можете получить таблицы значений $\pi_{tw}(x)$ и значений $C_{tw} \cdot x/(\log x)^2$ и сравнить, и получите, что практическая проверка подтверждает полученные значения. Таким образом, слово «правильность» здесь частично используется в физическом смысле. Но существует другой компонент понятия «правильный» (при отсутствии самого доказательства), а именно то, что предположение подходит под определенные сильные структурные размышления о том, что существует определенная последовательная структура, из которой делаются предположения, что структура сама по себе частично гипотетическая. Харди и Литтлвуд также опирались на эту структуру. Разумеется, такая структура может оказаться не единственной. В любом случае значение, полученное Харди – Литтлвудом равно

$$C_{tw} = 2 \prod_{p-\text{нечет.}} \left(1 - \frac{1}{(p-1)^2}\right).$$

Произведение берется по всем простым числам, не равным 2, т. е. по всем нечетным простым числам. Как вы можете видеть, эта константа несколько отличается от константы Бейтмана – Хорна. Различие возникает из-за альтернативного способа подсчета. Т. к. Харди, Литтлвуд и Бейтман, Хорн использовали сильно отличающиеся подходы к этой константе, но получили один и тот же ответ, это считается подтверждением того, что ответ, полученный Харди – Литтлвудом является правильным, поскольку он структурно подходит к другим исследованиям.

Теперь мы приведем значение, полученное Харди – Литтлвудом для константы C_Q, входящей в гипотетическую асимптотическую формулу для количества простых чисел меньше либо равных x, представленных в виде $p = n^2 + 1$. Гипотетическое значение равно

$$C_Q = \prod_{p\,-\,\text{нечет.}} \left(1 - \chi(p)\frac{1}{p-1}\right),$$

где $\chi(p) = 1$, если -1 является квадратом по модулю p, т. е. если $n^2 + 1 \equiv 0 \bmod p$ имеет решение, и $\chi(p) = -1$, если -1 не является квадратом по модулю p [см. также [LaT 76], часть II, особенно стр. 81, пример 1, где эта константа выводится из совершенно другой вероятной модели.]

[*Студент спрашивает, как возникают такие гипотезы-формулы и как делаются предположения.*]

Да, давайте займемся этим вопросом. Во-первых, существует необъяснимый компонент, а именно мозг математика, который знает, как посмотреть и где посмотреть, и он силен в предположениях. Наличие такого таланта — один из факторов, отличающих математика-исследователя от других людей. Затем тяжелая работа по проверке предположений, их структурированию и по написанию исследовательской статьи.

Давайте рассмотрим вопрос, как можно прийти к идее, что существует бесконечно много двойных простых чисел или простых чисел в виде $k^2 + 1$. Считаете ли вы, что мы получили нечто вроде ответа? Да? Ни то ни се? По крайней мере это что-то, этот ответ не только дает вам знание, что их бесконечно много, но и количественную меру — насколько много. Кто-нибудь может попробовать улучшить вероятностную модель так, чтобы вы также получили значение для константы, чтобы увидеть более точно, как много их должно быть.

Не так уж легко пройти по всему пути, как вы можете увидеть из факта, что первоклассные математики, такие как Сильвестер и Брун не смогли получить правильных ответов при первой попытке. Поэтому работа Харди – Литтлвуда — очень хорошая работа.

Я покажу вам сейчас один из подходов, который в конце концов приведет к константам. Рассуждения о необходимости являются значительно более сложными. Мы будем разбираться лишь с обычным случаем теоремы о простых числах.

Возьмем все числа $1, 2, 3, 4, 5, 6, \ldots, x$. Как много их них делится на 2? Примерно $\frac{x}{2}$. Количество целых чисел, меньших x, не делящихся на 2 приблизительно равно $\left(1 - \frac{1}{2}\right)x$, и множитель перед x равен $\left(1 - \frac{1}{2}\right)$. Количество целых чисел,

не делящихся на 3
$$\left(1 - \frac{1}{3}\right)x.$$
Количество целых чисел, *не* делящихся на 5
$$\left(1 - \frac{1}{5}\right)x.$$
Поэтому, чтобы получить количество $\pi(x)$ простых чисел меньше x, мы хотим знать, сколько целых чисел не делится на 2, не делится на 3, не делится на 5 и т. д. Как много вы получите? Ответ грубо равен
$$G(x) \cdot x, \quad \text{где } G(x) = \prod_{p \leqslant x}\left(1 - \frac{1}{p}\right).$$
Это первое приближение к определению правильной константы для $\pi(x)$, количества простых чисел меньших x, за исключением того, что мы предположили, что условия неделимости независимы, а они не независимы. Поэтому мы должны взять множитель
$$\prod_{p \leqslant x}\left(1 - \frac{1}{p}\right) = G(x)$$
и умножить на некоторую константу из-за сложных условий зависимости. Константа в этом случае равна e^{γ}, где γ — константа, называемая константой Эйлера. По теореме Мертенса выполнено асимптотическое соотношение
$$e^{\gamma} G(x) \sim \frac{1}{\log x},$$
и, таким образом,
$$\pi(x) \sim e^{\gamma} G(x) x \sim \frac{x}{\log x}.$$
Вы можете найти обсуждение того, почему возникает e^{γ} в статье Харди–Литтлвуда, и также в работе Харди и Райта [HaW 80], гл. 22, 22.20.

Так что вы видите, что простое обсуждение обычной теоремы о простых числах становится все более техническим и сложным. И еще сложнее точные вероятностные рассуждения Харди–Литтлвуда для двойных простых чисел, для простых чисел вида n^2+1 и в других случаях. А теперь, если вы до сих пор заинтересованы, идите и читайте Бейтмана–Хорна и Харди–Литтлвуда.

Некоторые графики

Я благодарен Стиву Миллеру за построение некоторых графиков, иллюстрирующих экспериментальное поведение ζ-функции и других трех функций $\frac{x}{\log x}$, $\pi(x)$ и $\mathrm{Li}(x)$ в определенных областях.

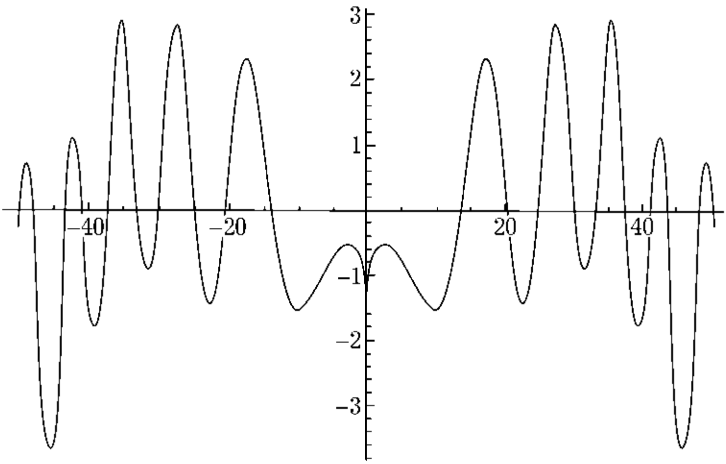
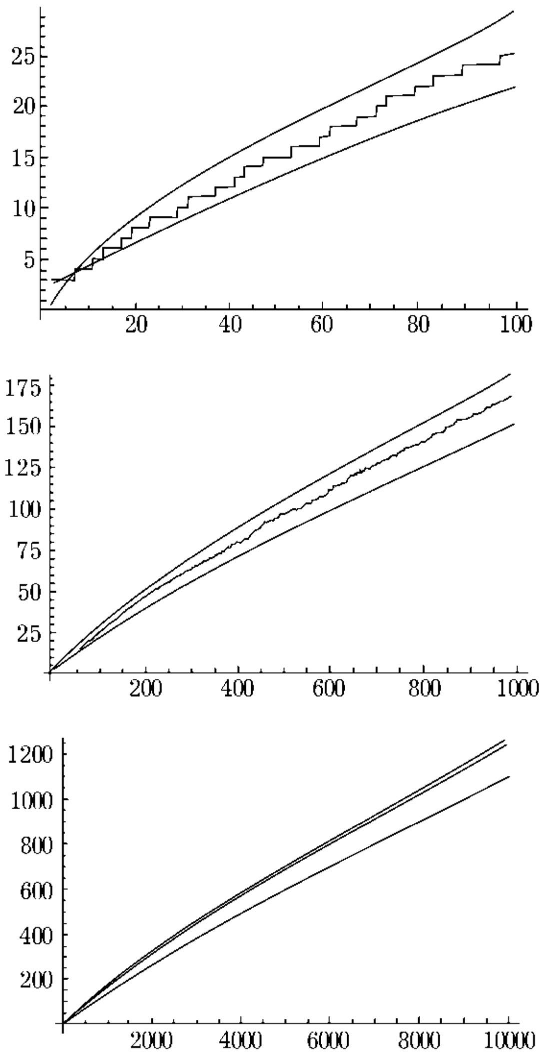


График $\xi\left(\frac{1}{2}+it\right)$, $-50<t<50$

Первый график — график функции $\xi\left(\frac{1}{2}+it\right)$, которая оказывается действительной функцией переменной t из-за функционального уравнения.

Следующие три графика сравнивают функции $\frac{x}{\log x}$, $\pi(x)$ и $\mathrm{Li}(x)$ на трех различных интервалах переменной x, а именно от 0 до 100, 1000 и 10 000 соответственно. Заметьте, что при продолжении становится очевидным из эксперимен-



Графики $\mathrm{Li}(x)$ (верхний), $\pi(x)$ (средний), $x/\log x$ (нижний)

та, что $\mathrm{Li}(x)$ дает лучшую аппроксимацию функции $\pi(x)$, чем $\dfrac{x}{\log x}$. Также очевидно, что $\mathrm{Li}(x)$ лежит над $\pi(x)$. До каких пределов, как вы думаете, будет продолжаться такое статистическое поведение? Сходите и спросите какого-нибудь специалиста по теории чисел. Или посмотрите книгу Ингама [Ing 32], гл. 5, теорема 35.

Литература

[BaH 62] P. Bateman and R. Horn. A heuristic asymptotic formula concerning the distribution of prime numbers. *Math. Comp.* **16** (1962), pp. 363–367.

[Bou 1854] V. J. Bouniakowsky. Sur les diviseurs numériques invariables des fonctions rationnelles entières. *Mem. Sci. Math et Phys.* **T. VI** (1854), pp. 307–329.

[Br 15] V. Brun, Über das Goldbachsche Gesetz und die Anzahl der Primzahlpaare. *Archiv for Mathematik* (Christiania) **34**, Part 2 (1915), pp. 1–15.

[HaR 74] H. Halberstam and H. Richert. *Sieve Methods*, Academic Press, 1974.

[HaL 23] G. H. Hardy and J. E. Littlewood. Some problems of Partitio Numerorum, *Acta Math.* **44** (1923), pp. 1–70.

[HaW 80] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*, Fourth edition, Oxford University Press, 1980.

[Ing 32] A. E. Ingham. *The Distribution of Prime Numbers*, Cambridge University Press, 1932.

[La 93] S. Lang. *Complex Analysis*, Third edition, Springer-Verlag, 1993 (Fourth edition, 1999).

[LaT 76] S. Lang and H. Trotter. *Frobenius Distributions in* GL_2*-Extensions, Springer Lecture Notes* **504**, Springer-Verlag, 1976.

[New 80] D. J. Newman. Simple analytic proof of the prime number theorem, *Amer. Math. Monthly* **87** (1980), pp. 693–696.

[Pat 88] S. J. Patterson. *An Introduction to the Theory Riemann Zeta Function*, Cambridge Studies in Advanced Mathematics, vol. 14, Cambridge University Press, 1988.

[SchS 58] A. Schinzel and W. Sierpinsky. Sur certaines hypothèses concernant les nombres premiers. *Acta Arith.* **4** (1958), pp. 185–208.

[Scho 76] L. Shoenfeld, Some sharper Tchebychev estimates II, *Math. Comp.* **30** (1976), pp. 337–360.

[Syl 1871] J. Sylvester, On the partition of an even number into two prime numbers. *Nature* **56** (1896–1897), pp. 196–197 (=*Math. Papers* **4**, pp. 734–737).

Гипотеза abc

Давайте начнем с теоремы о полиномах. Вы, вероятно, считаете, что вам известно все о полиномах. Большинство математиков думали бы так же, включая меня. К всеобщему удивлению Р. Мейсон в 1983 году обнаружил новый очень интересный факт о полиномах [Ma 83]. Что еще более примечательно, этот факт в действительности уже был обнаружен другим математиком В. Стотерсом [St 81], но математики не обратили внимание на это, и я узнал о статье Стотерса лишь намного позже от У. Занира [Za 95], который также переоткрыл некоторые из результатов Стотерса. Так что история математики не всегда течет гладко.

В любом случае, хотя история математики весьма поучительна, давайте отложим историю для того, чтобы сформулировать и доказать теорему Мейсона–Стотерса. Затем мы обсудим, как другие математики преобразовали эту теорему в гипотезу об обычных целых числах. Мы рассматриваем полиномы с комплексными коэффициентами. Множество всех таких полиномов от переменной t будем обозначать $\mathbb{C}[t]$. Запишем ненулевой элемент $\mathbb{C}[t]$ в виде

$$f(t) = c_1 \prod_{i=1}^{r} (t - \alpha_i)^{m_i},$$

где $\alpha_1, \alpha_2, \ldots, \alpha_r$ — различные корни, f и c_1 — константа, $c_1 \neq 0$. Целые числа m_i $(i = 1, \ldots, r)$ — кратности корней и степень полинома f равна

$$\deg f = m_1 + \ldots + m_r.$$

Количество различных корней f будет обозначаться $n_0(f)$, поэтому по определению $n_0(f) = r$.

Очевидно, что степень f может быть очень большой, а $n_0(f)$ может быть маленькой. Например, $f(t) = (t - \alpha)^{1000}$ —

степени 1000, а $n_0(f) = 1$. Если f, g — два ненулевых полинома, тогда в общем случае

$$n_0(fg) \leqslant n_0(f) + n_0(g).$$

Если дополнительно потребовать, что f, g — взаимно просты, тогда мы действительно получим равенство

$$n_0(fg) = n_0(f) + n_0(g).$$

Сформулируем теорему Мейсона–Стотерса:

Теорема. *Пусть f, g, $h \in \mathbb{C}[t]$ не равные константе взаимно простые полиномы, удовлетворяющие равенству $f + g = h$. Тогда*

$$\max(\deg f, \deg g, \deg h) \leqslant n_0(fgh) - 1.$$

Теорема демонстрирует в точности, каким образом отношение $f + g = h$ накладывает ограничение на степени полиномов f, g, h, а именно, они не превосходят количества различных корней полинома fgh, даже с добавлением -1.

Перед тем, как мы приведем доказательство Мейсона, мы упомянем некоторые приложения, которые покажут, насколько сильна эта теорема. Вы все знаете последнюю теорему Ферма, которая была доказана Вайлсом [Wi 95] десятилетием позже:

Теорема. *Пусть n — целое число $\geqslant 3$. Не существует решений уравнения*

$$x^n + y^n = z^n$$

в ненулевых целых числах x, y, z.

Аналогичная теорема для полиномов была известна еще в прошлом столетии и была доказана методами алгебраической геометрии. Сейчас мы получим намного более простое доказательство, используя теорему Мейсона–Стотерса.

Теорема. *Пусть n — целое число $\geqslant 3$. Не существует решений уравнения*

$$x(t)^n + y(t)^n = z(t)^n$$

с не равными константе взаимно простыми полиномами x, y, $z \in \mathbb{C}[t]$.

Доказательство.

Пусть $f(t) = x(t)^n$, $g(t) = y(t)^n$ и $h(t) = z(t)^n$. Тогда по теореме Мейсона – Стотерса получаем

$$\deg x(t)^n \leqslant n_0(x(t)^n y(t)^n z(t)^n) - 1.$$

Однако $\deg x(t)^n = n \cdot \deg x(t)$ и $n_0(x(t)^n) = n_0(x(t)) \leqslant \deg x(t)$. Следовательно,

$$n \cdot \deg x(t) \leqslant \deg x(t) + \deg y(t) + \deg z(t) - 1.$$

Тем же способом мы получаем аналогичные неравенства для $y(t)$ и $z(t)$:

$$n \cdot \deg y(t) \leqslant \deg x(t) + \deg y(t) + \deg z(t) - 1,$$

$$n \cdot \deg z(t) \leqslant \deg x(t) + \deg y(t) + \deg z(t) - 1.$$

Складывая три неравенства, получаем

$$n(\deg x(t) + \deg y(t) + \deg z(t)) \leqslant 3(\deg x(t) + \deg y(t) + \deg z(t)) - 3.$$

Таким образом,

$$(n-3)(\deg x(t) + \deg y(t) + \deg z(t)) \leqslant -3.$$

При $n \geqslant 3$ этого очевидно не может быть, потому что левая часть больше либо равна 0, а правая часть отрицательна. Итак, мы доказали последнюю теорему Ферма для полиномов. ■

Доказательство последней теоремы Ферма для полиномов намного труднее без использования теоремы Мейсона – Стотерса. Не ясно, сколько времени займет у вас или у случайно выбранного математика доказательство теоремы каким-нибудь способом. Попробуйте это с вашими друзьями и посмотрите, сколько времени им потребуется. Приведенное выше доказательство с использованием теоремы Мейсона – Стотерса является простым и элегантным. Позднее мы приведем различные обобщения. Увидев силу этой теоремы, теперь давайте докажем ее.

Доказательство теоремы Мейсона – Стотерса. В утверждении теоремы в левой части неравенства используется степень, а в правой части число n_0, которое обозначает количество различных корней полинома. Поэтому нам необходимо научиться понимать и управлять кратностями корней. Чтобы сделать это, мы разделим уравнения $f+g=h$ на h и получим

$$\frac{f}{h}+\frac{g}{h}=1.$$

Обозначим $R=\dfrac{f}{h}$ и $S=\dfrac{g}{h}$. Тогда $R+S=1$. Теперь вы должны почувствовать непреодолимое желание сделать что-нибудь с этим уравнением. Что вы делаете с функциями? Вы вычисляете их производные. Таким образом, мы получаем новое уравнение $R'+S'=0$, которое мы перепишем в виде

$$\frac{R'}{R}R+\frac{S'}{S}S=0. \tag{1}$$

Рассмотрим частное $\dfrac{g}{f}$. Используя наши обозначения и уравнение (1), мы получим

$$\frac{g}{f}=\frac{S}{R}=-\frac{R'/R}{S'/S}. \tag{2}$$

При этом мы выразили $\dfrac{g}{f}$ как частное логарифмических производных. Действительно, в математическом анализе, если F — функция, тогда $\dfrac{F'}{F}$ называется ее логарифмической производной. Используя эту логарифмическую производную, мы сможем управлять кратностями корней. Нам нужно использовать свойство логарифмической производной, а именно то, что она переводит произведения в суммы. Другими словами для двух функций F, G имеем

$$\frac{(FG)'}{FG}=\frac{F'}{F}+\frac{G'}{G}.$$

Это немедленно следует из правила дифференцирования произведения. Затем для частных получим

$$\frac{(F/G)'}{F/G}=\frac{F'}{F}-\frac{G'}{G}.$$

Таким образом, логарифмическая производная преобразует частные в разности. Затем по индукции мы получаем подобные отношения для сумм или частных нескольких сомножителей. К примеру, если $f_1, \ldots, f_n$ — полиномы, тогда

$$\frac{(f_1 \cdots f_n)'}{f_1 \cdots f_n} = \frac{f_1'}{f_1} + \ldots + \frac{f_n'}{f_n}.$$

Применяем это правило к разложениям на множители наших исходных полиномов f, g, h. Мы знаем для них следующие разложения на множители

$$f(t) = c_1 \prod (t - \alpha_i)^{m_i},$$
$$g(t) = c_2 \prod (t - \beta_j)^{n_j},$$
$$h(t) = c_3 \prod (t - \gamma_k)^{l_k}.$$

Заметим, что логарифмическая производная константы равна 0. Для полинома $(t - \alpha)$ логарифмическая производная равна $\dfrac{1}{(t-\alpha)}$. Поэтому вычисляя логарифмические производные $f(t)$, $g(t)$, $h(t)$, получаем

$$\frac{f'}{f} = \sum \frac{m_i}{(t - \alpha_i)}, \qquad \frac{g'}{g} = \sum \frac{n_j}{(t - \beta_j)}, \qquad \frac{h'}{h} = \sum \frac{l_k}{(t - \gamma_k)}.$$

Применяя правила для логарифмических производных к $R = f/h$ и $S = g/h$, получаем из (2):

$$\frac{g}{f} = -\frac{\dfrac{f'}{f} - \dfrac{h'}{h}}{\dfrac{g'}{g} - \dfrac{h'}{h}} = -\frac{\sum \dfrac{m_i}{t - \alpha_i} - \sum \dfrac{l_k}{t - \gamma_k}}{\sum \dfrac{n_j}{t - \beta_j} - \sum \dfrac{l_k}{t - \gamma_k}}. \tag{3}$$

Приведем выражение к общему знаменателю, для этого обозначим

$$D(t) = \prod (t - \alpha_i) \prod (t - \beta_j) \prod (t - \gamma_k).$$

Очевидно, что $\deg D(t) = n_0(fgh)$. Поэтому

$$\deg\Big(\frac{D(t)}{(t - \alpha_i)}\Big) = n_0(fgh) - 1 = \deg\Big(\frac{D(t)}{(t - \beta_j)}\Big) = \deg\Big(\frac{D(t)}{(t - \gamma_k)}\Big).$$

Умножим числитель и знаменатель выражения (3) на $D(t)$, после чего получим

$$\frac{g}{f} = -\frac{\sum \frac{m_i}{t-\alpha_i} - \sum \frac{l_k}{t-\gamma_k}}{\sum \frac{n_j}{t-\beta_j} - \sum \frac{l_k}{t-\gamma_k}} \cdot \frac{D(t)}{D(t)} =$$

$$= \frac{\text{полином степени} \leqslant n_0(fgh)-1}{\text{полином степени} \leqslant n_0(fgh)-1}.$$

Таким образом, $\frac{g}{f}$ равно частному двух полиномов, чьи степени $\leqslant n_0(fgh)-1$. Поскольку f, g — взаимно простые, из этого следует что степени g и f не превосходят $n_0(fgh)-1$. И наконец, т. к. $h = f + g$, мы приходим к заключению, что $\deg h$ не превосходит $n_0(fgh)-1$. Этим завершается доказательство теоремы Мейсона – Стотерса.

Примечательно, что такой простой, но сильный результат о полиномах был открыт лишь в начале 80-х, в конце двадцатого века.

Теперь мы собираемся преобразовать теорему в утверждение о целых числах. Вам, возможно, уже известно, что существует глубокая аналогия между целыми числами и полиномами. К примеру, обе системы удовлетворяют алгоритму Евклида, поэтому в обеих системах существует однозначное разложение на неприводимые или простые элементы. Мы ищем число, связанное с целым числом таким способом, чтобы оно было аналогично числу n_0 полинома. Мы также ищем число, аналогичное степени полинома. При умножении степени складываются, т. е. степень произведения полиномов равна сумме степеней индивидуальных сомножителей. Опыт показывает, что аналогом степени является логарифм абсолютного значения целого числа. Поэтому для простого числа p аналогом степени неприводимого полинома является обычный логарифм $\log p$. Для двух положительных целых чисел m, n мы получаем $\log(mn) = \log m + \log n$. Если $m = \prod p_i^{m_i}$ — разложение числа m на простые множители, тогда

$$\log m = \sum m_i \log p_i.$$

Теперь мы хотим получить аналог числа n_0. Для полинома

$$f(t) = \prod (t - \alpha_i)^{m_i}$$

давайте определим

$$N_0(f) = \prod (t - \alpha_i),$$

то есть мы заменим кратности m_i на 1 в разложении многочлена $f(t)$. Тогда по определению

$$n_0(t) \text{ равна степени } N_0(f(t)).$$

Так каково же правильное определение для $n_0(m)$ для положительного целого числа m? Предположим, что

$$m = p_1^{m_1} \cdots p_r^{m_r} \tag{4}$$

разложение числа m на простые. Чему же равно $n_0(m)$?

Студент. Я думаю $n_0(m)$ должно быть количеством различных простых множителей, поэтому $n_0(m) = r$.

Серж Ленг. Это неплохой ответ, но и не достаточно хороший. Когда мы работаем с комплексными полиномами, тогда они обладают однозначным разложением на простые, т. е. на неприводимые полиномы, которые имеют степень 1. Если вы работаете с полиномами с другими типами коэффициентов, к примеру, с полиномами с рациональными коэффициентами, тогда неприводимый полином не обязательно имеет степень 1. Нечто аналогичное происходит с простыми числами. Это означает, что нам необходимо приписать вес простому числу, и этот вес равен $\log p$. Поэтому для положительного целого числа m с разложением на простые множители (4), мы определим

$$n_0(m) = \sum_{i=1}^{r} \log p_i.$$

Мы также можем записать это выражение в виде

$$n_0(m) = \sum_{p|m} \log p.$$

Затем мы определим

$$N_0(m) = \prod_{p|m} p = \prod_{i=1}^{r} p_i.$$

Таким образом, $N_0(m)$ есть произведение простых чисел, делящих m, но взятых с кратностью 1. Будем называть $N_0(m)$ *радикалом* числа m.

Теперь, когда у нас есть определение $n_0(m)$ и $N_0(m)$, каким будет аналог теоремы Мейсона – Стотерса? Легко перевести гипотезу:

Гипотеза. *Пусть a, b, c — ненулевые, взаимно простые целые числа, такие, что $a + b = c$.*

Теперь нам нужно неравенство вида

$$\max(|a|, |b|, |c|) \leqslant ???.$$

Что должно быть в правой части неравенства?

Студент. Может $N_0(abc) - 1$?

Серж Ленг. Все не так просто. Во-первых, что касается -1: в случае полиномов это был подарок богов, и мы не можем ожидать чего-нибудь столь же точного для целых чисел. Кроме того, мы здесь действуем мультипликативно, а не аддитивно. Так что давайте взглянем на неравенство

$$\max(|a|, |b|, |c|) \leqslant N_0(abc).$$

К сожалению, оказывается, что в общем случае это неравенство не верно. Это было впервые обнаружено Стюартом – Тийдеманом [StT 86], которые также показали, что по-прежнему неравенство не будет выполняться, если мы заменим правую часть на выражение: константа умножить на $N_0(abc)$ независимо от того, насколько велика константа. Другими словами, *не существует такой константы K, чтобы неравенство*

$$\max(|a|, |b|, |c|) \leqslant K \cdot N_0(abc)$$

было верным для всех взаимно простых целых a, b, c, когда $a + b = c$. Два студента из Йельского университета Стейл

Войтек Ястрзебовский и Дэн Шпильман привели мне следующее простое доказательство этого факта.

Мы рассматриваем уравнение $a_n + b_n = c_n$, где

$$a_n = 3^{2^n}, \quad b_n = -1 \quad \text{и} \quad c_n = 3^{2^n} - 1.$$

Записав $3 = 1 + 2$, легко доказать по индукции, что 2^n делит $3^{2^n} - 1$. Поэтому

$$N_0(a_n b_n c_n) \leqslant 3 \cdot 2 \cdot \frac{c_n}{2^n}.$$

Однако неравенство

$$3^{2^n} \leqslant K \cdot 3 \cdot 2 \frac{3^{2^n}}{2^n}$$

не может выполняться для всех n независимо от того, насколько большой мы выберем константу K, поскольку $K \cdot 3 \cdot \dfrac{2}{2^n}$ стремится к 0 при увеличении n.

Поэтому необходимо изменить и еще ослабить неравенство. Правильная гипотеза, названная *гипотезой abc*, была получена Массером и Остерле. Это одна из лучших гипотез столетия.

Гипотеза *abc* (Masser, Oesterle, 1986). *При данном $\varepsilon > 0$ существует константа $K(\varepsilon)$ такая, что для всех ненулевых взаимно простых целых чисел a, b, c, таких, что $a + b = c$, верно неравенство*

$$\max(|a|, |b|, |c|) \leqslant K(\varepsilon)(N_0(abc))^{1+\varepsilon}.$$

Заметим, что Стюарт–Тийдеман привели некоторые оценки снизу для ε [StT 86].

В сущности, теми же самыми аргументами, что и для полиномов, можно показать, что гипотеза abc влечет последнюю теорему Ферма при достаточно больших n. Мы приведем подробное доказательство. Без потери общности мы можем предположить, что a, b, c — положительные целые числа, поэтому нам не нужно писать знак абсолютной величины.

Пусть x, y, z — положительные взаимно простые целые числа такие, что
$$x^n + y^n = z^n.$$
Пусть $a = x^n$, $b = y^n$ и $c = z^n$. Тогда
$$N_0(x^n y^n z^n) = N_0(xyz) \leqslant xyz.$$
Применив гипотезу abc, мы получаем
$$x^n \ll (xyz)^{1+\varepsilon}, \quad y^n \ll (xyz)^{1+\varepsilon}, \quad z^n \ll (xyz)^{1+\varepsilon}, \tag{5}$$
где знак $\ll$ — аббревиатура для неравенства $\leqslant K(\varepsilon)$ с константой $K(\varepsilon)$, зависящей от ε. Поэтому левая часть отношения $\ll$ является меньше либо равной произведению $K(\varepsilon)$ на правую часть. Перемножая неравенства в (5), получим
$$(xyz)^n \leqslant (xyz)^{3+\varepsilon\cdot 3}.$$
Взяв логарифм, получаем
$$(n - 3 - 3\varepsilon)\log(xyz) \leqslant \log K$$
с константой $K = K(\varepsilon)$. Т. к. $xyz \geqslant 2$, это последнее неравенство дает верхнюю границу для n, тем самым доказывая последнюю теорему Ферма для всех достаточно больших n.

Граница для n зависит от константы $K(\varepsilon)$. На сегодняшний день нет гипотез об эффективной оценке $K(\varepsilon)$. Разумеется, просто для эффективности мы можем взять $\varepsilon = 1$, поэтому эффективное вычисление $K(1)$ даст эффективное доказательство последней теоремы Ферма, сводя ее к вычислениям для конечного числа случаев. Естественно, доказательство Вайлса верно для всех n, но существуют уравнения, похожие на уравнение Ферма, для которых не известно аналога теоремы Вайлса, и подход через гипотезу abc может сработать.

Студент. Пытался ли кто-нибудь привести вычисления, которые могут привести к контрпримеру для гипотезы abc?

Серж Ленг. Обычно так не происходит. Таблицы разложений на простые числа в различных случаях действительно подтверждают гипотезу. Заметьте, что показатель $1+\varepsilon$ в экспоненте делает гипотезу очень сильной. Среди других вещей гипотеза abc показывает, что если в разложениях чисел a, b, c существуют простые числа с большими показателями, тогда эти простые числа компенсируются большим количеством маленьких простых чисел, возникающих в разложении лишь с показателем 1. Например, существуют таблицы для разложения на простые числа выражений $2^n \pm 1$ и похожих чисел в статье [BLSTW 83]. Эти таблицы явно показывают, что почти все простые числа возникают лишь с показателем 1. Если существуют маленькие простые числа с большими показателями, то они компенсируются большими простыми числами с показателем 1.

Студент. А влечет ли последняя теорема Ферма гипотезу abc?

Серж Ленг. Нет. Последняя теорема Ферма — это только частный случай. Гипотеза abc намного сильнее и дает больше информации о том, как ограничиваются показатели простых чисел в разложениях гипотезы abc. Чтобы прояснить обстановку, для последней теоремы Ферма нам не требуется использовать гипотезу abc с показателем $1 + \varepsilon$. Любого фиксированного показателя будет достаточно.

Студент. Каким образом возникают такие гипотезы?

Серж Ленг. Массер и Остерле не просто получили эту гипотезу. Они работали над намного более общей задачей, которая ни в коей мере не элементарна. То, как я представил гипотезу здесь, не отражает того, как она была открыта исторически. Настоящая жизнь намного сложнее. Гипотеза возникла из глубокого изучения алгебраической геометрии и теории модулярных функций, а не просто из связи с теоремой Мейсона Стотерса. Эти изыскания слишком сложны для того, чтобы я мог рассказать о них сейчас. Тем не менее я хотел бы сделать еще несколько комментариев, которые откроют некоторые возможности.

Рассмотрим уравнение вида

$$u^3 - v^2 = k.$$

Будем искать его решения во взаимно простых целых числах u, v и k. Это уравнение впервые рассматривалось М. Холлом [Ha 71]. Он предложил следующую гипотезу.

Гипотеза Холла. *Для целых чисел u, v и k, удовлетворяющих уравнению $u^3 - v^2 = k \neq 0$, необходимо, чтобы*

$$|u|^3 \ll |k|^{6+\varepsilon} \quad \text{и} \quad |v|^2 \ll |k|^{6+\varepsilon}.$$

В действительности Холл предложил свою гипотезу без ε! В то время он не видел необходимости в добавлении ε. Некоторое время я думал, что ε необходимо. А сейчас для меня не ясно, нужно или не нужно добавлять ε в гипотезу Холла. Возможно, формулировка Холла была совершенно правильной из-за того, что он рассматривал уравнение специального вида. В любом случае, «гипотеза Холла для полиномов» уже доказана в 1965 г. Давенпортом [Da 65], и даже в более точном виде, без неопределенных констант в оценке.

Теорема Давенпорта. *Пусть f, g — два полинома, не равные константе, удовлетворяющие выражению $f^3 - g^2 \neq 0$. Тогда*

$$\frac{1}{2}\deg f \leqslant \deg(f^3 - g^2) - 1$$

и

$$\frac{1}{3}\deg g \leqslant \deg(f^3 - g^2) - 1.$$

Вы можете доказать эти неравенства, непосредственно применяя теорему Мейсона–Стотерса, когда полиномы f и g взаимно простые. Будет хорошим упражнением по алгебре доказательство неравенств также в случае, когда f и g не являются взаимно простыми. Тогда вам необходимо шаг за шагом избавляться от общих множителей, чтобы свести теорему к случаю, когда f и g взаимно простые. Поэтому вам придется рассмотреть более общее уравнение вида

$$Af^3 + Bg^2 = h$$

с полиномиальными коэффициентами A и B. Ограничение на степени f, g, h в таком уравнении будет тогда зависеть от A и B.

Давайте вернемся к целым числам. Мы снова рассмотрим уравнение

$$u^3 - v^2 = k$$

с взаимно простыми числами u, v и k. Применяя гипотезу abc, мы получили ограничения

$$|u|^3 \ll (N_0(k))^{6+\varepsilon} \quad \text{и} \quad |v|^2 \ll (N_0(k))^{6+\varepsilon}. \tag{6}$$

Подробный вывод является легким упражнением. Гипотеза Холла с ε затем следует из гипотезы abc, по крайней мере для взаимно простых u, v, k, поскольку $N_0(k) \leqslant |k|$. В более общем случае рассмотрим уравнение

$$Au^3 + Bv^2 = k$$

с ненулевыми целыми коэффициентами A, B. Тогда гипотеза abc дает такие же оценки, как в (6). Разумеется, внутренние константы, возникающие в неравенствах (6), будут тогда зависеть от A и B. И еще более общий случай, когда мы можем рассматривать уравнения более высоких степеней, а именно

$$Au^n + Bv^m = k.$$

Показатели n, m — положительные целые числа. Мы предполагаем, что $mn \neq m + n$. Затем, как легкое упражнение на применение гипотезы abc, выводится неравенство

$$|u|^n \ll (N_0(k))^{mn(1+\varepsilon)/[mn-(m+n)]}$$

и аналогично для $|v|^m$. Как уже упоминалось, в этом неравенстве внутренние константы зависят от A и B. Возвращаясь к $n = 3$ и $m = 2$, мы видим, что существуют специальные значения A и B, которые особенно интересны. Например, $A = -4$ и $B = -27$. Тогда выражение

$$\Delta = -4u^3 - 27v^2$$

хорошо известно как детерминант полинома

$$X^3 + uX + v.$$

Для этих специальных значений A и B соответствующая гипотеза известна как обобщенная гипотеза Шпиро (Szpiro). Вообще говоря, Шпиро в исходной гипотезе использовал не число N_0, а более сложный инвариант N, который возникает в теории эллиптических кривых, т. е. в теории уравнений вида

$$Y^2 = X^3 + uX + v.$$

Эту теорию сложно объяснить на уровне нашего обсуждения, и мы не будем этим заниматься. В любом случае Шпиро пришел к своей гипотезе через глубокое изучение и осмысление алгебраической геометрии и теории чисел. В то же время, из-за того, что я использовал только простой инвариант N_0, который легко определить, и поскольку мы связывали гипотезу Шпиро лишь с теоремой Мейсона–Стотерса и гипотезой abc, я не придерживался исторического порядка. В этом смысле я обманывал вас, поскольку я отбросил большие области математики, которые играли важную роль в историческом развитии всех гипотез. Из целого я выделил то, что может быть просто и понятно объяснено в течение часа. Но математики, которые получили эти гипотезы, не пришли к ним так прямо и просто, а только после тяжелой и продолжительной работы с этими глубокими и обширными теориями. (Для дальнейших приложений гипотезы abc в теории эллиптических кривых и дальнейших комментариев, см. [La 90], в которой также содержится более обширная библиография.)

В заключение я отмечу, что исходная гипотеза Шпиро в действительности относилась не к неравенству (6), а к более слабому неравенству

$$|\Delta| \ll (N_0(\Delta))^{6+\varepsilon}. \tag{7}$$

Более сильная гипотеза (6), которая дает оценки для $|u|$ и $|v|$, а не только для $|\Delta|$, была сформулирована лишь позднее. Поэтому мы ссылаемся на (6) как на обобщенную гипотезу Шпиро.

Таким образом, как видите, потребовалось много времени для развития гипотезы abc, чтобы осознать ее центральное положение в теории чисел и в теории уравнений. История следовала не по прямой линии, а проходила через обходные пути и аналогии, где теорема Давенпорта и теоремы о полиномах также, как и гипотеза Холла, сыграли свою роль. Но именно так развивается математика!

Приложение

[*Весной* 1998 *г. я имел возможность поговорить по телефону с учеником выпускного класса Ноа Снайдером, кото-*

рый интересовался математикой. Я рассказал ему о гипотезе abc и о том факте, что она была доказана для полиномов. Я предложил ему подумать об этой задаче и предложил попытаться построить доказательство для полиномов. Весьма примечательно, что он нашел свое собственное доказательство, которое проще, чем то, которое я узнал от Мейсона. Я очень благодарен Н. Снайдеру за разрешение опубликовать его доказательство в этой книге. Сейчас он студент Гарварда. С. Ленг.]

Другое доказательство гипотезы *abc* для полиномов, полученное Н. Снайдером.

Нам понадобится следующая лемма о полиномах, например с коэффициентами из комплексных чисел. Обозначим НОД двух полиномов f, g как

$$\text{НОД}(f, g) = (f, g).$$

Пусть $n_0(f)$ — количество различных корней полинома f.

Лемма. *Пусть f ненулевой полином. Тогда*

$$\deg f = \deg(f, f') + n_0(f).$$

Доказательство.

Во-первых, сделаем замечание о корнях полинома f. Пусть α — корень f, т.е. комплексное число, такое, что $f(\alpha) = 0$. Мы полагаем, что f — полином, не равный константе, поэтому f имеет степень $d > 0$. Мы можем записать $f(t)$ как полином от $(t - \alpha)$, т.е. как сумму убывающих степеней $(t - \alpha)$,

$$f(t) = c_d(t-\alpha)^d + \ldots + c_e(t-\alpha)^e,$$

с коэффициентами $c_d, \ldots, c_e$ и $e \geqslant 0$. Поскольку α — корень, то мы в действительности получаем $e \geqslant 1$ и $c_e \neq 0$. Тогда

$$f'(t) = dc_d(t-\alpha)^{d-1} + \ldots + ec_e(t-\alpha)^{e-1},$$

и $ec_e \neq 0$. Таким образом, мы получили, что $(t-\alpha)^{e-1}$ — наибольшая степень $(t-\alpha)$, делящая $f'(t)$. Теперь обозначим через $\alpha_1, \ldots, \alpha_r$ различные корни f и запишем

$$f(t) = c(t-\alpha_1)^{e_1} \ldots (t-\alpha_r)^{e_r} \text{ с константой } c \neq 0.$$

Затем только множители степени 1 выражения (f, f') должны иметь вид $(t-\alpha_j)$ для некоторого j, т.е. для некоторой ненулевой константы c',

$$(f, f') = c'(t-\alpha_1)^{k_1} \ldots (t-\alpha_r)^{k_r}.$$

Используя однозначность разложения на множители, первое замечание показывает, что $k_j = e_j - 1$ для каждого $j = 1, \ldots, r$. Тогда мы получим

$$\begin{aligned}\deg f &= e_1 + \ldots + e_r = \\ &= (e_1 - 1) + \ldots + (e_r - 1) + r = \deg(f, f') + n_0(f),\end{aligned}$$

что и доказывает лемму. ■

Теорема. *Пусть f, g, h — взаимно простые полиномы, такие, что $f + g = h$. Без потери общности, пусть h — полином наибольшей степени среди f, g, h. Тогда*

$$\deg(h) \leqslant n_0(fgh) - 1.$$

Доказательство.

Нам дано, что $f + g = h$. Поэтому $f' + g' = h'$. Таким образом,

$$f'g - fg' = f'(f+g) - f(f'+g') = f'h - fh'.$$

Заметим, что (f, f') делит левую часть, (g, g') делит левую часть и (h, h') делит правую часть. Поэтому, т.к. f, g, h взаимно просты,

$$(f, f')(g, g')(h, h') \quad \text{делит} \quad f'g - fg'.$$

Тогда

$$\begin{aligned}&\deg(f, f') + \deg(g, g') + \deg(h, h') \leqslant \\ &\qquad \leqslant \deg(f'g - fg') \leqslant \deg f + \deg g - 1.\end{aligned}$$

Прибавляя $\deg h$ к обеим частям и переставляя слагаемые, мы получаем, что

$$\deg h \leqslant \deg f - \deg(f, f') + \deg g - \deg(g, g') + \deg h - \deg(h, h') - 1.$$

Согласно лемме получаем

$$\deg h \leqslant n_0(f) + n_0(g) + n_0(h) - 1 = n_0(fgh) - 1,$$

т. к. f, g, h взаимно простые. Теорема доказана. ■

Литература

[BLSTW 83] J. Brillhart, D. H. Lehmer, J. L. Selfridge, B. Tuckerman, S. S. Wagstaff. Factorization of $b^n \pm 1$, $b = 2, 3, 5, 6, 7, 10, 11$ up to high powers, *Contemporary Math.*, Vol. 22, AMS, 1983.

[Da 65] H. Davenport. On $f^3(t) - g^2(t)$. *K. Norske Vid. Selsk. Forrh.* (Trondheim), **38** (1965), pp. 86–87.

[Ha 71] M. Hall. The diophantine equation $x^3 - y^2 = k$. *Computers in Number Theory* (A. O. L. Atkin, B. J. Birch, eds.), Academic press, 1971, pp. 173–198.

[La 87] S. Lang. *Undergraduate Algebra,* Springer-Verlag, 1987, 1990. See especially Chapter IV, § 9.

[La 90] S. Lang. Old and new conjectured diophantine inequalities. *Bull. AMS* **23** (1990), pp. 37–75.

[Ma 83] R. C. Mason. *Diophantine Equations over Function Fields.* London Math. Soc. Lecture Note Series, Vol. 96, Cambridge University Press, 1984.

[StT 86] C. L. Stewart, R. Tijdeman. On the Oesterle–Masser conjecture, *Monatshefte Math.* **102** (1986), pp. 251–257.

[St 81] W. Stothers, Polynomial identities and hauptmoduln. *Quart. Math. Oxford* (2) **32** (1981), pp. 349–370.

[Wi 95] A. Wiles. Modular elliptic curves and Fermat's Last Theorem. *Annals of Math.* **142** (1995), pp. 443–551.

[Za 95] U. Zannier. On Davenport's bound for the degree of $f^3 - g^2$ and Riemann's existence theorem. *Acta Arithm.* **LXXI.2** (1995), pp. 107–137.

Глобальное интегрирование локально интегрируемых векторных полей

Материал этой беседы происходит из представления Артина теоремы Коши, но в действительности принадлежит, в связи с элементарным вещественным анализом, к курсу математического анализа нескольких (вообще говоря двух) переменных. Не предполагается никаких знаний комплексного анализа. Некоторые элементарные знания о частных производных считаются известными, но не более, чем приведено в [La 87] глава VIII, § 4. Более полное обсуждение можно найти в [La 97] глава XV и XVI.

В основном используются прямоугольные пути. Если вы желаете априорно принять понятие о числе обходов для замкнутого прямоугольного пути, тогда вы можете прочесть теорему 5 и ее доказательство независимо от всего остального. Результат технически находится на уровне геометрии плоскости, хотя концептуально изложение материала находится на более усложненном уровне. В действительности это результат о замкнутых контурах на плоскости. Когда может замкнутый контур быть представлен как сумма границ прямоугольников?

Глобальное интегрирование локально интегрируемых векторных полей

Я собираюсь обсудить тему, аналог которой обычно рассматривается на курсах комплексного анализа, но эта тема в основном связана с вещественным анализом на плоскости $\mathbb{R}^2$. Пусть U — связное открытое множество в $\mathbb{R}^2$. Рассмотрим векторное поле F на U. Это означает, что F — это отображение

$$F\colon U \to \mathbb{R}^2,$$

заданное в координатных функциях

$$F(x, y) = (p(x, y), q(x, y)).$$

Мы полагаем здесь и в дальнейшем, что функции p и q, задающие компоненты вектора, имеют непрерывные частные производные, или, как говорится, принадлежат классу C^1. Пусть γ — кусочно гладкий C^1-путь в U. Когда мы говорим ***путь***, всегда подразумеваем параметризованный путь, т. е. C^1-кривая в U является отображением

$$\gamma\colon [a, b] \to U$$

с непрерывными производными. Кусочно гладкий C^1-путь является последовательностью C^1-кривых

$$\gamma_i\colon [a_i, b_i] \to U, \qquad i = 1, \ldots , r,$$

таких, что $\gamma_i(b_i) = \gamma_{i+1}(a_i)$ для $i = 1, \ldots , r - 1$. Другими словами, конечная точка i-го отображения является начальной точкой $(i + 1)$-го отображения. Таким образом путь проходит из начальной точки $\gamma_1(a_1)$ в конечную точку $\gamma_r(b_r)$. Разумеется, можно параметризовать весь путь заново так, чтобы он определялся на одном интервале $\gamma\colon [a, b] \to U$, где $b_i = a_{i+1}$ $(i = 1, \ldots , r - 1)$.

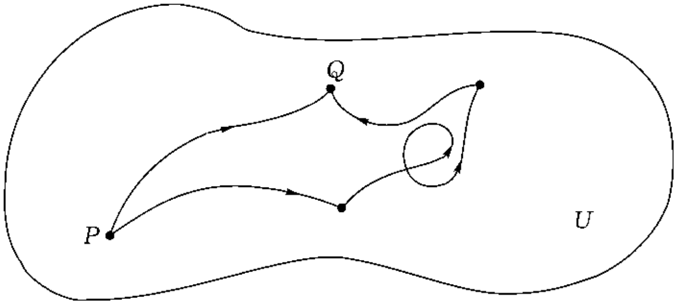
На нашем заданном векторном поле F мы можем определить интеграл от F вдоль пути

$$\int\limits_\gamma F = \sum_{i=1}^{r} \int\limits_i F.$$

Каждое слагаемое определено обычным способом

$$\int\limits_{\gamma_i} F = \int\limits_{a_i}^{b_i} F(\gamma_i(t))\gamma_i'(t)\,dt.$$

Интеграл от F по γ в общем случае зависит от пути, а не только от конечных точек.



Ситуация упрощается для важного класса векторных полей. Функция $\varphi\colon U \to \mathbb{R}$ называется *потенциалом* или *потенциальной функцией* векторного поля F, если F — градиент φ, т. е. $F = \operatorname{grad}\varphi$. Другими словами,

$$(p,\, q) = \operatorname{grad}\varphi = \left(\frac{\partial\varphi}{\partial x}, \frac{\partial\varphi}{\partial y}\right).$$

Векторное поле с потенциалом называется *консервативным* векторным полем. Простая теорема, стандартная в курсах анализа гласит, что *если F — консервативное, φ — потенциальная функция и P и Q — начальная и конечная точки пути γ, тогда*

$$\int\limits_{\gamma} F = \varphi(0) - \varphi(P).$$

Поэтому в этом случае интеграл не зависит от пути между конечными точками. Обратное тоже верно. Точная формулировка:

Теорема 1. *Пусть F — векторное поле на открытом связном множестве U в $\mathbb{R}^2$. Тогда F имеет потенциальную функцию на U тогда и только тогда, когда при двух заданных точках P и Q в U и для любого пути γ из P в Q, лежащего в U, интеграл $\int\limits_{\gamma} F$ не зависит от γ.*

Теорема 1 — это фундаментальный результат, обычно доказываемый в стандартных курсах анализа. Идея заключается в том, что фиксируется точка P_0 и определяется

$$\varphi(Q) = \int_{P_0}^{Q} F$$

вдоль произвольного пути из P_0 в Q. Если предположить, что интеграл не зависит от пути, получается, что этот интеграл является хорошо определенной функцией, для которой легко доказать, что $\operatorname{grad}\varphi = F$. Мы не будем здесь воспроизводить доказательство. Смотри к примеру [La 83/97] глава XV, теорема 4.2 или [La 87], глава VIII, § 4.

Также легко доказать, что две потенциальные функции на U отличаются на константу, т. е. разность равна константе. Смотри [La 87], глава VII, теорема 1.1. Поэтому потенциальная функция является однозначно определенной с точностью до константы. Заметим, что эта константа интегрирования исчезает, когда мы, вычисляя разность $\varphi(Q) - \varphi(P)$, получим значение интеграла от F, когда F имеет потенциальную функцию φ.

Можно по-другому переформулировать теорему 1. Путь называется *замкнутым*, если конечная точка совпадает с начальной точкой.

Теорема $1'$. *Пусть U — открытое связное множество в $\mathbb{R}^2$ и пусть F — векторное поле на U. Тогда F обладает потенциальной функцией тогда и только тогда, когда для всех замкнутых путей γ на U мы имеем $\int_\gamma F = 0$.*

Чтобы доказать эквивалентность теоремы 1 и теоремы $1''$, заметим, что при движении по пути из P в Q, а затем по пути из Q в P, получаем замкнутый путь. Доказательство простое и обычно приводится в курсах анализа.

Теперь предположим, что векторное поле принадлежит классу C^2, другими словами, его координатные функции $p(x, y)$ и $q(x, y)$ имеют непрерывные частные производные второго порядка. Предположим, что F имеет потенциальную функцию φ, поэтому по определению

$$p(x, y) = \frac{\partial\varphi}{\partial x} \quad \text{и} \quad q(x, y) = \frac{\partial\varphi}{\partial y}.$$

Элементарным фактом анализа является то, что

$$\frac{\partial p}{\partial y} = \frac{\partial q}{\partial x}, \tag{2}$$

поскольку частные производные коммутируют, т. е.

$$\frac{\partial p}{\partial y} = \frac{\partial^2 \varphi}{\partial y\, \partial x} = \frac{\partial^2 \varphi}{\partial x\, \partial y} = \frac{\partial q}{\partial x}.$$

Поэтому условие (2) является необходимым условием для того, чтобы F имело потенциальную функцию. Это условие называется *условием интегрируемости.* В курсах элементарного анализа обычно доказывается, что локально условия интегрируемости достаточно для существования потенциальной функции. Более точно:

Теорема 2. *Пусть F — векторное поле класса C^2 на U, удовлетворяющее условию интегрируемости* (2). *Пусть R — прямоугольник, содержащийся в U (соответственно D — круг, содержащийся в U), тогда F имеет потенциальную функцию φ на R (соответственно на D).*

Доказательство.

Зафиксируем точку (x_0, y_0) в R, скажем, центр прямоугольника, или в случае круга пусть (x_0, y_0) будет центром круга. Для точки (x, y) в R (соответственно в D) определим

$$\varphi(x, y) = \int_{x_0}^{x} p(t, y)\, dt + \int_{y_0}^{y} q(x_0, u)\, du.$$

Второй интеграл не зависит от x. По основной теореме анализа получаем тождество

$$\frac{\partial \varphi}{\partial x} = p(x, y),$$

которое является первым из двух условий для потенциальной функции. Чтобы найти частную производную $\dfrac{\partial \varphi}{\partial y}$, мы должны продифференцировать под знаком интеграла, что вы должны знать из курса элементарного математического анализа. Затем мы получим

$$\begin{aligned}\frac{\partial}{\partial y}\varphi(x, y) &= \int\limits_{x_0}^{x} \frac{\partial}{\partial y}\varphi(t, y)\,dt + q(x_0, y) = \\ &= \int\limits_{x_0}^{x} \frac{\partial}{\partial t}q(t, y)\,dt + q(x_0, y) = \\ &= q(x, y) - q(x_0, y) + q(x_0, y) = q(x, y).\end{aligned}$$

Что и доказывает теорему. ■

Векторное поле, которое удовлетворяет условию интегрируемости (2), будем называть *локально интегрируемым* в свете теоремы 2. Можно ослабить это определение, потребовав лишь, чтобы у данной точки существовала окрестность, в которой векторное поле имеет потенциальную функцию, но для простоты мы сохраним данное ранее определение.

Целью нашего текущего обсуждения является исследование условий, при которых F имеет потенциальную функцию на всем U. Другими словами, нам нужна глобальная интегрируемость. В общем случае условия интегрируемости не достаточно. Существует стандартный пример, а именно векторное поле, заданное выражением

$$G(x, y) = \left(\frac{-y}{x^2+y^2}, \frac{x}{x^2+y^2}\right) = \left(\frac{-y}{r^2}, \frac{x}{r^2}\right), \text{ где } r^2 = x^2 + y^2.$$

Разумеется, векторное поле не определено в точке начала координат $(0, 0) = O$. Оно определено на плоскости с выколотой точкой $\mathbb{R}^2 - \{O\}$. Простые вычисления показывают, что $G(x, y)$ удовлетворяет условию интегрируемости (2) и поэтому локально интегрируемо. Однако, G не имеет потенциальной функции на всей области определения. Чтобы показать это, мы рассмотрим замкнутую кривую γ такую, что

$$\int\limits_{\gamma} G \neq 0.$$

В действительности, как мы сейчас покажем, можно выбрать в качестве γ любую окружность с центром в начале координат.

Прежде всего вычислим $p\,dx + q\,dy$ в полярных координатах. При этом оказывается, что для $p = \dfrac{-y}{r^2}$, $q = \dfrac{x}{r^2}$

$$p\,dx + q\,dy = d\theta.$$

Действительно, положим $x = r\cos\theta$ и $y = r\sin\theta$ в полярных координатах. Тогда

$$dx = \frac{\partial x}{\partial r}\,dr + \frac{\partial x}{\partial \theta}\,d\theta = \cos\theta\,dr - r\sin\theta\,d\theta,$$

$$dy = \frac{\partial y}{\partial r}\,dr + \frac{\partial y}{\partial \theta}\,d\theta = \sin\theta\,dr - r\cos\theta\,d\theta.$$

Простые упрощения показывают, что

$$\left(\frac{-y}{r^2}\right)dx + \left(\frac{x}{r^2}\right)dy = d\theta.$$

Из этих вычислений следует, что полярный угол θ является локальной потенциальной функцией для G. Если R — прямоугольник в $\mathbb{R}^2 - \{O\}$ или D — круг в $\mathbb{R}^2 - \{O\}$, тогда θ — потенциал для G в R, соответственно в D.

Пусть γ — путь, не содержащий точку начала координат O. Тогда

$$\int\limits_{\gamma} G = \int\limits_{\gamma} d\theta.$$

Если γ — замкнутый путь в проколотой плоскости $\mathbb{R}^2 - \{O\}$, тогда вдоль пути θ изменяется от начальной точки до конечной точки на интеграл, кратный 2π. Таким образом,

$$\int\limits_{\gamma} G = 2\pi k, \tag{3}$$

где k — целое число. В частности, для окружности с центром в начале координат, ориентированной против часовой стрелки, скажем, для единичной окружности $x = \cos\theta$, $y = \sin\theta$, мы получим

$$G = \int_0^{2\pi} d\theta = 2\pi \neq 0.$$

Как вы можете проверить, то же самое выполнено для любой окружности с центром в начале координат. Из этого, в частности, следует, что G не имеет потенциала на всей $\mathbb{R}^2 - \{O\}$.

Отношение (3) является фундаментальным и служит основой для следующего определения. Пусть γ — замкнутый путь на плоскости, не содержащий начала координат O. Мы определим *число обходов* пути

$$W(\gamma, O) = \frac{1}{2\pi} \int_\gamma G.$$

Это число обходов относительно начала координат. Для произвольной точки P мы можем определить число обходов, выполнив параллельный перенос. Пусть $P = (x_0, y_0)$. Пусть G_P — векторное поле, определенное как

$$G_P(x, y) = G(x - x_0, y - y_0).$$

Пусть γ — замкнутый путь, не содержащий P. Мы определим число обходов $W(\gamma, P)$ формулой

$$W(\gamma, P) = \frac{1}{2\pi} \int_\gamma G_P.$$

Позднее мы увидим, что G и его переносы G_P составляют единственные препятствия для того, чтобы локально интегрируемое векторное поле имело глобальную потенциальную функцию. Этот фундаментальный результат основан на следующей теореме, которая, как будет показано, представляет собой чисто топологический результат, но сначала мы выразим ее аналитически.

Теорема 3. *Пусть U — открытое связное множество в $\mathbb{R}^2$. Пусть γ — замкнутый путь в U. Предположим, что для всех точек $P \notin U$ число обходов $W(\gamma, P)$ равно 0. Пусть F — локально интегрируемое векторное поле на U. Тогда $\int_\gamma F = 0$.*

В действительности, вместо того, чтобы работать с путями в строгом смысле, как мы их определили, удобнее работать с формальными линейными комбинациями

$$\gamma = \sum_{i=1}^{n} m_i \gamma_i,$$

где каждый γ_i — замкнутый путь и m_i — целое число для каждого i. Интеграл от F по γ определяется просто как сумма интегралов:

$$\int_\gamma F = \sum_{i=1}^{n} m_i \int_{\gamma_i} F.$$

Такая формальная линейная комбинация будет называться *замкнутой цепью*. Тогда и теорема, и лемма выполняются для замкнутых цепей.

Пример. Пусть γ_1 — окружность радиуса 1 с центром в начале координат и γ_2 — окружность радиуса 2. Пусть

$$\gamma = \gamma_1 - \gamma_2.$$

Каждая из окружностей ориентирована против часовой стрелки. Возьмем $F = G$ — векторное поле, определенное перед теоремой 3, и пусть $U = \mathbb{R}^2 - \{O\}$ — плоскость с выколотой точкой. Единственная точка, не принадлежащая U, — это начало координат, и тогда число обходов $W(\gamma, O)$ будет равняться

$$2\pi W(\gamma, O) = \int_{\gamma_1} G - \int_{\gamma_2} G = 2\pi - 2\pi = 0.$$

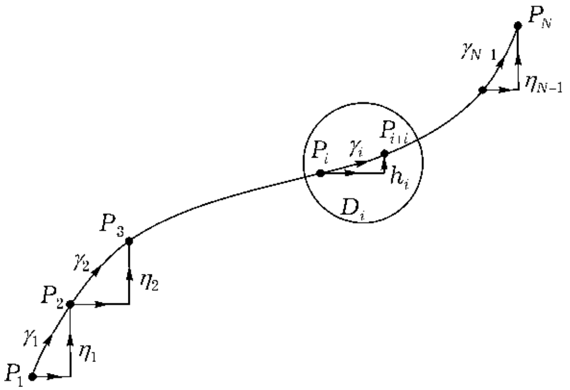
Теперь мы приведем доказательство теоремы 3. Затем мы покажем ее приложения.

Доказательство основной теоремы

Доказательство будет проходить в несколько этапов. На первом шаге сведем теорему к тому, что мы называем прямоугольными путями. Более точно мы будем называть путь *прямоугольным*, когда он состоит из конечного количества отрезков прямых, которые параллельны координатным осям, т. е.

отрезки прямых, которые горизонтальны или вертикальны. Тогда мы получим лемму редукции:

Лемма 4. *Если теорема 3 выполняется для прямоугольных замкнутых цепей, тогда она будет выполняться для произвольных замкнутых цепей (всегда подразумеваем кусочно непрерывно дифференцируемые).*



Чтобы доказать лемму, нам нужно сначала показать, как произвольный путь γ в U может быть заменен прямоугольным путем η в U, имеющим ту же самую начальную точку и ту же самую конечную точку и таким, что для любого локально интегрируемого векторного поля F в U мы получаем

$$\int_\eta F = \int_\gamma F.$$

Предположим, что $\gamma\colon [a,\, b] \to U$ определен над замкнутым интервалом $[a,\, b]$. Ввиду равномерной непрерывности γ можно использовать разбиение интервала

$$a = a_1 \leqslant a_2 \leqslant \ldots \leqslant a_N = b,$$

причем $a_{i+1} - a_i$ достаточно мало, так что образ $[a_i,\, a_{i+1}]$ в γ содержится в круге $D_i \subset U$. Пусть $\gamma_i\colon [a_i,\, a_{i+1}] \to U$ и пусть $P_i = \gamma(a_i)$.

Теперь зададим прямоугольный путь, заменяя каждое γ_i двумя отрезками прямых, параллельных осям, как показано на рисунке. Таким образом, η_i также содержится в D_i, и $\eta_1, \ldots, \eta_{N-1}$ составляют прямоугольный путь η из P_1 до P_N. Более того, для каждого i путь η_i также содержится в D_i. Поскольку P_i, P_{i+1} — начальная и конечная точка как пути γ_i, так и η_i и поскольку F — локально интегрируема, то, следовательно,

$$\int\limits_{\gamma_i} F = \int\limits_{\eta_i} F \quad \text{для} \quad i = 1, \ldots, N-1.$$

Поэтому

$$\int\limits_{\gamma} F = \int\limits_{\eta} F.$$

В частности, это применимо к векторному полю G и поэтому

$$\int\limits_{\gamma} G = \int\limits_{\eta} G.$$

По определению, это означает, что число обходов

$$W(\eta, P) = W(\gamma, P) = 0$$

для любой произвольной точки P не в U. Поэтому если теорема 3 верна для прямоугольного пути η, то она также верна для γ, и лемма 4 доказана.

Осталось доказать теорему 3 для прямоугольных путей или цепей.

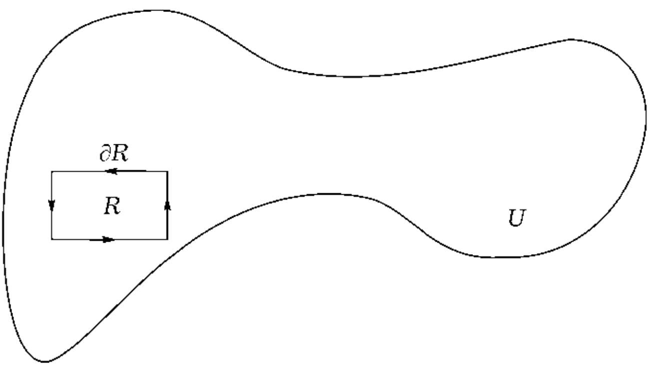
Предположим сначала, что R — замкнутый прямоугольник, содержащийся в U. Пусть ∂R — его граница, ориентированная против часовой стрелки, как показано на рисунке.

Т. к. векторное поле F локально интегрируемо, то из теоремы 2 следует, что

$$\int\limits_{\partial R} F = 0.$$

Если путь или цепь заданы как сумма границ прямоугольников, скажем

$$\eta = \sum_i m_i \, \partial R_i,$$



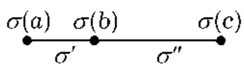
и каждый прямоугольник содержится в U, тогда

$$\int_{\eta} F = \sum_i m_i \int_{\partial R_i} F = 0.$$

Из фундаментального результата, содержащегося в следующей теореме, мы получим то, что нам нужно для замкнутых прямоугольных цепей. По лемме 4, тем самым мы также докажем теорему 3.

Прямоугольные замкнутые пути. Теорема о контурах на плоскости

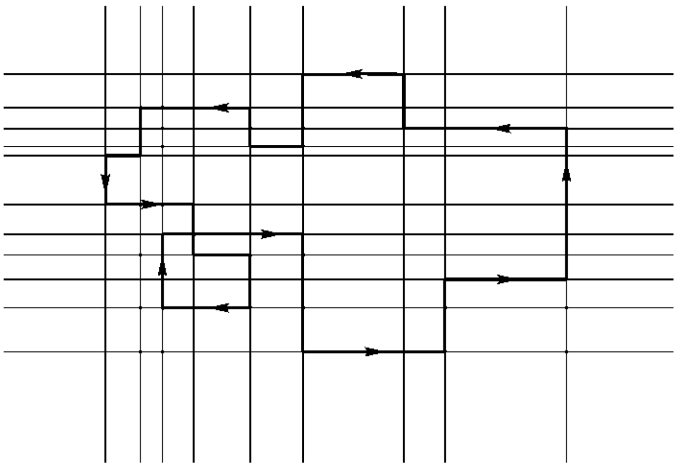
Пусть U — открытое связное множество в $\mathbb{R}^2$ и пусть η — прямоугольный путь в U. Тогда η состоит из последовательных горизонтальных и вертикальных отрезков. Пусть σ — такой отрезок, скажем $\sigma\colon [a, b] \to U$. Подберем $a < c < b$. Тогда σ делится на два отрезка $\sigma'\colon [a, c] \to U$ и $\sigma''\colon [c, b] \to U$:



Если мы заменим σ на $\sigma' + \sigma_1''$ и повторим эту процедуру конечное число раз, используя различные отрезки, составляющие η, мы получим то, что мы будем называть *разбиением* пути η и обозначим $\mathrm{Subd}\,\eta$. Мы теперь можем сформулировать фундаментальную теорему, в которой появляется только число обходов, но за исключением этого ничего о векторных полях.

Теорема 5. *Пусть η — замкнутая прямоугольная цепь в U. Предположим, что $W(\eta, P) = 0$ для любой точки P, не лежащей в U. Тогда существует конечное количество прямоугольников R_i $(i = 1, \dots, N)$, содержащихся в U и целые числа m_i такие, что для некоторого разбиения пути η мы имеем*

$$\operatorname{Subd} \eta = \sum_i m_i \, \partial R_i.$$



Доказательство.

Задана прямоугольная цепь η. Мы рисуем все вертикальные и горизонтальные линии, проходящие через стороны этой цепи, как показано на следующем рисунке. Тогда эти вертикальные и горизонтальные линии разделяют плоскость на прямоугольники и прямоугольные области, продолжающиеся в бесконечность в вертикальном и горизонтальном направлениях. Пусть R_i — один из прямоугольников и пусть P_i — точка внутри R_i. Пусть

$$m_i = W(\eta, P_i).$$

Для некоторых прямоугольников $m_i = 0$, а для некоторых прямоугольников $m_i \neq 0$. $R_1, \ldots, R_N$ — те прямоугольники, для которых $m_1, \ldots, m_N$ не равны 0, и пусть ∂R_i — граница R_i для $i = 1, \ldots, N$, ориентированная против часовой стрелки. Нам нужно доказать два утверждения:

1°. Всякий прямоугольник R_i такой, что $m_i \neq 0$ содержится в U.

2°. Некоторое разбиение η равно

$$\sum_{i=1}^{N} m_i \, \partial R_i.$$

Тем самым будет доказана нужная нам теорема.

Утверждение 1. По предположению P_i должно быть в U, поскольку $W(\eta, \rho) = 0$ для всякой точки P, не лежащей в U. Заметим, что функция количества обходов

$$P \mapsto W(\eta, P)$$

является непрерывной и поэтому является постоянной на связных множествах, поскольку она принимает целые значения. Поэтому для всех P во внутренности R_i будем иметь $W(\eta, P) = W(\eta, P_i) \neq 0$. Поэтому внутренность R_i содержится в U. Если граничная точка R_i лежит на η, тогда она находится в U. Если граничная точка R_i не лежит на η, тогда количество обходов относительно η определено и снова не равно 0 по непрерывности, приближаясь к граничной точке из внутренности. Это доказывает, что весь прямоугольник, включая границу, содержится в U, и тем самым доказано наше первое утверждение.

Утверждение 2. Для доказательства этого утверждения мы прежде всего отметим, что если P_k — точка вне прямоугольника R_i, то

$$W(\partial R_i, P_k) = 0.$$

Однако для Q внутри R_i будет выполнено $W(\partial R_i, Q) = 1$.

Теперь мы перейдем к доказательству этого утверждения. Все вертикальные и горизонтальные линии, проходящие через нашу цепь η, могут рассечь некоторые из горизонтальных и вертикальных путей, возникающих в η, на несколько частей. Таким образом формируется разбиение η. Для целей интегрирования это разбиение равнозначно самому η. Теперь мы рассмотрим цепь

$$C = \eta - \sum_i m_i \, \partial R_i.$$

Утверждение 2 равносильно тому, что цепь C не содержит никакого отрезка, другими словами, тому, что цепь C является нулевой цепью. Предположим, что какой-то отрезок σ находится в C и σ — сторона прямоугольника R_k. Тогда мы сможем записать

$$C = \eta - \sum_i m_i \, \partial R_i = m\sigma + \text{слагаемые без } \sigma,$$

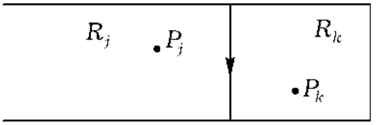
где m — какое-то целое число, которое не равно 0, если σ содержится в C. В действительности мы должны доказать, что $m = 0$, поэтому ни одного отрезка не может остаться в C. Мы рассмотрим цепь

$$C' = \eta - \sum m_i \, \partial R_i - m \, \partial R_k = C - m \, \partial R_k,$$

и вычислим количество обходов относительно P_k, которая лежит внутри R_k. Мы получим

$$\begin{aligned} &W(\eta - \sum_i m_i \, \partial R_i - m \, \partial R_k, P_k) = \\ &= W(\eta, P_k) - \sum_i m_i W(\partial R_i, P_k) - mW(\partial R_k, P_k) = \\ &= m_k - m_k - m = -m. \end{aligned}$$

Однако цепь C' больше не содержит σ, поскольку σ является стороной R_k и мы вычли $m\sigma$ из C, вычитая $m \, \partial R_k$ из C. В нашем разбиении плоскости на прямоугольники σ принадлежит границе двух прямоугольников. Если R_k находится по одну сторону σ, тогда какой-то прямоугольник R_j находится по другую сторону, как показано на рисунке.



Точки P_k и P_j можно соединить отрезком прямой, который не имеет ни единой точки, общей с цепью C'. Поэтому

$$\begin{aligned}-m &= W(\eta - \sum m_i \, \partial R_i - m \, \partial R_k, \ P_k) = \\ &= W(\eta - \sum m_i \, \partial R_i - m \, \partial R_k, \ P_j) = \\ &= W(\eta, P_j) - \sum_i m_i W(\partial R_i, P_j) - mW(\partial R_k, P_j) = \\ &= m_j - m_j - 0 = 0.\end{aligned}$$

Это доказывает утверждение 2 и завершает доказательство теоремы 5. ■

Замечание. Приведенное выше доказательство не содержит никаких сведений из анализа, за исключением определения количества обходов, которое было удобно задать с помощью интеграла. За исключением этого аргументы носят чисто комбинаторный характер, и теорема является результатом о контурах на плоскости.

Два приложения

Мы можем применить теорему 3 двумя способами. Первый покажет, как интеграл вдоль замкнутого пути или цепи может быть вычислен как сумма интегралов по маленьким окружностям. Другой приведет к условию, показывающему, что единственное препятствие для того, чтобы локально интегрируемое векторное поле имело потенциальную функцию, возникает из-за нашего векторного поля G и переносов G_P на определенные точки P.

Теорема 6. *Пусть U — открытое множество в $\mathbb{R}^2$ и пусть γ — замкнутая цепь в U, такая, что $W(\gamma, P) = 0$ для всех $P \notin U$. Пусть $P_1, \ldots, P_n$ — конечное число различных* *точек U, не принадлежащих γ. Пусть γ_i $(i = 1, \ldots, n)$ граница замкнутого круга $\overline{D}_i$, содержащегося в U, содержащего P_i и ориентированного против часовой стрелки. Мы предполагаем, что $\overline{D}_i$ не пересекает γ и не пересекает $\overline{D}_j$, если $i \neq j$. Пусть*

$$m_i = W(\gamma, P_i).$$

Пусть U^* — множество, полученное удалением $P_1, \ldots, P_n$ из U. Пусть F — локально интегрируемое векторное поле на U^*. Тогда

$$\int_\gamma F = \sum_{i=1}^{n} m_i \int_{\gamma_i} F.$$

Доказательство.

Пусть

$$C = \gamma - \sum m_i \gamma_i.$$

Мы утверждаем, что для всех точек P вне U^* выполнено $W(C, P) = 0$. Если $P \notin U$, тогда

$$W(C, P) = W(\gamma, P) - \sum m_i W(\gamma_i, P) = 0 - 0 = 0.$$

Затем для одной из точек P_k вместо P будет выполнено

$$W(\gamma_i, P_k) = \begin{cases} 1 & \text{если } i = k, \\ 0 & \text{если } i \neq k. \end{cases}$$

Следовательно,

$$W(C, P_k) = W(\gamma, P_k) - \sum m_i W(\gamma_i, P_k) = W(\gamma, P_k) - m_k = 0.$$

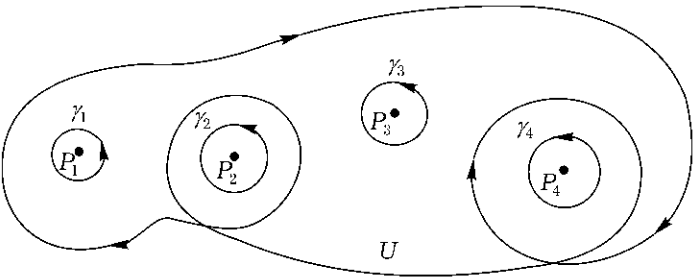
Тогда мы применяем теорему 3 к открытому множеству U^* и цепи C. Мы получим

$$0 = \int_C F = \int_\gamma F - \sum_{i=1}^{n} m_i \int_{\gamma_i} F,$$

что и доказывает теорему 6. ■

Пример 1. На следующем рисунке выполнено

$$\int_{\gamma} F = -\int_{\gamma_1} F - 2\int_{\gamma_2} F - \int_{\gamma_3} F - 2\int_{\gamma_4} F.$$



Пусть P — точка на $\mathbb{R}^2$, и пусть D_P — круг с центром в P. Пусть D_P^* — круг с дыркой, полученный удалением P. Пусть γ_P — окружность с центром в P, лежащая в круге D_P и ориентированная против часовой стрелки. Пусть F — локально интегрируемое векторное поле на D_P^*. *Вычетом* F в точке P мы будем называть

$$\operatorname{res}_P(F) = \frac{1}{2\pi}\int_{\gamma_P} F.$$

Вспоминая теорему 6, мы видим, что этот вычет не зависит от выбора окружности γ_P. Применение теоремы 6 приводит к формуле

$$\int_{\gamma} F = \sum_i 2\pi m_i \cdot \operatorname{res}_{P_i}(F).$$

Другими словами интеграл F вдоль γ равен 2π умножить на сумму вычетов, сосчитанных с кратностями m_i.

Эта терминология используется в теории функций комплексного переменного, где вы обнаружите теорему, соответствующую теореме 6, сформулированную для определенных комплексных функций. Но когда вы будете изучать курс теории функций комплексного переменного, вам следует проверить для себя, что теорема 6, которую мы доказали, содержит соответствующую теорему о комплексных функциях как частный случай. Эта теорема называется *теоремой о вычетах* или *теоремой Коши о вычетах.*

Мы перейдем ко второму приложению, приведя препятствие к существованию потенциальной функции.

Теорема 7. *Пусть U — связное открытое множество в $\mathbb{R}^2$, такое, что для каждого замкнутого пути γ в U и в каждой точки $P \notin U$ выполнено*

$$W(\gamma, P) = 0.$$

Пусть $P_1, \ldots, P_n$ — различные точки U, и пусть $U^* = U - \{P_1, \ldots, P_n\}$ — множество, полученное удалением $P_1, \ldots, P_n$ из U. Пусть F — локально интегрируемое векторное поле на U^*. Как обычно, пусть

$$G(x, y) = \left(\frac{-y}{x^2+y^2}, \frac{x}{x^2+y^2}\right).$$

Пусть G_P — перенос G в точку P. Тогда существуют константы $a_1, \ldots, a_n$ и функция φ на U^* такая, что

$$F - \sum_{i=1}^{n} a_i G_{P_i} = \operatorname{grad} \varphi.$$

Доказательство.

Пусть γ_i — маленькие окружности с центрами в точках P_i, ориентированные против часовой стрелки. Мы предполагаем, что окружности достаточно малы, чтобы при $k \neq i$ замкнутые круги, ограниченные γ_i и γ_k, отделены друг от друга. Тогда выполнено

$$\int_{\gamma_i} G_{P_j} = \begin{cases} 2\pi & \text{если } i = j, \\ 0 & \text{если } i \neq j. \end{cases} \qquad (*)$$

Пусть $a_i = \frac{1}{2\pi} \int_{\gamma_i} F = \operatorname{res}_{P_i}(F)$ по нашей последней терминологии. Пусть

$$F_1 = F - \sum_{i=1}^{n} a_j G_{P_j}.$$

По теореме 1′ и теореме 3 достаточно доказать, что для всех замкнутых путей γ в U^* выполнено

$$\int\limits_{\gamma} F_1 = 0.$$

Но используя теорему 6 и условия (∗), получим

$$\begin{aligned}\int\limits_{\gamma} F_1 &= \sum_i m_i \int\limits_{\gamma_i} F_1 = \sum_i m_i \int\limits_{\gamma_i} \Big(F - \sum_j a_j G_{P_j}\Big) = \\ &= \sum_i m_i 2\pi a_i - \sum_i \sum_j \int\limits_{\gamma_i} m_i a_j G_{P_j} = \\ &= \sum_i m_i 2\pi a_i - \sum_i m_i a_i 2\pi = 0,\end{aligned}$$

что доказывает теорему 7. ∎

Литература

[Ar 65] E. Artin, On the theory of complex functions, *Collected papers of Emil Artin*, Addison-Wesley, 1965, pp. 513–522. Reprinted by Springer-Verlag, Second edition, 1996.

[La 87] S. Lang, *Calculus of Several Variables,* Springe-Verlag, 1987.

[La 97] S. Lang, *Undergraduate Analysis,* Second edition, Springer-Verlag, 1997.

[La 99] S. Lang, *Complex Analysis,* Fouth edition, Springer-Verlag, 1999.

Теоремы анализа об аппроксимации

Анализ в большой степени заключается в получении оценок. Но также существует часть анализа, которая занимается проверкой тождественности различных выражений, что в сущности представляет собой алгебраическую структуру, в отличие от оценок. Последовательности Дирака и семейства Дирака используют оба аспекта.

Примеры показывают широкую область применения. Кроме того, я считаю естественным упомянуть тэта-ряды в связи с совершенно различными контекстами, чтобы показать, как казалось бы различные части математики, в действительности, близко связаны. Курсы обычно посвящаются лишь одной из областей математики из-за ограниченности по времени, а также потому, что они являются подготовкой к другим курсам, так что обычно не хватает времени, чтобы смотреть в окно и бродить без цели. В настоящей книге у нас нет ни фиксированного расписания, ни программы курса, ни контрольных работ и окружение совершенно другое чем то, что во время курса. Этот материал дает возможность провести целую серию бесед.

Теоремы анализа об аппроксимациях

Я собираюсь начать с очень общей теоремы об аппроксимации, которая относится к так называемым последовательностям Дирака. Затем я собираюсь обсудить примеры и приложения, такие как:

— Теорема Вейерштрасса об аппроксимации, в которой говорится, что непрерывная функция в конечном замкнутом интервале равномерно аппроксимируется полиномами.

— Ряды Фурье.

— Гармонические функции и семейства Пуассона.

— Решения уравнения теплопроводности.

В каждом случае частный результат в каждой конкретной области будет представлен как специальный случай общей теоремы.

За исключением теоремы Вейерштрасса об аппроксимации, эти результаты были изначально открыты в областях, общих для физики и математики. Но только потому, что что-то было открыто в связи с физикой, не означает, что это физика. Все из вышеперечисленных результатов имели значение во многих областях математики, возможно даже во всех ее областях. Я хотел бы сказать, к примеру, что тепловой источник, который мы обсудим позднее, является большим взрывом математической вселенной. Вы увидите, что он используется повсюду.

Теперь вернемся к математике. Мы работаем с вещественной прямой $\mathbb{R}$. Начнем с определения свертки двух функций. Пусть f и g — две функции на $\mathbb{R}$ (вещественные числа), непрерывные или кусочно-непрерывные, комплекснозначные и такие, что при $x \to \pm\infty$ значения $f(x)$ стремятся к нулю достаточно быстро, так, что все следующие интегралы сходятся. Например, если $f(x) = 0$ вне конечного интервала, тогда f удовлетворяет вышеперечисленным требованиям. Мы определим ***свертку*** $f * g$ формулой

$$f * g(x) = \int\limits_{-\infty}^{+\infty} f(x-t)g(t)\,dt.$$

Очевидно, что свертка коммутативна, т. е.

$$f * g = g * f.$$

Мы просто сделаем замену $u = x - t$ и $t = x - u$, $du = -dt$. Рассмотрим свертку как произведение, которое является билинейным, т. е.

$$f * (g_1 + g_2) = f * g_1 + f * g_2 \quad \text{и} \quad (f_1 + f_2) * g = f_1 * g + f_2 * g.$$

Более того, для любого комплексного числа α

$$(\alpha f) * g = f * (\alpha g) = \alpha(f * g).$$

Также, меняя порядок интегрирования, можно проверить ассоциативность трех функций f, g, h, а именно

$$(f * g) * h = f * (g * h).$$

Поэтому свертка удовлетворяет обычным правилам умножения. Будем говорить, что кусочно непрерывные функции с достаточно быстрым вырождением на бесконечности образуют коммутативную алгебру над $\mathbb{C}$, и вещественнозначные функции образуют коммутативную алгебру над $\mathbb{R}$.

Возникает первый вопрос: существует ли единичный элемент в этих алгебрах? Другими словами, существует ли функция δ такая, что $\delta * f = f$ для всех функций f, скажем, непрерывных функций, обращающихся в нуль вне некоторого конечного интервала? Кто скажет «да»? [*Несколько рук поднимается.*] Кто скажет «нет»? [*Снова поднимается несколько рук. Большинство хранит молчание.*] Я дам вам ответ. Ответ — НЕТ. Но мы собираемся определить нечто, почти столь же замечательное, как единичный элемент, а именно последовательность Дирака.

По определению, *последовательность Дирака* — это последовательность непрерывных функций $\{K_n\}$, обладающих следующими свойствами:

DIR 1. Функции K_n — неотрицательные, т. е. $K_n \geqslant 0$ при всех $n = 1, 2, \ldots$

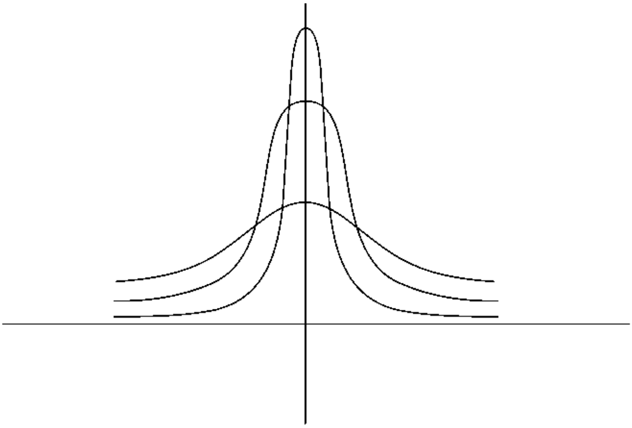
DIR 2. При всех n выполнено

$$\int_{-\infty}^{+\infty} K_n = 1.$$

DIR 3. При заданных $\varepsilon > 0$ и $\delta > 0$ существует n_0 такое, что при всех $n \geqslant n_0$ выполнена оценка

$$\int_{|x| \geqslant \delta} K_n(x)\, dx < \varepsilon.$$

Последовательность Дирака может выглядеть так, как показано на следующем рисунке.



На практике последовательности Дирака состоят из четных функций, т. е. график симметричен относительно вертикальной оси, как нарисовано на рисунке.

Последовательности Дирака обладают примечательным аппроксимирующим свойством, которое мы сейчас сформулируем и докажем.

Теорема 1 (Общая теорема об аппроксимации). *Пусть f — ограниченная кусочно-непрерывная функция на $\mathbb{R}$. Тогда последовательность $\{K_n * f\}$ сходится к f равномерно на любом компактном множестве, где f непрерывна. То есть для любой точки x, в которой f непрерывна, выполнено*

$$\lim_{n\to\infty}(K_n * f)(x) = f(x),$$

и сходимость равномерная, как сказано выше.

В некотором смысле, последовательность $\{K_n\}$ «сходится» к единичному элементу, хотя и очевидно из рисунка, что последовательность не имеет предельной функции. Из-за вышеуказанного свойства предела обычно говорят, что последовательность Дирака является *аппроксимацией единичного элемента.*

Доказательство.

Примечательно, что доказательство весьма простое. По определению

$$K_n * f(x) = f * K_n(x) = \int\limits_{-\infty}^{+\infty} K_n(t) f(x-t)\,dt.$$

По свойству DIR 2 получаем

$$f(x) = f(x)\int\limits_{-\infty}^{+\infty} K_n(t)\,dt = \int\limits_{-\infty}^{+\infty} K_n(t) f(x)\,dt.$$

Вычитая, получим

$$K_n * f(x) - f(x) = \int\limits_{-\infty}^{+\infty} K_n(t)[f(x-t) - f(x)]\,dt.$$

Пусть значение B ограничивает функцию f на $\mathbb{R}$, т.е.

$$|f(x)| \leqslant B \quad \text{при всех } x \in \mathbb{R}.$$

Пусть S — компактное множество, на котором f непрерывна. Тогда f равномерно непрерывна на S, и это единственное свойство, которое мы будем использовать. Поэтому при заданном ε, существует δ такое, что при всех $x \in S$ и $|t| \leqslant \delta$ выполнено

$$|f(x-t) - f(x)| < \varepsilon.$$

Затем мы оценим интеграл свертки, который мы разобьем на две части:

$$|K_n * f(x) - f(x)| = \int\limits_{|t|\leqslant\delta} K_n(t)|f(x-t) - f(x)|\,dt +$$

$$+ \int\limits_{|t|\geqslant\delta} K_n(t)|f(x-t) - f(x)|\,dt \leqslant \int\limits_{-\infty}^{+\infty} K_n(t)\varepsilon\,dt + \varepsilon 2B \leqslant \varepsilon + \varepsilon 2B$$

при $n \geqslant n_0$, как в свойстве DIR 3. Это доказывает, что $\{K_n * f\}$ аппроксимирует f равномерно на S и завершает доказательство теоремы. ■

Перечислим несколько частных приложений общей теоремы в различных областях математики.

Теорема Вейерштрасса об аппроксимации

Теорема. *Пусть f — непрерывная функция в конечном замкнутом вещественном интервале $[a, b]$. Тогда f может быть равномерно аппроксимирована полиномами на $[a, b]$.*

Доказательство.

Используя параллельный перенос и преобразование подобия, с точностью до замены переменной мы можем предположить без потери общности, что интервал равен $[0, 1]$. Пусть L — такая линейная функция, что $L(0) = f(0)$ и $L(1) = f(1)$. Поскольку L — полином (степени 1), достаточно доказать теорему для функции $f - L$ вместо функции f. Преимущество, которое мы при этом получаем, состоит в том, что $f - L$ обращается в нуль в конечных точках интервала. Таким образом, мы свели нашу задачу к доказательству теоремы, когда $f(0) = f(1) = 0$, что мы и будем предполагать с этого момента.

Затем мы определим *последовательность Ландау*

$$K_n(x) = \begin{cases} \dfrac{1}{c_n}(1 - x^2)^n, & \text{если } |x| \leqslant 1, \\ 0, & \text{если } |x| \geqslant 1. \end{cases}$$

Выберем константу c_n так, чтобы

$$\int_{-1}^{1} K_n(x)\,dx = 1, \quad \text{поэтому} \quad c_n = \int_{-1}^{1} (1 - x^2)^n\,dx.$$

Тогда K_n удовлетворяет свойству DIR 2. Тривиально, что $K_n(x) \geqslant 0$ при всех x. Легко доказать, что свойство DIR 3 выполнено. Вы можете проработать детали самостоятельно или посмотреть их в [La 83/97], глава XI, § 2, стр. 288–289. Поэтому $\{K_n\}$ — последовательность Дирака. По общей теореме об аппроксимации получаем, что $K_n * f$ сходится равномерно к f на интервале $[0, 1]$. Осталось доказать лишь то, что $K_n * f$ — полином. Это просто, и мы сейчас проведем вычисления. Мы имеем

$$K_n * f(x) = \int_{-1}^{1} K_n(x - t) f(t)\,dt.$$

Разложим $K_n(x-t)$ как полином по x и по t, а именно

$$K_n(x-t) = \sum a_{ij}^{(n)} x^i t^j \quad \text{с коэффициентами } a_{ij}^{(n)} \in \mathbb{R}.$$

Таким образом,

$$K_n * f(x) = \sum a_{ij}^{(n)} x^i \int_{-1}^{1} t^j f(t)\,dt = \sum_i \sum_j a_{ij}^{(n)} b_j x^i,$$

где

$$b_j = \int_{-1}^{1} t^j f(t)\,dt.$$

Поэтому $K_n * f(x)$ — это полином по переменной x, и теорема доказана. ■

Кто из вас раньше видел доказательство теоремы Вейерштрасса?

Студент. Мы видели другое доказательство на курсе анализа.

Серж Ленг. Какое доказательство?

Студент. С использованием теоремы Вейерштрасса – Стоуна.

Серж Ленг. Да, теорема Вейерштрасса – Стоуна превосходная общая теорема, весьма полезная во многих ситуациях. Но есть существенная разница между применением теоремы Вейерштрасса – Стоуна и вышеприведенным доказательством. Теорема Вейерштрасса – Стоуна не дает никакой информации об аппроксимирующей последовательности, а при использовании последовательности Ландау возникает такая последовательность, явно выраженная в терминах начальной функции f. Теорема Вейерштрасса – Стоуна, к примеру, не дает вам информации о степенях аппроксимирующих полиномов как функции от индекса n и не дает оценки на коэффициенты аппроксимирующих полиномов как функции от n и, разумеется, начальной функции f. Явная конструкция показывает вам точно и явно, как происходит аппроксимация f. Использование функций Ландау дает явную аппроксимирующую последовательность и в этом смысле лучше, чем доказательство с использованием теоремы Вейерштрасса–Стоуна, если вам нужен эффективный, конструктивный результат.

Ряды Фурье

В этом приложении мы рассмотрим периодические функции периода 2π. Пусть f — такая функция. Интегралы в последующем будут браться на интервале $[-\pi, \pi]$ вместо $(-\infty, +\infty)$ в определении последовательности Дирака. Основная теорема об аппроксимации останется той же, за исключением этого изменения. Также временами у нас будет возникать нормализующий множитель $\frac{1}{2\pi}$.

Для любого целого n примем

$$\chi_n(x) = e^{inx}.$$

Мы рассмотрим свертку χ_n и функции f (непрерывной, если вы хотите, в худшем случае кусочно непрерывной), а именно

$$(\chi_n * f)(x) = \int_{-\pi}^{\pi} \chi_n(x-t) f(t)\, dt = \int_{-\pi}^{\pi} e^{in(x-t)} f(t)\, dt =$$

$$= e^{inx} \int_{-\pi}^{\pi} e^{-int} f(t)\, dt = 2\pi c_n e^{inx},$$

где

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-int} f(t)\, dt$$

известен как n-й *коэффициент Фурье* функции f. *Ряд Фурье* от f тогда по определению

$$S_f(x) = \sum_{-\infty}^{+\infty} c_n e^{inx}.$$

Он может сходиться или расходиться, поскольку мы не сделали никаких предположений о f, кроме предположения о непрерывности. Мы определим *последовательность Дирихле* D_n

как конечную сумму

$$D_n = \frac{1}{2\pi} \sum_{k=-n}^{n} \chi_k \quad \text{или также} \quad D_n = \frac{1}{2\pi} \sum_{k=-n}^{n} e^{ikx}.$$

Тогда $S_{f,n} = D_n * f$ — n-я частичная сумма ряда Фурье, а именно

$$S_{f,n}(x) = \sum_{k=-n}^{n} c_k e^{ikx}.$$

К сожалению, $\{D_n\}$ не является последовательностью Дирака, и, как мы уже говорили, частичные суммы $S_{f,n}$ не обязательно сходятся, тем более к функции f. Однако мы собираемся построить другую последовательность тригонометрических полиномов, которая является последовательностью Дирака и сходится к функции f.

Мы определим K_n как усреднение последовательности Дирихле, т. е.

$$K_n = \frac{1}{n}(D_0 + D_1 + \ldots + D_{n-1})$$

или в терминах переменной x

$$K_n(x) = \frac{1}{n}(D_0(x) + D_1(x) + \ldots + D_{n-1}(x)).$$

Простым вычислением с использованием тригонометрических тождеств и тождеств для конечных геометрических последовательностей находим, что

$$K_n(x) = \frac{1}{2\pi n} \frac{\sin^2\left(\frac{nx}{2}\right)}{\sin^2\left(\frac{x}{2}\right)}.$$

Наличие квадратов в правой части выражения показывает, что $K_n \geqslant 0$, что является первым свойством DIR 1 для последовательности Дирака.

Чтобы проверить свойство DIR 2, мы сначала почленно проинтегрируем $D_n(x)$. Получим

$$\int_{-\pi}^{\pi} \chi_k(x)\, dx = \int_{-\pi}^{\pi} e^{ikx}\, dx = \begin{cases} 2\pi, & \text{если } k = 0, \\ 0, & \text{если } k \neq 0. \end{cases}$$

Затем сразу получаем

$$\int_{-\pi}^{\pi} D_n(x)\,dx = 1$$

после интегрирования каждого из членов и вычисления суммы. Это доказывает, что последовательность $\{D_n\}$ удовлетворяет свойству DIR 2. Но вычисляя затем сумму, определяющую K_n из $D_0, \ldots, D_{n-1}$, мы приходим к заключению, что

$$\int_{-\pi}^{\pi} K_n(x)\,dx = 1,$$

т. е. K_n удовлетворяет свойству DIR 2.

И, наконец, нам нужно проверить свойство DIR 3. При $0 < \delta \leqslant \leqslant |x| \leqslant \pi$ значения

$$\left|\frac{1}{\sin^2\left(\frac{x}{2}\right)}\right|$$

ограничены, поскольку $\frac{x}{2}$ отделено от 0. Пусть B — граничное значение, т. е.

$$\left|\frac{1}{\sin^2\left(\frac{x}{2}\right)}\right| \leqslant B.$$

При заданном ε существует n_0 такое, что при $n \geqslant n_0$ получим

$$\frac{1}{2\pi n} \int_{\delta \leqslant |x| \leqslant \pi} \left|\frac{\sin^2\left(\frac{nx}{2}\right)}{\sin^2\left(\frac{x}{2}\right)}\right| dx \leqslant \frac{1}{n} B < \varepsilon.$$

Это доказывает свойство DIR 3, и таким образом мы показали, что

Последовательность $\{K_n\}$ — последовательность Дирака (для периодических функций).

Применяя общую теорему об аппроксимации, мы получим один из основных результатов из теории рядов Фурье.

Теорема Фейера – Чезаро (Fejer – Cesaro). *Пусть f — кусочно-непрерывная периодическая функция. Пусть $\{K_n\}$ — среднее значение частичных сумм ряда Фурье функции f. Тогда $K_n(x)$ сходится равномерно к $f(x)$ при всех x из компактного множества, на котором f — непрерывна, т. е.*

$$\frac{1}{n}(S_0(x) + S_1(x) + \ldots + S_n(x)) \to f(x)$$

равномерно, как сказано выше.

Суммирование средних значений частичных сумм рядов Фурье известно как *суммирование Чезаро.*

Гармонические функции на круге

В наших следующих примерах мы воспользуемся разновидностью понятия последовательности, а именно, мы введем понятие семейства вместо последовательности. Таким образом, мы определим семейство Дирака $\{K_r\}$, где $0 \leqslant r < 1$, и пусть r стремится к 1 вместо n, стремящегося к бесконечности. Мы будем называть $\{K_n\}$ *семейством Дирака*, если оно удовлетворяет трем условиям:

DIR 1. Для всех r, $0 \leqslant r < 1$ выполнено $K_r(t) \geqslant 0$.

DIR 2. Для этих значений r выполнено

$$\int\limits_{-\pi}^{\pi} K_r(t)\, dt = 1.$$

DIR 3. При заданных ε, δ существует r_0, $0 < r_0 < 1$ такое, что для всех r, $r_0 \leqslant r < 1$ выполнено

$$\int\limits_{-\pi}^{-\delta} K_r(t)\, dt + \int\limits_{\delta}^{\pi} K_r(t)\, dt < \varepsilon.$$

В точности тем же способом, которым мы доказали первую теорему об аппроксимации для последовательностей, можно доказать соответствующую версию для семейств.

Теорема об аппроксимации для семейств Дирака. *Пусть $\{K_r\}$ — семейство Дирака, как определено выше. Пусть f — кусочно-непрерывная функция. Тогда $K_r * f$ сходится к f при $r \to 1$ равномерно на любом компактном множестве, на котором f непрерывна.*

Разумеется, нам нужны практические примеры семейств Дирака. Используя полярные координаты (r, θ), определим

$$P_r(\theta) = P(r, \theta) = \frac{1}{2\pi} \sum_{-\infty}^{+\infty} r^{|n|} e^{in\theta} \quad \text{для} \quad 0 \leqslant r < 1.$$

Нам нужно доказать, что $\{P_r\}$, называемое *семейством Пуассона*, является семейством Дирака. Во-первых, заметим, что ряд, определяющий $P_r(\theta)$ — абсолютно сходящийся в выбранной нами области $0 \leqslant r < 1$ равномерно на интервале $0 \leqslant r \leqslant r_0$, если $r_0 < 1$, поскольку ряд можно сравнить с геометрической последовательностью. Простые тригонометрические тождества показывают, что $P(r, \theta)$ выражается как

$$P(r, \theta) = \frac{1}{2\pi} \frac{1 - r^2}{1 - 2r\cos\theta + r^2}.$$

Из этого тождества мы можем проверить выполнение свойства DIR 1, поскольку и числитель, и знаменатель в вышеприведенном выражении больше либо равны 0. Действительно, знаменатель принимает наименьшее значение при $\cos\theta = 1$, в этом случае знаменатель равен $(1 - r)^2$.

Теперь покажем, что свойство DIR 2 тоже выполнено. Продифференцируем почленно ряд для $P(r, \theta)$ и воспользуемся теми же значениями, которые мы получили при изучении рядов Фурье, а именно

$$\int_{-\pi}^{\pi} e^{in\theta}\, d\theta = \begin{cases} 2\pi, & \text{если } n = 0, \\ 0, & \text{если } n \neq 0. \end{cases}$$

Из этого следует, что только интеграл слагаемого с индексом $n = 0$ дает ненулевую добавку к интегралу ряда, и мы получим, что

$$\int_{-\pi}^{\pi} P_r(\theta)\, d\theta = 1.$$

Это доказывает свойство DIR 2.

И, наконец, покажем, что $\{P_r\}$ удовлетворяет свойству DIR 3. При $|\theta| \geqslant \delta > 0$ выполнено неравенство

$$\frac{1}{1 - 2r\cos\theta + r^2} \leqslant \frac{1}{1 - 2r\cos\delta + r^2}.$$

Исследуя производную выражения $1 - 2r\cos\delta + r^2$, мы получим, что минимум знаменателя в правой части достигается при $r = \cos\delta$. Минимальное значение знаменателя равно $1 - (\cos\delta)^2$. Следовательно, дробь

$$\frac{1}{1 - 2r\cos\theta + r^2}$$

как функция от r равномерно ограничена. Следовательно,

$$\lim_{r\to 1} \frac{1 - r^2}{1 - 2r\cos\theta + r^2} = 0,$$

и выполняется равномерно при $|\theta| \geqslant \delta > 0$. Следовательно, существует r_0 такое, что при $r_0 \leqslant r < 1$ выполнено

$$\int_{\delta\leqslant|\theta|\leqslant\pi} P_r(\theta)\, d\theta < \varepsilon.$$

Это доказывает, что семейство $\{P_r\}$ удовлетворяет свойству DIR 3 и, следовательно, является семейством Дирака.

Как следствие из общей теоремы об аппроксимации мы получаем:

Теорема Пуассона об аппроксимации. *Пусть f — периодическая кусочно-непрерывная функция. Тогда*

$$P_r * f$$

сходится к f при $r \to 1$ равномерно на любом компактном множестве, где f — непрерывная.

Далее отметим, что в примере Пуассона используется дополнительная структура, а именно дифференциальное уравнение в частных производных. Начнем с нескольких общих замечаний о дифференцировании интеграла свертки

$$(g * f)(x) = \int g(x-t)f(t)\,dt,$$

где g — бесконечно дифференцируемая функция. Мы можем заниматься дифференцированием в каждом из случаев, на вещественной прямой $\mathbb{R}$ или на интервале $[-\pi, \pi]$, если имеем дело с периодическими функциями. На вещественной прямой мы должны предполагать, что рассматриваемые интегралы абсолютно сходятся. Пусть D обозначает производную, которая при вышеописанном выборе переменных равна $D = \frac{d}{dx}$. Тогда при подходящих условиях абсолютной сходимости мы можем дифференцировать под знаком интеграла и получим формулу

$$D(g * f) = (Dg) * f,$$

или в терминах переменной x

$$\frac{d}{dx}\int g(x-t)f(t)\,dt = \int \frac{d}{dx}g(x-t)f(t)\,dt.$$

Повторное дифференцирование дает для каждого целого числа m

$$D^m(g * f) = (D^m g) * f,$$

при подходящих условиях абсолютной сходимости.

Применим все это к функции $P(r, \theta)$. Обозначим через Δ *оператор Лапласа*, который в прямоугольных координатах равен

$$\Delta = \left(\frac{\partial}{\partial x}\right)^2 + \left(\frac{\partial}{\partial y}\right)^2.$$

В действительности нам нужно выражение для оператора Лапласа в полярных координатах. Кто знает это выражение?

[*Нет ответов.*]

Я всегда задаю этот вопрос на моих лекциях, и никто не знает ответа. Нет никаких отличий, происходит ли это здесь или где-нибудь еще. В действительности я тоже не знал ответа, пока мне не пришлось преподавать математический анализ. В любом случае ответ следующий

$$\Delta = \left(\frac{\partial}{\partial r}\right)^2 + \frac{1}{r}\left(\frac{\partial}{\partial r}\right)^2 + \frac{1}{r^2}\left(\frac{\partial}{\partial \theta}\right)^2.$$

Вывод этого выражения является простым упражнением, при использовании выражений $x = r\cos\theta$, $y = \sin\theta$ и цепного правила. Затем вы можете дифференцировать ряд для $P(r, \theta)$ почленно с $\dfrac{\partial}{\partial r}$ или $\dfrac{\partial}{\partial \theta}$, и повторяя эти частные производные, получите

$$(\Delta P)(r, \theta) = 0.$$

То, что почленное дифференцирование является законным, следует из факта, что ряд и его частные производные равномерно абсолютно сходятся при r в пределах $0 \leqslant r \leqslant r_0 < 1$ для всех θ.

Бесконечно дифференцируемая функция f, удовлетворяющая уравнению $\Delta f = 0$, называется *гармонической*. Выражение для свертки $P_r * f$ на интервале $[-\pi, \pi]$ мы можем дифференцировать под знаком интеграла и поэтому

$$\Delta(P * f)(r, \theta) = \Delta \int_{-\pi}^{\pi} P(r, \theta - t) f(t)\, dt = \int_{-\pi}^{\pi} (\Delta P)(r, \theta - t) f(t)\, dt = 0.$$

Следовательно, $P * f$ — гармоническая функция.

Вышеприведенные вычисления находят приложение во всей математике и других областях. Физики особенно заинтересованы в гармонических функциях и в том, что называется решением краевой задачи на окружности. Предположим, что задана непрерывная функция на окружности, которая является границей круга радиуса 1, с центром в начале координат. Поэтому функцию f можно рассматривать как функцию $f(\theta)$ от переменной θ, рассматриваемой как обычный угол. Физики хотят найти гармоническую функцию F на всем круге D, такую, что функция $F(r, \theta)$ с $0 \leqslant r < 1$ продолжается до непрерывной функции на замкнутом круге и имеет заданное значение $f(\theta)$ на границе, так, чтобы

$$\lim_{r \to 1} F(r,\, \theta) = f(\theta)$$

при каждом θ. Теперь мы можем решить эту задачу. Обозначим

$$F(r,\, \theta) = (P_r * f)(\theta) = (P * f)(r,\, \theta) \quad \text{для} \quad \theta \in \mathbb{R}.$$

По свойству DIR 3 мы знаем, что

$$\lim_{r \to 1} (P * f)(r,\, \theta) = f(\theta),$$

Поэтому F имеет непрерывное граничное значение $f(\theta)$ для каждого θ. Более того, мы видели, что $\Delta F = 0$ во внутренности диска, поэтому F — гармоническая, как требовалось.

Гармонические функции на верхней полуплоскости

Верхняя полуплоскость состоит из всех точек $(x,\, y) \in \mathbb{R}^2$ с произвольным x и $y > 0$. Мы представим семейство Дирака, расположенное в верхней полуплоскости, граница которой состоит из прямой линии $y = 0$. Поэтому мы определим *семейство Дирака* $\{K_y\}$ на $\mathbb{R}$ с $y > 0$ как и раньше, за исключением того, что $y \to 0$ вместо $r \to 1$. Приведем пример.

Обозначим

$$K_y(x) = \frac{1}{\pi} \frac{y}{x^2 + y^2} \quad \text{при} \quad y \to 0.$$

Все функции $K_y > 0$ и простая замена переменных показывает, что свойство DIR 2 выполняется. Также простая оценка показывает, что выполняется и свойство DIR 3. Оставим проверку вам на самостоятельную работу.

Общая теорема об аппроксимации для семейств Дирака говорит нам, что для каждой непрерывной ограниченной функции f на $\mathbb{R}$ свертка

$$(K_y * f)(x)$$

сходится к $f(x)$ равномерно по x в компактном множестве.

Мы получим ту же дополнительную структуру, как для семейства Пуассона. Здесь мы используем оператор Лапласа в прямоугольных координатах

$$\Delta = \left(\frac{\partial}{\partial x}\right)^2 + \left(\frac{\partial}{\partial y}\right)^2.$$

Простое дифференцирование показывает, что функция $(x, y) \mapsto K_y(x)$ удовлетворяет уравнению Лапласа, т. е.

$$\Delta K = 0.$$

Другими словами, K — гармоническая функция на верхней полуплоскости. Дифференцируя под знаком интеграла, также, как в предыдущем примере, мы приходим к заключению, что свертка $K * f$ — гармоническая при любой ограниченной непрерывной функции f. Поэтому мы доказали следующий результат:

Теорема. *Пусть f — ограниченная непрерывная функция на вещественной оси. Определим*

$$F(x, y) = (K_y * f)(x) = \frac{1}{\pi} \int\limits_{\mathbb{R}} \frac{y}{(x-t)^2 + y^2} f(t)\, dt.$$

Тогда F — гармоническая на верхней полуплоскости, т. е. $\Delta F = 0$ на верхней полуплоскости и

$$\lim_{y \to 0} F(x, y) = f(x) \quad \textit{для} \quad \forall x \in \mathbb{R}.$$

В теореме говорится, что заданная функция f является граничным значением гармонической функции F в верхней полуплоскости, таким образом, мы явно построили гармоническую функцию с заданным граничным значением.

Приятно получить независимую конструкцию в данном случае, но если вы изучили курс комплексного анализа, вам должно быть известно, что круг и верхняя полуплоскость аналитически изоморфны. Действительно, отображение

$$z \mapsto \frac{z - i}{z + i}$$

показывает аналитический изоморфизм верхней полуплоскости с кругом. Тогда семейство Пуассона в предыдущем примере соответствует семейству $K(x, y)$ при этом изоморфизме. Хорошим упражнением будет проверка этого утверждения на курсе комплексного анализа.

Тепловой источник на вещественной прямой

Мы собираемся привести один из наиболее важных примеров семейств Дирака, может быть, самый важный пример. В этом параграфе будем использовать переменные t, x, при $t > 0$, $t \mapsto 0$. Таким образом, семейство Дирака — это семейство непрерывных функций $\{K_t\}$ на $\mathbb{R}$, индексированное вещественными положительными числами t и удовлетворяющее следующим свойствам:

DIR 1. $K_t \geqslant 0$ при всех $t > 0$.

DIR 2. $\int\limits_{\mathbb{R}} K_t(x)\,dx = 1$ при всех $t > 0$.

DIR 3. При заданных ε и δ существует $t_0 > 0$ такое, что если $t < t_0$, тогда

$$\int\limits_{|x|\geqslant\delta} K_t(x)\,dx < \varepsilon.$$

Мы также будем использовать обозначение $K_t(x) = K(t, x)$. Общая теорема об аппроксимации для семейства Дирака $\{K_t\}$ верна и утверждает, что:

Пусть f — ограниченная функция на $\mathbb{R}$, кусочно-непрерывная на каждом конечном интервале. Тогда $(K_t * f)(x)$ сходится равномерно к $f(x)$ при $t \to 0$ на каждом компактном подмножестве, на котором f непрерывна, или на любом подмножестве, на котором f равномерно непрерывна.

Теорема. *Пусть при $t > 0$*

$$K_t(x) = K(t, x) = \frac{1}{\sqrt{4\pi t}} e^{-x^2/4t}\,dx.$$

Тогда $\{K_t\}$ — семейство Дирака.

Проверка аксиом не представляет проблемы. Для свойства DIR 1 очевидно, что $K(t, x) > 0$ при всех t, x. Для свойства DIR 2 нам нужно доказать, что

$$\int\limits_{-\infty}^{\infty} K(t, x)\, dx = \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi t}} e^{-x^2/4t}\, dx = 1.$$

Сделаем замену переменных $y^2 = \frac{x^2}{4t}$, поэтому $y = \frac{x}{2\sqrt{t}}$, $dy = \frac{dx}{2\sqrt{t}}$. Затем нам нужно использовать стандартное значение из интегрального исчисления, а именно

$$\int\limits_{-\infty}^{\infty} e^{-y^2}\, dy = \sqrt{\pi},$$

и получаем свойство DIR 2.

Свойство DIR 3 также несложно доказать, нам лишь нужно провести некоторые оценки. Заданы ε, δ, нам нужно доказать, что при t, достаточно близком к нулю:

$$\frac{1}{2\sqrt{t}} \int\limits_{\delta}^{\infty} e^{-x^2/4t}\, dx < \varepsilon. \qquad (*)$$

Делаем замену переменных $x = 2\sqrt{t}y$, как и выше. Тогда оцениваемый интеграл $(*)$ становится

$$\int\limits_{\delta/(2\sqrt{t})}^{\infty} e^{-y^2}\, dy.$$

При $t \to 0$ нижний предел интегрирования $\frac{\delta}{2\sqrt{t}}$ стремится к бесконечности и, поскольку,

$$\int\limits_{-\infty}^{\infty} e^{-y^2}\, dy$$

конечен, то

$$\lim_{A\to\infty}\int_A^\infty e^{-y^2}\,dy = 0,$$

что и доказывает свойство DIR 3.

Функция $(t, x) \mapsto K(t, x)$ называется *тепловым источником* на $\mathbb{R}$. Терминология пришла из физики, и можно предположить, что она имеет какое-то отношение к теплопроводности (что возможно), но с нашей точки зрения $K(t, x)$ — просто важный пример семейств Дирака, который возникает повсюду, имеет ли это отношение к теплопроводности или нет.

Также существует структура для дифференциального уравнения. Определим дифференциальный оператор $\mathbf{H}$ на $\mathbb{R}^+ \times \mathbb{R}$:

$$\mathbf{H}_{t,x} = -\left(\frac{\partial}{\partial x}\right)^2 + \frac{\partial}{\partial t}.$$

Мы будем называть $\mathbf{H}$ *оператором теплопроводности.*

Теорема. *Тепловой источник удовлетворяет уравнению теплопроводности, т. е.*

$$\mathbf{H}K = 0,$$

или в терминах переменных

$$\mathbf{H}_{t,x}K(t, x) = 0.$$

Чтобы увидеть это, оставайтесь хладнокровными, спокойными и собранными. Продифференцируйте дважды по x и один раз по t. Вы получите то же самое значение, если вы не сделаете ошибок в вычислениях, которые я часто делаю. После вычитания вы получите 0.

Дифференцирование под знаком интеграла приводит к:

Следствие. *Пусть f — ограниченная непрерывная функция на $\mathbb{R}$. Пусть*

$$F(t, x) = (K_t * f)(x).$$

Тогда $\mathbf{H}F = 0$, т. е. F удовлетворяет уравнению теплопроводности.

Это лишь частный случай общей возможности дифференцирования свертки под знаком интеграла, т.е.

$$D(g * f) = (Dg) * f$$

при условии абсолютной сходимости. Эти условия здесь выполнены.

Таким образом, мы видим, что нам удалось решить краевую задачу на верхней полуплоскости, с данным краевым значением f для оператора теплопроводности, а не для оператора Лапласа, с которым мы встречались раньше. Утверждается, что K является ***фундаментальным решением*** уравнения теплопроводности, поскольку свертка с K позволяет получить решение уравнения теплопроводности с заданными краевыми условиями.

Работать на $\mathbb{R}^n$ не сложнее, чем на $\mathbb{R}$. Используя n-мерную версию:

$$K_t^{\mathbb{R}^n}(x) = K^{\mathbb{R}^n}(t, x) = \frac{1}{(4\pi t)^{n/2}} e^{-x^2/4t},$$

где $x = (x_1, \ldots, x_n)$ — это n-ка из вещественных чисел, и

$$x^2 = x \cdot x = x_1^2 + \ldots + x_n^2$$

обычное скалярное произведение x с самим собой. Проверка того, что выполняются три условия DIR 1, DIR 2 и DIR 3, является технической операцией, в точности, как в одномерном случае. Не запутайтесь, вычисляя частные производные.

Оператор теплопроводности на $\mathbb{R}^n$ тогда будет

$$\mathbf{H}_{t,x} = -\Delta_x + \frac{\partial}{\partial t},$$

где

$$\Delta_x = \left(\frac{\partial}{\partial x_1}\right)^2 + \ldots + \left(\frac{\partial}{\partial x_n}\right)^2$$

оператор Лапласа. Они применяются к функции

$$F(t, x) = F(t, x_1, \ldots, x_n).$$

Тепловой источник на окружности

Также, как мы рассматривали периодические функции при обсуждении рядов Фурье, рассмотрим периодические функции в связи с тепловым источником. Таким образом, мы хотим найти семейство Дирака для периодических функций, удовлетворяющих уравнению теплопроводности. Оператор теплопроводности тот же самый, что и на $\mathbb{R}$. Нам нужно семейство $\{K_t\}$ периодических функций

$$K(t, x) = K_t(x), \quad \text{удовлетворяющих} \quad K(t, x + 2\pi) = K(t, x)$$

и следующим трем условиям:

DIR 1. $K_t \geqslant 0$ при всех $t > 0$.

DIR 2. $\int\limits_0^{2\pi} K_t(x)\,dx = 1$ при всех $t > 0$.

DIR 3. При заданных ε, δ между 0 и 1, существует t_0, $0 < t_0 < 1$, такое, что при $0 < t < t_0$ выполнено

$$\int\limits_{-\pi}^{-\delta} K_t(x)\,dx + \int\limits_{\delta}^{\pi} K_t(x)\,dx < \varepsilon.$$

Дополнительно, если мы обозначим

$$\mathbf{H} = -\left(\frac{\partial}{\partial x}\right)^2 + \frac{\partial}{\partial t},$$

тогда мы также хотим, чтобы K удовлетворяло уравнению в частных производных:

$$\mathbf{H}K = 0.$$

Чтобы отличать такие периодические функции K от непериодических функций для тепловых источников на $\mathbb{R}$, мы будем записывать $K^{\mathbb{R}}$ для тепловых источников на $\mathbb{R}$, т. е.

$$K^{\mathbb{R}}(x) = \frac{1}{\sqrt{4\pi t}} e^{-x^2/4t}.$$

Естественный подход к построению периодического решения в нашей задаче — это периодизация непериодического решения на $\mathbb{R}$. Поэтому мы определим

$$K^S(t, x) = \sum_{-\infty}^{\infty} K^{\mathbb{R}}(t, x + 2\pi n) = \sum_{n\in\mathbb{Z}} K^{\mathbb{R}}(t, x + 2\pi n). \qquad (1)$$

Верхний индекс S обозначает окружность. Стандартное обозначение целых чисел $\mathbb{Z}$ и сумма берется по всем целым числам $n \in \mathbb{Z}$.

При каждом значении $t > 0$, члены ряда (1) экспоненциально убывают, поэтому ряд сходится очень быстро. Оставим вам выписать все детали.

Тогда следует, что $K^S(t, x)$ — периодическая, т. е.

$$K^S(t, x + 2\pi) = K^S(t, x)$$

при всех $t > 0$ и $x \in \mathbb{R}$.

Теорема. *Функция K^S определяет семейство Дирака периодических функций и удовлетворяет уравнению теплопроводности $\mathbf{H}K^S = 0$.*

Доказательство.

Условие положительности DIR 1, очевидно, выполняется даже в обычной строгой форме $K^S(t, x) > 0$ при всех t, x (строгая положительность).

Для проверки свойства DIR 2 мы сводим интеграл над $[0, 2\pi]$ к интегралу над $\mathbb{R}$. Мы получаем

$$\int_0^{2\pi} K^S(t, x)\, dx = \int_0^{2\pi} \frac{1}{\sqrt{4\pi t}} \sum_{n\in\mathbb{Z}} e^{-\frac{(x+2\pi n)^2}{4t}}\, dx =$$

$$= \frac{1}{\sqrt{4\pi t}} \sum_{n\in\mathbb{Z}} \int_0^{2\pi} e^{-\frac{(x+2\pi n)^2}{4t}}\, dx = \frac{1}{\sqrt{4\pi t}} \sum_{n\in\mathbb{Z}} \int_{2\pi n}^{2\pi(n+1)} e^{-y^2/4t} dy =$$

$$= \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-y^2/4t} dt = \int_{-\infty}^{\infty} K_t^{\mathbb{R}}(y)\, dy = 1.$$

Таким образом, мы свели свойство DIR 2 на окружности к свойству DIR 2 для теплового источника $K^{\mathbb{R}}$ на прямой, интеграл от которого, как мы уже видели, равен 1.

Осталось доказать свойство DIR 3. Мы снова сводим это свойство к соответствующему свойству $K^{\mathbb{R}}$ на прямой. Мы должны показать, что интеграл от K_t^S на интервале $[\delta, \pi]$ стремится к 0, когда $t \to 0$. Чтобы сделать это, мы снова поменяем местами сумму и интеграл, а именно

$$\frac{1}{\sqrt{4\pi t}} \sum_{n=0}^{\infty} e^{-\frac{(x+2\pi n)^2}{4t}} dx = \\ = \sum_{n=0}^{\infty} \frac{1}{\sqrt{4\pi t}} \int\limits_{\delta+2\pi n}^{2\pi(n+1)} e^{-y^2/4t} dy \leqslant \int\limits_{\delta}^{\infty} K^{\mathbb{R}}(t, x)\, dy.$$

По свойству DIR 3 для теплового источника на $\mathbb{R}$, мы знаем, что существует t_0, $0 < t_0 < 1$, такое, что это последнее выражение при $0 < t < t_0$ оказывается меньше ε. Аналогичный аргумент доказывает неравенство для интервала $[-\pi, -\delta]$. Таким образом, мы доказали свойство DIR 3 для K^S.

И, наконец, рассмотрим дифференциальное уравнение. Сходимость периодического ряда

$$\sum_{-\infty}^{\infty} e^{-\frac{(x+2\pi n)^2}{4t}}$$

является достаточно быстрой для того, чтобы мы могли почленно продифференцировать и в случае с $\dfrac{\partial}{\partial t}$ и в случае с $\dfrac{\partial}{\partial x}$ и $\left(\dfrac{\partial}{\partial x}\right)^2$. Но если c — произвольная константа и f удовлетворяет уравнению теплопроводности, т. е.

$$\mathbf{H}_{t,x} f(t, x) = 0 \quad \text{для всех} \quad t, x,$$

тогда функция g, определяемая как

$$g(t, x) = f(t, x + c)$$

также удовлетворяет уравнению. Следовательно, каждое слагаемое

$$f_n(t, x) = \frac{1}{\sqrt{4\pi t}} e^{-\frac{(x+2\pi n)^2}{4t}}$$

удовлетворяет уравнению теплопроводности $\mathbf{H}f_n = 0$. Следовательно, $\mathbf{H}K^S = 0$, как и требовалось показать.

Затем мы установим соотношение между сверткой на окружности (т. е. для периодических функций) и сверткой на вещественной прямой. На окружности свертка задается формулой

$$f * g(x) = \int_0^{2\pi} f(y)g(x-y)\,dy,$$

для периодических функций f, g с периодом 2π. Пусть f — периодическая. Тогда

$$\begin{aligned}
K_t^S * f(x) &= \int_0^{2\pi} K_t^S(y)f(x-y)\,dy = \\
&= \int_0^{2\pi} \sum_{n\in\mathbb{Z}} \frac{1}{\sqrt{4\pi t}} e^{-\frac{(y+2\pi n)^2}{4t}} f(x-y)\,dy = \\
&= \frac{1}{\sqrt{4\pi t}} \sum_{n\in\mathbb{Z}} \int_0^{2\pi} e^{-\frac{(y+2\pi n)^2}{4t}} f(x-y)\,dy = \\
&= \frac{1}{\sqrt{4\pi t}} \sum_{n\in\mathbb{Z}} \int_{2\pi n}^{2\pi(n+1)} e^{-y^2/4t} f(x-y)\,dy = \\
&\text{(из-за периодичности) } f \\
&= \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-y^2/4t} f(x-y)\,dy.
\end{aligned}$$

Таким образом, мы получили соотношение между сверткой на окружности и сверткой на $\mathbb{R}$. Мы можем подытожить вычисления в теореме. ■

Теорема. *Пусть f — непрерывная периодическая функция периода 2π. Тогда*

$$K_t^S * f(x) = K_t * f(x),$$

где в левой части свертка на $[0, 2\pi]$, *а в правой — на* $(-\infty, \infty)$.

Заметьте, что формула

$$\mathbf{H}(K^S * f) = (\mathbf{H}K^S) * f = 0$$

также может быть получена из этой последней теоремы.

Ряды Фурье и инверсия Пуассона

Так как K_t^S — периодическая, у нее есть ряд Фурье, который может быть легко вычислен. В действительности, K_t^S бесконечно дифференцируема, как можно проверить, дифференцируя ряд почленно и показывая, что продифференцированный ряд сходится абсолютно и равномерно на каждом ограниченном интервале $|x| \leqslant A$. По общим теоремам о рядах Фурье, ряд Фурье от K_t^S сходится к функции K_t^S. Следующая теорема показывает нам, какой этот ряд Фурье.

Теорема. *Ряд Фурье для* K_t^S *задается выражением*

$$\frac{1}{\sqrt{4\pi t}} \sum_{n\in\mathbb{Z}} e^{-\frac{(x+2\pi n)^2}{4t}} = \frac{1}{2\pi} \sum_{n\in\mathbb{Z}} e^{-n^2 t} e^{inx}.$$

Доказательство.

m-й коэффициент Фурье определяется интегралом для $m \in \mathbb{Z}$

$$\int_0^{2\pi} K_t^S(x) e^{-imx} dx = \sum_{n\in\mathbb{Z}} \frac{1}{\sqrt{4\pi t}} \int_0^{2\pi} e^{-\frac{(x+2\pi n)^2}{4t}} e^{-imx} dx =$$

$$= \sum_{n\in\mathbb{Z}} \frac{1}{\sqrt{4\pi t}} \int_{2\pi n}^{2\pi(n+1)} e^{-y^2/4t} e^{-imx} dy,$$

где мы сделали подстановку $y = x + 2\pi n$ и воспользовались $e^{im2\pi n} = 1$. Последнее выражение является равным

$$\frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-y^2/4t} e^{-imy} dy.$$

Мы сделали подстановку $y = \sqrt{2t}u$, $dy = \sqrt{2t}du$ и использовали стандартное значение из математического анализа, а именно

$$\frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} e^{-u^2/2} e^{-iuv} du = e^{-v^2/2}.$$

Затем получаем

$$\int\limits_{0}^{2\pi} K_t^S(x) e^{-imx} dx = e^{-m^2 t}.$$

Это и доказывает теорему. ■

Эта теорема получена Пуассоном и называется *формулой инверсии Пуассона.* Очень важный частный случай возникает при выборе $x = 0$, тогда e^{inx} становится единицей. В этом случае мы получим то, что известно как *формула суммирования Пуассона.*

Следствие.

$$\frac{1}{\sqrt{4\pi t}} \sum_{n \in \mathbb{Z}} e^{-\frac{(2\pi n)^2}{4t}} = \frac{1}{2\pi} \sum_{n \in \mathbb{Z}} e^{-n^2 t}.$$

Существует много приложений для этой формулы в анализе и теории чисел. Мы собираемся привести одно из этих приложений ниже, показав, как Риман доказал функциональное уравнение ζ-функции. В анализе формула инверсии Пуассона возникает в спектральной теории. Для очень общего обсуждения формул инверсии см. [JoL 94].

Ряд Фурье теплового источника имеет форму, которая интересна сама по себе и называется θ-рядом. Мы приведем характерные свойства таких θ-рядов в отдельном параграфе.

θ-ряды и свертка

Начнем с ряда Фурье теплового источника на окружности. Не зная ничего, мы можем определить *θ-функцию* $\theta_t(x)$

(при $x \in \mathbb{R}$ и $t > 0$) с помощью ряда

$$\theta_t(x) = \frac{1}{2\pi} \sum_{n \in \mathbb{Z}} e^{-n^2 t} e^{inx}.$$

Разумеется, периодический тепловой источник K_t^S равен θ-ряду, но нам не нужно знать этого для того, что будет получено дальше. Мы просто начнем с вышеприведенного определения, поэтому аргументы, доказывающие следующую теорему самодостаточны.

Теорема. *Для положительных вещественных чисел t, s выполнено соотношение*

$$\theta_t * \theta_s = \theta_{t+s}.$$

Доказательство.

Выше мы записали ряд для $\theta_t(x)$, а теперь запишем ряд с s вместо t,

$$\theta_s(x) = \frac{1}{2\pi} \sum_{m \in \mathbb{Z}} e^{-m^2 s} e^{imx}.$$

Из-за быстрого убывания $e^{-n^2 t}$ и $e^{-m^2 s}$ свертка задается выражением

$$\theta_t * \theta_s(x) = \int_0^{2\pi} \theta_t(x-y)\theta_s(y)dy =$$

$$= \frac{1}{(2\pi)^2} \sum_{m,n} \int_0^{2\pi} e^{-n^2 t} e^{-m^2 s} e^{in(x-y)} e^{imy} dy =$$

$$= \frac{1}{(2\pi)^2} \sum_{m,n} e^{inx} e^{-(n^2 t + m^2 s)} \int_0^{2\pi} e^{i(m-n)y} dy.$$

Обычные аргументы показывают, что последний интеграл равен 0, если $m \neq n$, и равен 2π, если $m = n$. Следовательно, только члены с $m = n$ остаются в сумме, и мы получаем

$$\theta_t * \theta_s(x) = \frac{1}{(2\pi)^2} \sum_{n \in Z} e^{-n^2(t+s)} \cdot 2\pi = \theta_{t+s}(x),$$

что и доказывает теорему. ■

Замечание. Разумеется, поскольку θ-функция равна тепловому источнику, та же самая формула применима и к тепловому источнику. Другими словами,

$$K_t^S * K_{t'}^S = K_{t+t'}^S.$$

Оператор свертки является ассоциативным, т. е. для трех функций h, g, f мы имеем

$$(h * g) * f = h * (g * f).$$

Вы можете проверить ассоциативность заменой порядка интегрирования. Для каждого t функция K_t^S может рассматриваться как интегральный оператор

$$f \mapsto K_t^S * f.$$

Тогда вышеприведенная формула свертки $K_t^S * K_{t'}^S = K_{t+t'}^S$ говорит вам, что оператор составленный из K_t^S и $K_{t'}^S$ задается сверткой с $K_{t+t'}^S$. Поэтому говорится, что множество тепловых источников $\{K_t^S\}$ является полугруппой операторов, поскольку оно замкнуто по композиции.

Таким образом, мы получили все основные свойства теплового источника на вещественной прямой. Если вы будете изучать математику, то в дальнейшем вы осознаете, что эти свойства являются типичными в очень общих ситуациях. То, что мы делали, это не случайно, это прототип для больших и лучших теорий.

Дальше мы оставим изучение теплового источника и переключимся на, казалось бы, совершенно отличные аспекты θ-функций.

Формула суммирования Пуассона и функциональное уравнение ζ-функции

Одно из приложений формулы суммирования Пуассона — это риманово доказательство функционального уравнения ζ-функции, которое мы сейчас приведем. Сначала мы введем функцию

$$\psi(t) = \sum_{n \in \mathbb{Z}} e^{-n^2 \pi t},$$

которая с точностью до замены переменных и постоянного множителя является θ-функцией. Более точно

$$\psi(t) = 2\pi\theta(\pi t, 0).$$

Тогда формула суммирования Пуассона может быть записана в виде

$$\psi(t^{-1}) = \sqrt{t}\,\psi(t). \tag{1}$$

Затем ζ-*функция* ζ определяется рядом

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

Этот ряд сходится абсолютно при всех вещественных $s > 1$, как известно из курса математического анализа и с помощью интегрального теста. Вы также можете взять комплексное s с $\mathrm{Re}(s) > 1$, если вы знаете о комплексных числах. Также сходимость является равномерной при

$$\mathrm{Re}(s) \geqslant c > 1.$$

Но мы можем рассматривать только вещественные s, если вам этого хочется. Задача заключается в том, чтобы получить выражение для $\zeta(s)$, которое бы сходилось при всех значениях s с единственно возможными исключениями в $s = 0$ и $s = 1$. Это то, что сделал Риман, и мы будем следовать рассуждениям Римана. В доказательстве используется так называемое *преобразование Меллина* $\mathbf{M}g$ подходящей функции g, определяемое как

$$(\mathbf{M}g)(s) = \int_0^{\infty} g(t)t^s \,\frac{dt}{t}.$$

Также при $s > 1$ введем функцию

$$F(s) = \pi^{-s/2}\Gamma\left(\frac{s}{2}\right)\zeta(s),$$

где *гамма-функция* определяется обычным интегралом

$$\Gamma(s) = \int_0^{\infty} e^{-t}t^s \,\frac{dt}{t}.$$

Теорема Римана. *Функция F выражается аналитически при всех $s \neq 0,\ 1$ и удовлетворяет уравнению*

$$F(s) = F(1-s).$$

Доказательство.

Определим функцию

$$g(t) = \sum_{n=1}^{\infty} e^{-n^2\pi t} = \frac{1}{2}(\psi(t) - 1) \tag{2}$$

или также

$$2g(t) = \psi(t) - 1.$$

Тогда

$$(\mathbf{M}g)\left(\frac{s}{2}\right) = \int\limits_0^{\infty} g(t) t^{s/2} \frac{dt}{t}.$$

Лемма. *Выполнено равенство*

$$(\mathbf{M}g)\left(\frac{s}{2}\right) = F(s) \quad \text{при} \quad s > 1.$$

Доказательство.

Подставим ряд для $g(t)$ и поменяем порядок суммирования и интегрирования. Получим

$$(\mathbf{M}g)\left(\frac{s}{2}\right) = \sum_{n=1}^{\infty} \int\limits_0^{\infty} e^{-n^2\pi t} t^{s/2} \frac{dt}{t}.$$

Сделаем замену переменных $t \mapsto \dfrac{t}{n^2\pi}$, т. е. $u = n^2\pi t,\ du = n^2\pi\, dt$. Тогда

$$(\mathbf{M}g)\left(\frac{s}{2}\right) = \sum_{n=1}^{\infty} \frac{1}{n^s \pi^{s/2}} \int\limits_0^{\infty} e^{-u} u^{s/2} \frac{du}{u} = \sum_{n=1}^{\infty} \pi^{-s/2} \frac{1}{n^s} \Gamma\left(\frac{s}{2}\right) = F(s),$$

что доказывает лемму. ■

Затем разложим интеграл Меллина $(\mathbf{M}g)\left(\frac{s}{2}\right)$ на два интеграла

$$(\mathbf{M}g)\left(\frac{s}{2}\right) = \int_1^\infty g(t)t^{s/2}\frac{dt}{t} + \int_0^1 g(t)t^{s/2}\frac{dt}{t} =$$
$$= \int_1^\infty g(t)t^{s/2}\frac{dt}{t} + \int_0^\infty g\left(\frac{1}{t}\right)t^{-s/2}\frac{dt}{t},$$

сделав преобразование $t \mapsto \dfrac{1}{t}$. Используя (2), получим

$$g\left(\frac{1}{t}\right) = \frac{1}{2}\left(\psi(\frac{1}{t}) - 1\right) = \frac{1}{2}\sqrt{t}\psi(t) - \frac{1}{2} = \sqrt{t}g(t) + \frac{1}{2}\sqrt{t} - \frac{1}{2}.$$

Подставляя это значение в интеграл, содержащий $g\left(\frac{1}{t}\right)$, и принимая во внимание, что $F(s) = (\mathbf{M}g)\left(\frac{s}{2}\right)$, получим

$$F(s) = \frac{1}{s-1} - \frac{1}{s} + \int_1^\infty g(t)(t^{s/2} + t^{(1-s)/2})\frac{dt}{t}.$$

Первые два слагаемых получаются прямым интегрированием и применением математического анализа. Теперь заметим, что интегральное слагаемое верно для всех значений s, а первые два слагаемых имеют особенности при $s = 1$ и $s = 0$, но в других случаях являются прекрасными простыми функциями. Более того, при замене s на $1-s$ и обратно, выражение

$$\frac{1}{s-1} - \frac{1}{s}$$

остается неизменным и также остается неизменным выражение под знаком интеграла. Это завершает доказательство функционального уравнения. ■

Постскриптум. Кстати, если вы сейчас вернетесь к подсчетам простых чисел и гипотезе Римана, вы увидите ту же ζ-функцию, которая использовалась в этом параграфе. Продолжив анализ немного дальше, Риман получил явную формулу для $\pi(x)$ в терминах нулей этой ζ-функции. Эквивалентность между гипотезой Римана, как мы ее сформулировали в терминах подсчетов, и тем, как Риман ее сформулировал, является прямым следствием явной формулы. Но вывод явной формулы находится вне уровня этих бесед.

θ-функции и комплексные двоякопериодические функции

Просто для того, чтобы показать, как разделы математики, которые на первый взгляд кажутся различными, на самом деле связаны, я собираюсь сейчас описать, каким образом θ-функции возникают в разложении определенных типов комплексно-аналитических функций, называемых эллиптическими функциями. Это разложение аналогично разложению полинома на линейные множители над комплексными числами. На более глубоком уровне это аналогично разложению положительного целого числа на простые числа. Но проявление этих аналогий требует все больше и больше пространства и времени, и нам придется где-то остановиться, поэтому мы остановимся после этого обсуждения.

Каждый знает периодические функции синус и косинус, которые связаны с геометрией окружности. Вне этой геометрии, как вы знаете из курса дифференциальных уравнений, они удовлетворяют дифференциальному уравнению

$$y'^2 = 1 - y^2, \tag{1}$$

которое является лишь другим способом записи выражений

$$\sin'(x) = \cos x \quad \text{и} \quad \sin^2(x) + \cos^2(x) = 1.$$

Более того $\sin x$ и $\cos x$ — периодические с одним основным периодом 2π. В восемнадцатом и девятнадцатом столетии математики натолкнулись на похожий феномен, но несколько более сложный, а именно, они рассматривали функции f, которые удовлетворяют дифференциальному уравнению

$$f'^2 = f^3 + Af + B. \tag{2}$$

Здесь A и B — константы. В дифференциальном уравнении для синуса и косинуса полином в правой части имеет степень 2. В похожей, но более сложной ситуации (2) полином имеет степень 3. Полиномиальное дифференциальное уравнение в действительности является довольно общим, поскольку заменой переменных полином степени 3 всегда может быть преобразован в другой, в котором нет квадратичного члена, т. е. к виду

$$P(T) = T^3 + AT + B,$$

где коэффициент при T^2 равен нулю. Поэтому уравнение (2) является прямым обобщением (1) на полиномы третьей степени в правой части. К этому дифференциальному уравнению пришли в теории комплексных, двоякопериодических функций, т. е. функций $f(z)$ комплексной переменной, которые имеют два периода ω_1, ω_2, которые линейно независимы на $\mathbb{R}$. Например, $\omega_1 = 1$ и $\omega_2 = i$ будет такой парой периодов. Мы хотим изучить аналитические функции f, такие, что

$$f(z + \omega_1) = f(z) \quad \text{и} \quad f(z + \omega_2) = f(z)$$

для всех комплексных z. На самом деле, такие функции должны иметь полюса, если они не являются константами, поэтому мы в действительности работаем с тем, что называют мероморфными функциями, обладающими двумя независимыми периодами. Такие функции называются *эллиптическими*. Примечательно, что изучение эллиптических функций приводит к θ-функциям, как мы опишем ниже. Эти θ-функции для определенных значений переменных удовлетворяют уравнению теплопроводности и, как мы видели раньше, являются фундаментальными решениями уравнения теплопроводности на окружности.

Для приложений к комплексной ситуации определим сначала *θ-функцию Римана*

$$\theta_1(\tau, z) = \sum_{n\in\mathbb{Z}} e^{\pi i n^2 \tau} e^{2\pi i n z}.$$

В этом определении τ, z комплексные числа. Мы запишем $\tau = u + it$, u, t — вещественные, поэтому $t = \mathrm{Im}(\tau)$ — мнимая часть τ. Мы предполагаем $t > 0$, чтобы ряд сходился. При использовании этого предположения видно, что

$$e^{\pi i n^2 \tau} = e^{-\pi n^2 t} e^{\pi i n^2 u}.$$

Т. к. $e^{\pi i n^2 u}$ по абсолютному значению равно 1, абсолютное значение $e^{\pi i n^2 \tau}$ просто равно $e^{-\pi n^2 t}$. Следовательно, ряд для $\theta_1(\tau, z)$ абсолютно сходится, и равномерно при $|t| \geqslant \delta > 0$. Для вещественного $z = x$ и чисто мнимого $\tau = it$ мы в сущности получим θ-ряд, который рассматривался раньше в связи с тепловым источником на окружности.

θ-функция Римана θ_1 обладает простым почти периодическим соотношением, которое мы приведем явно перед тем, как укажем ее приложение к эллиптическим функциям. Прежде всего заметим, что 1 является истинным периодом, т. е.

PER 1. $\theta_1(\tau, z+1) = \theta_1(\tau, z)$.

Это получается немедленно, поскольку

$$\theta_1(\tau, z+1) = \sum_{n \in \mathbb{Z}} e^{\pi i n^2 \tau} \cdot e^{2\pi i n(z+1)}$$

и

$$e^{2\pi i n(z+1)} = e^{2\pi i n z} \cdot e^{2\pi i n} = e^{2\pi i n z},$$

поскольку $e^{2\pi i n} = 1$ при всех $n \in \mathbb{Z}$. Далее, τ не вполне период, но все же имеем

PER 2. $\theta_1(\tau, z+\tau) = e^{-\pi i(\tau + 2z)} \cdot \theta_1(\tau, z)$.

Вычисления показаны ниже

$$\begin{aligned}
\theta_1(\tau, z+\tau) &= \sum_{n \in \mathbb{Z}} e^{\pi i n^2 \tau} \cdot e^{\pi i n z} \cdot e^{\pi i n \tau} = \\
&= \sum_{n \in \mathbb{Z}} e^{\pi i (n^2 + 2n + 1)\tau} \cdot e^{-\pi i \tau} \cdot e^{\pi i n z} = \\
&= e^{-\pi i(\tau + 2z)} \sum_{n \in \mathbb{Z}} e^{\pi i (n+1)^2 \tau} \cdot e^{2\pi i (n+1) z} = \\
&= e^{-\pi i(\tau + 2z)} \cdot \theta_1(\tau, z),
\end{aligned}$$

таким образом доказывая соотношение PER 2.

Эти два соотношения PER 1 и PER 2 позволяют нам построить двоякопериодические функции. Все, что нам нужно сделать — это придумать что-то, чтобы избавиться от лишнего множителя, возникающего в соотношении PER 2. Пусть $a_1, \ldots, a_n$ и $b_1, \ldots, b_n$ — комплексные числа, обладающие свойством

$$a_1 + \ldots + a_n = b_1 + \ldots + b_n.$$

Пусть

$$f(z) = \frac{\prod_j \theta_1(\tau, z - a_j)}{\prod_j \theta_1(\tau, z - b_j)}.$$

Тогда из соотношения PER 1 видно, что $f(z+1) = f(z)$, и из-за нашего предположения на числа $a_1, \ldots, a_n$, $b_1, \ldots, b_n$ из соотношения PER2 получаем, что $f(z + \tau) = f(z)$. Таким образом, мы построили множество двоякопериодических функций с периодами $1, \tau$.

Можно показать, что все двоякопериодические мероморфные функции могут быть построены вышеописанным способом с использованием θ-функций несколько более общих, чем те, которые мы описали выше, но с использованием в точности такой же процедуры. Для этого смотрите книги по эллиптическим функциям, например [La 87].

Чтобы узнать о других применениях θ-функций, смотрите обзор Гесцеси и Вейкарда [GeW 98], который вышел во время работы над этой книгой.

Литература

[GeW 98] F. Gesztesy, R. Weikard, Elliptic algebro geometric solutions of the KDV and AKNS hierarchies — an analytic approach, *Bull. AMS* **35**, № 4 (1998) pp. 271–317.

[JoL 94] J. Jorgenson, S. Lang. *Explicit formulas for regularized series and products,* Springer lecture Notes 1593, 1994.

[La 83/97] S. Lang, *Undergraduate Analysis,* Springer-Verlag, 1983; Second edition, 1997.

[La 87] S. Lang, *Elliptic Functions,* Springer-Verlag, 1987.

[La 99] S. Lang, *Complex Analysis,* Fourth edition, Springer-Verlag, 1999.

Пространства Брюа–Титса

Мы будем использовать понятия и вспомогательные теоремы, которые являются стандартными в основных курсах по линейной алгебре и анализу. Однако, такие курсы по своей структуре предназначены для того, чтобы как можно шире представить основы этих дисциплин, и в них не предусмотрено время на углубленное изучение в конкретных направлениях. Представленный материал дает возможность студентам вузов продвинуться в этих направлениях, собрав вместе понятия, которые разделены по курсам, такие, как понятие производной, как линейного отображения, ряды экспонент, применяемые к линейным отображениям, положительно определенные матрицы, длина кривых, и т. д. Один из способов использования представленного материала — попросить заинтересованных студентов самих прочитать его и представить другим на специальном студенческом коллоквиуме, семинаре или математическом клубе. Я сделал это в Йейльском университете с очень большим успехом.

Первая часть, посвященная закону полупараллелограмма, может быть темой одной беседы. Она может быть проведена в начале курса математического анализа сразу после определения полного метрического пространства. Для студентов, которые знают, что такое группа, или желают изучить это прямо сейчас, я рекомендую теорему Брюа–Титса о неподвижной точке. Здесь даже, может быть, что-то из экономики. Широко известно, что некоторые теоремы о неподвижной точке часто используются не только в математике, но также и в экономике. Должны быть какие-нибудь приложения теоремы Брюа–Титса о неподвижной точке в экономике. Весьма вероятно, что если вы их обнаружите, то получите Нобелевскую премию.

Вторая и третья часть требуют несколько более глубоких знаний из курса анализа и линейной алгебры. Это дает практическую иллюстрацию тому, что мы говорим студентам, что производная дифференцируемого отображения в точке — это линейное отображение, но большинство математиков следуют этому, по-прежнему вычисляя частные производные вместо использования линейного отображения как такового. Вторая часть определяет метрику на пространстве вещественных положительно определенных матриц. Третья часть доказывает, что это пространство удовлетворяет закону полупараллелограмма. Обе эти части дают практику по математическому анализу в векторных пространствах, смешанному с некоторыми знаниями из линейной алгебры, включая скалярные произведения. Нет ни времени, ни возможности предложить этот материал в стандартных курсах, но эта тема является наиболее классическим примером некоторых аспектов анализа и алгебры, а также геометрии, показывая, как экспоненциальное отображение «искривляет» обычное евклидово пространство и влияет на расстояния — в действительности увеличивает расстояния. Поэтому вторая и третья часть дают примеры дифференциальной геометрии, довольно отличные от обычных примеров, выделяющих кривые и поверхности. Показывается использование математического анализа в векторных пространствах. Эти части дают прекрасный пример объединения математического анализа, линейной алгебры, анализа в векторных пространствах и геометрии на уровне, доступном студенту. Все три части дают возможность провести ряд бесед.

В конце главы я опишу некоторые исторические события, лежащие за рассматриваемой нами математикой. Оказывается, что совершенно элементарные, красивые и сильные результаты могут быть извлечены из весьма причудливого контекста, что я и сделал. Даже история этого математического контекста не является элементарной. Такова жизнь, но мы не можем позволить древней истории прятать или ограждать математику, доступную на элементарном уровне.

Закон полупараллелограмма

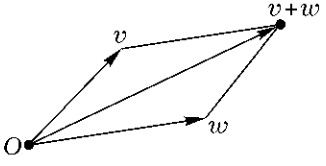
Обычный закон параллелограмма впервые был получен в контексте евклидовой плоскости, но является верным в векторном пространстве над вещественными числами с положительно определенным скалярным произведением. Из такого произведения мы получим норму на векторном пространстве обычным способом, а именно, для вектора v норма определяется как

$$|v| = \sqrt{v \cdot v}.$$

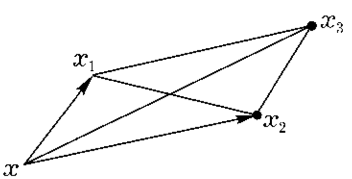
Закон параллелограмма утверждает, что для двух векторов выполнено

$$|v - w|^2 + |v + w|^2 = 2|v|^2 + 2|w|^2.$$

Или словами, сумма квадратов диагоналей равна два умножить на сумму квадратов сторон. Закон параллелограмма может быть проиллюстрирован следующим рисунком. Хорошо известно, что норма возникает из положительно определенного скалярного произведения тогда и только тогда, когда норма удовлетворяет закону параллелограмма, но нас не интересует здесь это направление. Мы должны получить несколько математических фактов в порядке, обратном историческому, а затем сделаем необходимые исторические комментарии.



В формуле, приведенной выше для закона параллелограмма, параллелограмм расположен в начале координат. Закон в действительности связывает длины сторон с длинами диагоналей. Он применяется к параллелограмму, расположенному в произвольной точке x, как показано на следующем рисунке:



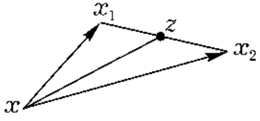
Расстояние между двумя точками x_1, x_2 задается выражением

$$d(x_1, x_2) = |x_1 - x_2|.$$

Тогда в законе параллелограмма утверждается, что

$$d(x_1, x_2)^2 + d(x, x_3)^2 = 2d(x, x_1)^2 + 2d(x, x_2)^2.$$

Приставка «полу» к закону параллелограмма может быть добавлена двумя способами: во-первых, мы остановимся в середине одной из диагоналей, и, во-вторых, мы запишем неравенство вместо равенства. При первом способе мы рисуем половину параллелограмма:



Пусть z — середина отрезка между x_1 и x_2. Тогда в терминах x, x_1, x_2, z закон параллелограмма может быть записан в форме

$$d(x_1, x_2)^2 + 4d(x, z)^2 = 2d(x, x_1)^2 + 2d(x, x_2)^2.$$

Затем мы добавим приставку «полу», записав неравенство $\leqslant$ вместо равенства, и мы также перейдем в более общее пространство, поэтому давайте начнем сначала.

Пусть X — метрическое пространство с функцией расстояния d, поэтому расстояние $d(x, y)$ определено между двумя точками x, y. Оно $\geqslant 0$ и $= 0$ тогда и только тогда, когда $x = y$. Верно равенство

$$d(x, y) = d(y, x),$$

и расстояние удовлетворяет неравенству треугольника

$$d(x, z) \leqslant d(x, y) + d(y, z)$$

для всех точек $x, y, z \in X$. Изначально средняя точка между двумя точками x_1, x_2 не определена. Теперь мы должны описать условие, которое позволит ввести среднюю точку.

Мы будем говорить, что X удовлетворяет ***закону полупараллелограмма***, если для любых двух точек $x_1, x_2 \in X$ существует точка z такая, что для всех $x \in X$

$$d(x_1, x_2)^2 + 4d(x, z)^2 \leqslant 2d(x, x_1)^2 + 2d(x, x_2)^2,$$

следовательно,

$$d(z, x_1) = d(z, x_2) = \frac{1}{2}d(x_1, x_2).$$

Это получается при выборе $x = x_1$ и $x = x_2$ в законе полупараллелограмма, чтобы получить неравенства $2d(x_1, z) \leqslant \leqslant d(x_1, x_2)$ и $2d(x_2, z) \leqslant d(x_1, x_2)$. Противоположные неравенства следуют из неравенства треугольника

$$d(x_1, x_2) \leqslant d(x_1, z) + d(z, x_2).$$

Заметьте, что точка z однозначно определяется точками x_1, x_2, поскольку если z' — другая такая точка, мы выберем $x = z'$ в законе, и увидим, что $z = z'$. Поэтому мы назовем z *средней точкой* между x_1 и x_2.

По определению, *пространство Брюа – Титса* — это полное метрическое пространство, которое удовлетворяет закону полупараллелограмма. «Полное» означает, что любая последовательность Коши в X сходится, т. е. имеет предел в X. Мы обозначим через $\mathbf{B}_R(x)$ открытый шар радиуса R с центром в точке x и через $\overline{\mathbf{B}}_R(x)$ замкнутый шар радиуса R с центром в точке x. Поэтому $\overline{\mathbf{B}}_R(x)$ — это множество точек $y \in X$ таких, что

$$d(x, y) \leqslant R.$$

Пусть X — метрическое пространство. Пусть S — ограниченное подмножество, т. е. существует замкнутый шар $\overline{\mathbf{B}}_R(x)$ для некоторого x, такой, что $S \in \overline{\mathbf{B}}_R(x)$. Пусть $r = \inf$ всех R, таких, что $S \in \overline{\mathbf{B}}_R(x)$ для всех возможных радиусов R и центров $x \in X$.

Вполне может оказаться, что не существует замкнутого шара радиуса r, содержащего S. Если такой шар радиуса r существует, то мы назовем его шаром минимального радиуса, содержащим S. Сейчас мы приведем условия, при которых такой шар существует.

Теорема 1.1 (Серр). *Пусть X — пространство Брюа – Титса. Пусть S — ограниченное подмножество X. Тогда существует единственный замкнутый шар $\overline{\mathbf{B}}_r(x_1)$ минимального радиуса, содержащий S.*

Доказательство.

Сначала докажем единственность.

Пусть существует два шара $\overline{\mathbf{B}}_r(x_1)$ и $\overline{\mathbf{B}}_r(x_2)$ минимального радиуса, содержащих S, но $x_1 \neq x_2$. Пусть x — произвольная точка S, поэтому $d(x, x_2) \leqslant r$ и $d(x, x_1) \leqslant r$. Пусть z — средняя точка отрезка между x_1 и x_2. По закону полупараллелограмма, получим

$$d(x_1, x_2)^2 \leqslant 4r^2 - 4d(x, z)^2.$$

Поскольку по определению r существуют точки $x \in S$ такие, что $d(x, z) \geqslant r - \varepsilon$ для всех $\varepsilon > 0$, следовательно,

$$d(x_1, x_2)^2 \leqslant 4r^2 - 4(r-\varepsilon)^2$$

для всех $\varepsilon > 0$. Вычислив предел при $\varepsilon \to 0$, мы получим, что $d(x_1, x_2) = 0$, следовательно, $x_1 = x_2$. Таким образом, центры шаров совпадают. Поскольку шары имеют один и тот же радиус r, шары совпадают, тем самым доказана единственность.

Что касается существования, пусть $\{x_n\}$ — последовательность точек, которые являются центрами шаров радиуса r_n, стремящихся к инфимуму всех радиусов таких, что $\overline{\mathbf{B}}_{r_n}(x_n)$ содержит S. Пусть r является таким инфимумом. Если последовательность $\{x_n\}$ является последовательностью Коши, то она сходится к некоторой точке, которая является центром замкнутого шара минимального радиуса, содержащего S, и тем самым мы докажем теорему. Несложно записать детально это доказательство. Оставим вам это на самостоятельную работу.

Основная идея заключается в том, чтобы показать, что центры образуют последовательность Коши, т. е. мы всегда получаем вышеописанную ситуацию. Даже в случае плоскости можно доказать нечто, что имеет отношение к геометрии, преподаваемой в старших классах. Мы приведем доказательство.

Пусть z_{mn} — средняя точка между x_n и x_m. Ввиду минимальности r при данном ε существует точка $x \in S$ такая, что

$$d(x, z_{mn})^2 \geqslant r^2 - \varepsilon.$$

Применим закон полупараллелограмма при $z = z_{mn}$. Тогда

$$d(x_m, x_n)^2 \leqslant 2d(x, x_m)^2 + 2d(x, x_n)^2 - 4d(x, z_{mn})^2 \leqslant \\ \leqslant 2r_m^2 + 2r_n^2 - 4r^2 + 4\varepsilon.$$

При $m, n \to \infty$, $2r_m^2 + 2r_n^2 - 4r^2 \to 0$. Тогда неравенство, приведенное выше, показывает, что $\{x_n\}$ — последовательность Коши, тем самым завершая доказательство теоремы. ■

Центр шара в теореме 1.1 называется *круговым центром* множества S.

Пусть X — метрическое пространство. Под *изометрией* X мы будем понимать биекцию

$$g\colon X \to X,$$

такую, что g сохраняет расстояние. Другими словами, для всех $x_1, x_2 \in X$ выполнено

$$d(g(x_1), g(x_2)) = d(x_1, x_2).$$

Если геометрия плоскости правильно преподавалась в старших классах, то вам должно быть известно, что параллельные переносы, вращения и отражения являются изометриями евклидова пространства, и что все изометрии плоскости могут быть получены композицией этих специальных изометрий. В любом случае эти отображения дают примеры изометрий. Заметьте, что если g_1, g_2 — изометрии, такой же будет их композиция $g_1 \circ g_2$. Также, если g — изометрия, то g имеет обратное отображение (поскольку g является биекцией), и из условия изометрии немедленно вытекает, что $g^{-1}\colon X \to X$ также является изометрией. Заметьте, что тождественное отображение $\mathrm{id}\colon X \to X$ является изометрией.

Пусть G — множество изометрий. Будем говорить, что G — *группа* изометрий, если G содержит тождественное отображение, G замкнуто по операции композиция (т. е. если $g_1, g_2 \in G$, то $g_1 \circ g_2 \in G$) и замкнуто по операции обращения (т. е. если $g \in G$, то $g^{-1} \in G$). Часто записывают $g_1 g_2$ вместо $g_1 \circ g_2$. Заметьте, что множество всех изометрий само является группой изометрии.

Пусть $x' \in X$. Подмножество Gx', состоящее из всех элементов $g(x')$, $g \in G$, называется *орбитой* x' на G. Пусть S обозначает такую орбиту. Тогда для всех $g \in G$ и всех элементов $x \in S$ следует, что $gx \in S$. Действительно, мы можем записать $x = g_1 x'$ для некоторого $g_1 \in S$. Тогда

$$g(g_1 x') = g(g_1(x')) = (g \circ g_1)(x') \in S,$$

и $g \circ g_1 \in G$ по предположению. Действительно, $g(S) = S$, поскольку данное $g_1 x'$ может быть записано как

$$g_1 x' = gg^{-1} g_1 x' = gx, \qquad x \in S.$$

После этих предварительных замечаний сформулируем следующую теорему:

Теорема 1.2 (Теорема Брюа–Титса о неподвижной точке). *Пусть X — метрическое пространство Брюа – Титса. Пусть G — группа изометрий X. Действие группы G обозначим $(g, x) \mapsto g \cdot x$. Предположим, что G имеет ограниченную орбиту. (Это происходит, если, к примеру, G является компактной, а действие непрерывным.) Тогда G имеет неподвижную точку, к примеру круговой центр орбиты.*

Доказательство.

Пусть $p \in X$ и пусть $G \cdot p$ — орбита. Пусть $\overline{\mathbf{B}}_r(x_1)$ — единственный замкнутый шар минимального радиуса, содержащий орбиту. Для любого $g \in G$ образ $g \cdot \overline{\mathbf{B}}_r(x_1) = \overline{\mathbf{B}}_r(x_2)$ — замкнутый шар того же радиуса, содержащий орбиту и $x_2 = g \cdot x_1$. Поэтому по свойству единственности из теоремы 1.1 следует, что x_1 — неподвижная точка, что и завершает доказательство. ∎

Для следующего следствия мы предполагаем, что вам известно определение фактор-группы и фактор-пространства G/H относительно подгруппы H. Если вы этого не знаете, пропустите следствие.

Следствие 1.3. *Пусть G — группа, H — подгруппа. Пусть K — подгруппа G такая, что действия K являются переносами на фактор-пространстве G/H. Предположим, что G/H имеет метрику (функцию расстояния) такую, что*

переносы элементами K являются изометриями, G/H — пространство Брюа – Титса, и одна из орбит ограничена. Тогда пространство, сопряженное K содержится в H.

Доказательство.

По теореме 1.2 действие K имеет неподвижную точку, т.е. существует фактор-множество xH такое, что $kxH = xH$ для всех $k \in K$. Тогда $x^{-1}KxH \subset H$, следовательно, $x^{-1}Kx \subset H$, что и требовалось доказать. ■

Заметьте, что теоремы 1.1 и 1.2 совершенно элементарны и могут быть изучены сразу после того, как вы узнаете определение полного метрического пространства. Последующие параграфы несколько более сложны и требуют несколько бо́льших знаний линейной алгебры, смешанной с анализом. Поэтому следующий материал может рассматриваться как интересная тема, дополняющая курсы математического анализа и топологии, изучаемые на старших курсах.

Пространство положительно определенных матриц

Сейчас мы опишем наиболее важный пример пространств Брюа–Титса. В действительности, это типичный пример. Пусть:

$$\mathrm{Mat}_n(\mathbb{R}) = \text{векторное пространство вещественных } n \times n \text{ матриц.}$$

$$\mathrm{GL}_n(\mathbb{R}) = \text{множество вещественных обратимых } n \times n \text{ матриц.}$$

Заметьте, что $\mathrm{GL}_n(\mathbb{R})$ содержится в $\mathrm{Mat}_n(\mathbb{R})$, и $\mathrm{GL}_n(\mathbb{R})$ образует группу, замкнутую по операции умножения и по операции вычисления мультипликативного обратного элемента. Тогда мы скажем, что $\mathrm{GL}_n(\mathbb{R})$ является группой, которую мы будем для краткости обозначать G, поэтому $G = \mathrm{GL}_n(\mathbb{R})$ по определению.

Векторное пространство $\mathrm{Mat}_n(\mathbb{R})$ имеет подпространство:

$$\mathrm{Sym}_n(\mathbb{R}) = \text{пространство симметричных матриц.}$$

Вспомните, что если v — матрица, тогда ее *транспозиция* ${}^t v$ получается по определению заменой строк на столбцы и обратно, и v называется *симметричной*, когда $v = {}^t v$. Мы часто опускаем $\mathbb{R}$ из обозначений и пишем просто Mat_n, GL_n и Sym_n. Конечномерное векторное пространство Sym_n имеет положительно определенное скалярное произведение, которое позволяет рассматривать Sym_n как евклидово пространство. Произведение определяется следующим образом. Вспомните, что след матрицы — это сумма диагональных элементов. Для $v, w \in \mathrm{Sym}_n$ определим их скалярное произведение как след их произведения

$$\langle v, w\rangle_{\mathrm{tr}} = \mathrm{tr}(vw) \quad \text{так, что} \quad |v|^2_{\mathrm{tr}} = \mathrm{tr}(v^2).$$

Здесь и далее tr обозначает след матрицы. Поскольку v — симметричная, след $\mathrm{tr}(v^2)$ является суммой квадратов — проверьте это, записав матрицу произведения. Поэтому немедленно проверяется, что $(v, w) \mapsto \mathrm{tr}(vw)$ — положительно определенное скалярное произведение, которое превращает Sym_n в евклидово пространство размерности $n(n+1)/2$.

Затем вспомним, что матрица p называется *положительной* или *положительно определенной*, если она симметричная и

$$\langle p\xi, \xi\rangle > 0 \qquad \forall \xi \in \mathbb{R}^n,\ \xi \neq 0.$$

Здесь $\langle\ ,\ \rangle$ — обычное скалярное произведение на $\mathbb{R}^n$. Обозначим

$$\mathrm{Pos}_n = \mathrm{Pos}_n(\mathbb{R}) = \text{множество всех симметричных}$$
$$\text{положительно определенных } n \times n \text{ матриц.}$$

Поэтому Pos_n является подмножеством Sym_n. Сразу из определения следует, что Pos_n является открытым подмножеством Sym_n, поэтому геометрически мы представляем Pos_n как открытое подмножество евклидова пространства. Также, как в обычном пространстве, касательное пространство в точке $p \in \mathrm{Pos}_n$ является переносом Sym_n. Не используя более сложного языка, мы можем обозначить его Sym_n также, как мы обозначаем $\mathbb{R}^n$ касательное пространство в любой точке $\mathbb{R}^n$.

Экспоненциальное отображение из Sym_n в Pos_n

Теперь мы установим связь между Sym_n и Pos_n посредством обычного экспоненциального отображения

$$\exp\colon \mathrm{Mat}_n \to \mathrm{GL}_n,$$

заданного обычным степенным рядом

$$\exp(v) = \sum_{k=0}^{\infty} \frac{v^k}{k!}.$$

Если v, w коммутируют (т. е. $vw=wv$), тогда $\exp(v+w) = \exp(v)\exp(w)$. Поэтому $\exp(v)$ обратимо, с обратным элементом, равным $\exp(-v)$.

Мы воспользуемся экспоненциальным отображением, чтобы «искривить» Sym_n. Более точно рассмотрим экспоненциальное отображение лишь на симметричных матрицах

$$\exp\colon \mathrm{Sym}_n \to \mathrm{Pos}_n,$$

поскольку образ Sym_n при экспоненциальном отображении лежит в Pos_n. Чтобы убедиться в этом, заметьте, что если v симметричная и мы выберем $q = \exp(v/2)$, тогда q является симметричной и $\exp(v) = q^2$. Следовательно, $\exp(v)$ является положительно определенной, поскольку для всех векторов $\xi \in \mathbb{R}^n$, $\xi \neq 0$ имеем

$$\langle q^2\xi, \xi\rangle = \langle q\xi, q\xi\rangle > 0.$$

Из линейной алгебры экспоненциальное отображение является дифференциальным (т. е. C^∞) изоморфизмом, в частности оно имеет обратное отображение из класса C^∞, которое может быть названо *логарифм*. Чтобы увидеть это, возьмем положительную матрицу p. Мы можем диагонализировать p, т. е. существует базис $\{\xi_1, \ldots, \xi_n\}$ в $\mathbb{R}^n$ и числа $\lambda_1, \ldots, \lambda_n > 0$ такие, что

$$p\xi_i = \lambda_i\xi_i, \quad i = 1, \ldots, n.$$

Тогда определим $\log p = v$ как линейное отображение, представленное диагональной матрицей

$$\begin{pmatrix} \log\lambda_1 & & 0 \\ & \ddots & \\ 0 & & \log\lambda_n \end{pmatrix}$$

по отношению к базису $\xi_1, \ldots, \xi_n$. Подобно можно определить квадратный корень из p как линейное отображение, представленное матрицей

$$\begin{pmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_n} \end{pmatrix}$$

относительно базиса $\xi_1, \ldots, \xi_n$.

Переменное положительно определенное скалярное произведение

Мы подошли к одной из наших целей, которая заключается в том, чтобы определить положительно определенное скалярное произведение на Sym_n, рассматриваемое как касательное пространство в точке $p \in \mathrm{Pos}_n$, таким способом, что это произведение зависит от p. Мы определим это произведение по следующей формуле

$$\langle v, w\rangle_p = \mathrm{tr}(p^{-1}vp^{-1}w) \quad \text{так, что} \quad |v|_p^2 = |v|_{p,\mathrm{tr}}^2 = \mathrm{tr}((p^{-1}v)^2).$$

Положительная определенность вытекает из факта, что

$$\mathrm{tr}(p^{-1}vp^{-1}v) = \mathrm{tr}(p^{-1/2}vp^{-1/2}p^{-1/2}vp^{-1/2}) = \mathrm{tr}(w^2)$$

с $w = p^{-1/2}vp^{-1/2}$ и $\mathrm{tr}(w^2) > 0$, если $v \neq 0$. Если $t \mapsto p(t)$ является кривой в Pos_n, то производная от длины задается выражением

$$\left(\frac{ds}{dt}\right)^2 = \mathrm{tr}(p^{-1}(t)p'(t))^2,$$

сокращенно

$$ds^2 = \mathrm{tr}((p^{-1}dp)^2).$$

Назовем введенную выше метрику ***метрикой следа.*** Если $\alpha\colon [a, b] \to \mathrm{Pos}_n$ — кривая в Pos_n, тогда ее длина определяется по формуле

$$L(\alpha) = \int_a^b |\alpha'(t)|_{\alpha(t),\mathrm{tr}} dt.$$

[*См. комментарий в конце* § 4.] Даны две точки $p_1, p_2 \in \mathrm{Pos}_n$, определим *расстояние* $d(p_1, p_2)$ как inf (наименьшая нижняя грань) длин всех кривых α, соединяющих p_1 и p_2 в Pos_n. Все кривые предполагаются кусочно C^1, т. е. с непрерывными производными. В качестве простого упражнения по математическому анализу можно доказать, что это определяет функцию расстояния на Pos_n, а именно, она больше либо равна нулю, симметрична и удовлетворяет неравенству треугольника. Таким образом, Pos_n является метрическим пространством.

Теперь перейдем к основной теореме.

Теорема 2.1. *Экспоненциальное отображение является неубывающим по метрике.*

Доказательство будет дано в следующем параграфе. Здесь мы сделаем несколько комментариев.

Прежде всего, свойство увеличения метрики может быть обнаружено в бесконечно малом масштабе. Т. е. для всех $v \in \mathrm{Sym}_n$ пусть

$$\exp'(v)\colon \mathrm{Sym}_n \to \mathrm{Sym}_n$$

обозначает производную экспоненциального отображения в точке v. Таким образом,

$$\exp'(v)w = \lim_{t\to 0} \frac{\exp(v+tw) - \exp(v)}{t}.$$

Тогда $\exp'(v)$ является линейным отображением, и свойство увеличения метрики утверждает, что для всех $w \in \mathrm{Sym}_n$ получим

$$|w|_{\mathrm{tr}} \leqslant |\exp'(v)w|_{\exp v}.$$

Следовательно, обратное отображение является невозрастающим по метрике. Это выполняется для производных, но тогда это также выполняется для функции расстояния, которая является inf длин кривых, и длина определена в терминах производной от кривой.

Следствие 2.2. Pos_n *является полным метрическим пространством.*

Доказательство.

Пусть $\{p_n\}$ — последовательность Коши в Pos_n, $p_n = \exp(v_n)$, $v_n \in \mathrm{Sym}_n$. По свойству невозрастания метрики, $\{v_n\}$ сходится к некоторому $v \in \mathrm{Sym}_n$. По непрерывности сразу получим, что $\{p_n\}$ сходится к $\exp(v)$.

Определим прямую, проходящую через начало координат в Sym_n как обычно. Возьмем «вектор» $v \in \mathrm{Sym}_n$, $v \neq 0$, и прямая, проходящая через начало координат в направлении v является кривой $t \mapsto tv$ $(t \in \mathbb{R})$. Для таких прямых существует более сильная версия теоремы, чем теорема 2.1.

Теорема 2.3. *Экспоненциальное отображение* $\exp\colon \mathrm{Sym}_n \to \mathrm{Pos}_n$ *сохраняет метрику на прямой, проходящей через начало координат. Если* $p \in \mathrm{Pos}_n$, $p = \exp v$, $v \in \mathrm{Sym}_n$, *тогда* $d(e, p) = |v|_{\mathrm{tr}}$.

Доказательство.

Такая прямая имеет вид $t \mapsto tv$ с некоторым $v \in \mathrm{Sym}_n$, $v \neq 0$. Нам нужно доказать, что

$$|v|^2_{\mathrm{tr}} = |\exp'(tv)v|^2_{\exp tv}.$$

Заметьте, что

$$\frac{d}{dt}\exp(tv) = \exp'(tv)v = \frac{d}{dt}\sum \frac{t^n v^n}{n!} = \\ = \sum \frac{t^{n-1}}{(n-1)!}v^n = \exp(tv)v.$$

Следовательно,

$$|\exp'(tv)v|^2_{\exp tv} = \mathrm{tr}(((\exp tv)^{-1}(\exp tv)v)^2) = \mathrm{tr}(v^2) = |v|^2_{\mathrm{tr}},$$

что и доказывает первое утверждение. ■

Что касается второго утверждения, т. е. общей версии теоремы, она получается как приложение теоремы 2.1. Пусть $p = \exp v$. Заметьте, что $|v|_{\mathrm{tr}}$ — длина отрезка прямой между 0 и v. Чтобы увидеть, что $d(e, p) = |v|_{\mathrm{tr}}$, пусть α — кусочно C^1 кривая в Pos_n между e и p длины L. Ее образ при $\exp^{-1}$ (т. е. log) является кривой между 0 и v. Как мы знаем $|v|_{\mathrm{tr}}$ — расстояние между 0 и v в Sym_n. По теореме 2.1 $\log\colon \mathrm{Pos}_n \to \mathrm{Sym}_n$ является неувеличивающим метрику. Следовательно,

$$|v|_{\mathrm{tr}} = d(0, v) \leqslant L.$$

Следовательно, образ при отображении exp отрезка прямой между 0 и v — это кривая между e и p с самой короткой длиной, что является расстоянием между e и p. Это доказывает общую версию теоремы 2.3. ■

Используя две вышеприведенные теоремы, мы можем доказать закон полупараллелограмма «в начале координат», а именно:

Теорема 2.4. *Пусть* $v_1 \in \mathrm{Sym}_n$, $v_1 \neq 0$; $v_2 = -v_1$; $x_1 = \exp(v_1)$, $x_2 = \exp(v_2)$ *и* $z = \exp(0) = e$. *Тогда для любого* $v \in \mathrm{Sym}_n$ *и* $x = \exp(v)$ *получим*

$$d(x_1, x_2)^2 + 4d(x, z)^2 \leqslant 2d(x, x_1)^2 + 2d(x, x_2)^2.$$

Доказательство.

В векторном пространстве Sym имеем в точности закон параллелограмма

$$d(v_1, v_2)^2 + 4d(v, 0)^2 = 2d(v, v_1)^2 + 2d(v, v_2)^2,$$

где $d(v, w) = |v - w|_{\mathrm{tr}}$ для $v, w \in \mathrm{Sym}_n$. При экспоненциальном отображении расстояния в левой части сохраняются по теореме 2.3, и расстояния в правой части увеличиваются, поэтому мы получим в точности закон полупараллелограмма в Pos_n, тем самым доказывая теорему. ■

Чтобы увидеть, что Pos_n является пространством Брюа – Титса, нам осталось показать, что вышеприведенный результат в начале координат верен повсюду, чтобы в точности соответствовать определению. Для этого требуется, чтобы Pos_n являлось однородным пространством, что мы сейчас и обсудим.

Группа $G = \mathrm{GL}_n(\mathbb{R})$ обратимых матриц действует на Pos_n следующим образом. Для $g \in G$ определим $[g]$: $\mathrm{Pos}_n \to \mathrm{Pos}_n$ по формуле

$$[g]p = gp^tg \quad \text{для} \quad p \in \mathrm{Pos}_n.$$

Как обычно ^tg означает транспонированное g. Заметим, что $[g]p$ действительно положительно, поскольку для всех $\xi \in \mathbb{R}^n$ для скалярного произведения $\langle\ ,\ \rangle$ на $\mathbb{R}^n$ имеем

$$\langle [g]p\xi, \xi\rangle = \langle gp^tg\xi, \xi\rangle = \langle p^tg\xi, {}^tg\xi\rangle > 0.$$

Немедленно проверяется, что для g, g_1, $g_2 \in G$ выполнено

$$[g^{-1}] = [g]^{-1} \quad \text{и} \quad [g_1g_2] = [g_1][g_2].$$

Поскольку, как мы отметили раньше, каждая положительная матрица является квадратом положительной матрицы, из этого следует, что G действует транзитивно на Pos_n. Это означает, что при заданных $p_1, p_2 \in \mathrm{Pos}_n$ существует $g \in G$ такая, что $[g]p_1 = p_2$. Действительно, если $p = q^2$ с положительной q, тогда $p = [q]e$. Если $p_1 = [q_1]e$ и $p_2 = [q_2]e$, тогда $p_2 = [q_2q_1^{-1}]p_1$.

Теорема 2.5. *Ассоциация $g \mapsto [g]$ является представлением G в группе изометрий Pos_n, т. е. каждая $[g]$ является изометрией.*

Доказательство.

Прежде всего заметим, что $[g]$ также может рассматриваться как отображение на всем векторном пространстве Sym_n, и это отображение является линейным, как функция от таких матриц. Следовательно, его производная дается формулой

$$[g]'(p)w = gw^tg \quad \text{для всех} \quad w \in \mathrm{Sym}_n.$$

Теперь проверим, что $[g]$ сохраняет скалярное произведение или норму. Имеем

$$\begin{aligned}|[g]'(p)w|^2_{[g]p} &= \mathrm{tr}((([g]p)^{-1}gw^tg)^2) = \mathrm{tr}(((gp^tg)^{-1}gw^tg)^2) = \\ &= \mathrm{tr}({}^tg^{-1}p^{-1}g^{-1}gw^tg^tg^{-1}p^{-1}g^{-1}gw^tg) = \\ &= \mathrm{tr}({}^tg^{-1}p^{-1}wp^{-1}w^tg) = \mathrm{tr}((p^{-1}w)^2) = |w|^2_p,\end{aligned}$$

что и доказывает теорему. ■

Теорема 2.6. *Отображение $p \mapsto p^{-1}$ является изометрией.*

Доказательство.

Доказательство остается в качестве упражнения читателю. Это просто анализ на векторных пространствах. ■

Теорема 2.7. *Пространство Pos_n с метрикой следа является пространством Брюа – Титса, т. е. оно удовлетворяет закону полупараллелограмма.*

Доказательство.

Доказательство заключается в переносе теоремы 2.4 на произвольные точки. Поскольку элементы G действуют как изометрии, достаточно будет доказать, что при заданных $x_1, x_2 \in \mathrm{Pos}_n$ существует $g \in G$ такая, что $g(x_1) = \exp(v_1)$ и $g(x_2) = \exp(-v_1)$ для какого-то $v_1 \in \mathrm{Sym}_n$. Но, во-первых, по транзитивности мы можем найти h_1 такое, что $h_1(x_2) = e$ и $h_1(x_1) = p$ для какого-то $p \in \mathrm{Pos}_n$. Теперь $p^{-1/4} \in \mathrm{Pos}_n$, и пусть h_2 будет отображением

$$h_2 \colon x \mapsto p^{-1/4} x p^{-1/4},$$

поэтому $h_2 = [p^{-1/4}]$. Тогда $h_2(e) = p^{-1/2}$, $h_2(p) = p^{1/2}$. Следовательно, $h_2 \circ h_1 = g$ отображает x_1 на $p^{1/2}$ и x_2 на $p^{-1/2}$. Т. к. существует какое-то v_1, такое, что $p = \exp(v_1)$, то мы получили нужный нам элемент $g \in G$. Это завершает доказательство. ■

Таким образом, единственная вещь, которую осталось сделать в связи с законом полупараллелограмма — это доказать теорему 2.1, что мы сделаем в следующем параграфе. Однако, существует еще один результат, который не зависит от этого закона и который является фундаментальным. Хотя закон полупараллелограмма остается в центре нашего обсуждения, мы приведем здесь эту фундаментальную теорему. Это явная формула, дающая расстояние между двумя точками p и q в Pos_n.

Теорема 2.8. *Пусть $a_1, \ldots, a_n$ — корни полинома $\det(tp - q)$. Тогда*

$$d(p, q) = \sum (\log a_i)^2.$$

Доказательство.

Способ доказательства похож на тот, который использовался при доказательстве теоремы 2.7 при использовании переносов.

Во-первых, положим $p = e$ (единичная матрица) и $q = d$ — диагональная матрица с компонентами $a_1, \ldots, a_n$. Пусть v — диагональная матрица с компонентами $\log a_i$ $(i = 1, \ldots, n)$. Тогда $d = \exp(v)$. По теореме 2.3 получим

$$d_{\mathrm{Pos}}(e, d) = d_{\mathrm{Sym}}(0, v) = |v^2|_{\mathrm{tr}} = \sum (\log a_i)^2,$$

таким образом, доказывая формулу в этом случае, мы сведем общий случай к вышеприведенному частному случаю. Мы утверждаем, что существует $g \in G$, такое, что $[g]p = e$ и $[g]q = d$ — диагональная. Действительно, мы можем сначала перенести p в e, используя некоторый элемент $h_1 \in G$, поэтому без потери общности мы можем предположить, что $p = e$. Существует ортонормированный базис в $\mathbb{R}^n$, который диагонализует q, поэтому существует диагональная матрица d и вещественная унитарная матрица k (сохраняющая евклидову норму в $\mathbb{R}^n$) такая, что $q = kdk^{-1} = kd^tk$, поскольку ${}^tk = k^{-1}$. Но $e = kk^{-1} = k^tk$, поэтому выбор $[k]q$ доказывает наше утверждение. Наконец, из уравнений $qp^tg = e$ и $gq^tg = d$ с диагональной матрицей d мы получим $p = g^{-1t}g^{-1}$ и $q = g^{-1}d^tg^{-1}$, поэтому

$$\det(tp - q) = \det(tg^{-1t}g^{-1} - g^{-1}d^tg^{-1}) = \det(g)^{-2}\det(te - d).$$

Поэтому корни $\det(tp - q)$ те же самые, что и корни $\det(te - d)$ и $\mathrm{dist}(p, q) = \mathrm{dist}(e, d)$. Это и доказывает теорему. ■

Свойство увеличения метрики для экспоненциального отображения

Этот параграф посвящен доказательству теоремы 2.1, следуя, в сущности, [Mo 53]. Используем сокращение

$$\mathrm{Mat}_n(\mathbb{R}) = \mathscr{A}.$$

Мы будем писать Sym вместо Sym_n.

Пусть v, $w \in \mathscr{A}$. Определим

$$F_v(w) = \exp\Bigl(-\frac{v}{2}\Bigr) \cdot \exp'(v)w \cdot \exp\Bigl(-\frac{v}{2}\Bigr) = e^{-v/2}\exp'(v)we^{-v/2}.$$

Заметьте, что

$$\exp'(v)w = \frac{d}{dt}\exp(v + tw)\Big|_{t=0}.$$

Непосредственно из определений получаем

$$\exp'(v)w = \sum_{n=0}^{\infty}\frac{1}{n!}\sum_{r+s=n-1} v^r w v^s. \qquad (1)$$

Т. к. степени элемента коммутируют между собой, отметим, что $\exp\Bigl(-\frac{v}{2}\Bigr)$ коммутирует со степенями v^r, v^s.

Лемма 3.1. *Отображения F_v и $\exp'(v)$ симметричны по отношению к tr-скалярному произведению на $\mathscr{A}$. Если $v \in \mathrm{Sym}$, тогда F_v и $\exp'(v)$ отображают Sym в само себя.*

Доказательство.

Обычная проверка дает для u, v, $w \in \mathscr{A}$:

$$\begin{aligned}\mathrm{tr}(F_v(w)u) &= \sum_{n=0}^{\infty}\frac{1}{n!}\sum_{r+s=n-1}\mathrm{tr}(\exp\Bigl(-\frac{v}{2}\Bigr)v^r w v^s \exp\Bigl(-\frac{v}{2}\Bigr)u) = \\ &= \sum_{n=0}^{\infty}\frac{1}{n!}\sum_{r+s=n-1}\mathrm{tr}(wv^s\exp\Bigl(-\frac{v}{2}\Bigr)u\exp\Bigl(-\frac{v}{2}\Bigr)v^r) = \\ &= \mathrm{tr}(wF_v(u)),\end{aligned}$$

т. к. $\exp\Bigl(-\frac{v}{2}\Bigr)$ коммутирует с v^r и v^s, это завершает доказательство того, что F_v симметрично по отношению к tr-скалярному произведению. Если $v \in \mathrm{Sym}$, то формула (1) показывает, что F_v отображает Sym в само себя. Утверждения о $\exp'(v)$ доказываются по тому же шаблону. ∎

Определим левое умножение $L_v\colon \mathscr{A} \to \mathscr{A}$, $L_v(w) = vw$ и R_v — правое умножение. Пусть $D_v = L_v - R_v$, поэтому

$$D_v(w) = vw - wv = [v, w] \quad \text{для всех} \quad v, w \in \mathscr{A}.$$

Лемма 3.2. *Пусть* $v \in \mathrm{Sym}$. *Тогда* D_v^2 *симметрично на Sym. Другими словами, для всех* $w, u \in \mathrm{Sym}$ *имеем*

$$\langle u, D_v^2 w\rangle_{\mathrm{tr}} = \langle D_v^2 u, w\rangle_{\mathrm{tr}},$$

или записанное другим способом

$$\mathrm{tr}(u\, D_v^2 w) = \mathrm{tr}(w\, D_v^2 u).$$

Доказательство.

Доказательство снова рутинное, а именно

$$\begin{aligned} D_v(w) &= vw - wv, \\ D_v^2(w) &= v^2 w - 2vwv + wv^2, \\ (D_v^2 w)u &= v^2 wu - 2vwvu + wv^2 u, \\ w(D_v^2 u) &= wv^2 u - 2wvuv + wuv^2. \end{aligned}$$

Применяя tr к этим двум последним выражениям и используя его основное свойство $\mathrm{tr}(xy) = \mathrm{tr}(yx)$, получаем доказательство леммы. ∎

Напомним, что симметричный оператор B на векторном пространстве с положительно определенным скалярным произведением называется *неотрицательным* и записывается $B \geqslant 0$, если выполняется

$$(Bw, w) \geqslant 0$$

для всех w в пространстве. Затем определим $B_1 \geqslant B_2$, если $B_1 - B_2 \geqslant 0$.

В последующих доказательствах воспользуемся двумя основными свойствами.

Спектральное свойство. Пусть M — симметричное линейное отображение конечномерного векторного пространства над $\mathbb{R}$ с положительно определенным скалярным произведением. Пусть $b \geqslant 0$. Пусть $f_0(t)$ — сходящийся степенной ряд, такой, что $f_0(t) \geqslant b$ для всех t в интервале, содержащем собственные числа M. Тогда $f_0(M) \geqslant bI$.

Доказательство.

Доказывается сразу же диагонализацией линейного отображения по отношению к базису. ∎

А также:

Свойство Шварца. Для всех v, $w \in \mathrm{Sym}$

$$\mathrm{tr}((vw)^2) \leqslant \mathrm{tr}(v^2 w^2).$$

Для удобства читателя мы напомним доказательство.

Доказательство.

Матрицы (линейные отображения) v и w могут быть одновременно диагонализованы, если одна из них положительно определена, и в этом случае неравенство эквивалентно обычному неравенству Шварца. Если обе матрицы сингулярны, то можно рассмотреть матрицу $w + \varepsilon e$, где e — единичная матрица и $\varepsilon > 0$. Тогда $w + \varepsilon e$ несингулярная для всех достаточно малых $\varepsilon \neq 0$, и тогда можно использовать предыдущий несингулярный случай с последующим вычислением предела при $\varepsilon \to 0$. Это завершает доказательство. ■

Определим формальный степенной ряд:

$$f(t) = \sum_{k=0}^{\infty} \frac{(t/2)^{2k}}{(2k+1)!} = \frac{\mathrm{sh}(t/2)}{t/2} = \frac{\exp(t/2) - \exp(-t/2)}{t}.$$

Заметим, что L_v и R_v коммутируют между собой и поэтому

$$\exp\Big(\frac{D_v}{2}\Big) = \exp\Big(\frac{L_v}{2}\Big)\exp\Big(-\frac{R_v}{2}\Big).$$

Мы можем вычислить $f(D_v)$. Т.к. лишь четные степени D_v возникают в степенном ряде для f, следовательно если $v \in \mathrm{Sym}$, то $f(D_v)$ отображает Sym само в себя и оператор

$$f(D_v)\colon \mathrm{Sym} \to \mathrm{Sym}$$

симметричный для tr-скалярного произведения.

Лемма 3.3. *Для любых $v \in \mathscr{A}$ имеем $D_v F_v = D_v f(D_v)$.*

Доказательство.

Пусть $t \mapsto x(t)$ — гладкая кривая в $\mathscr{A}$. Тогда

$$x(\exp x) = (\exp x)x.$$

Дифференцирование обеих сторон дает

$$x' \exp x + x(\exp x)' = (\exp x)'x + (\exp x)x',$$

и поэтому

$$x' \exp x - (\exp x)x' = (\exp x)'x - x(\exp x)'.$$

Умножая левую и правую часть на $\exp\left(-\frac{x}{2}\right)$ и используя тот факт, что x коммутирует с $\exp\left(-\frac{x}{2}\right)$, приходим к выражению

$$\exp\left(-\frac{x}{2}\right)x' \exp\left(\frac{x}{2}\right) - \exp\left(\frac{x}{2}\right)x' \exp\left(-\frac{x}{2}\right) = \tag{2}$$
$$= \exp\left(-\frac{x}{2}\right)(\exp x)' \exp\left(-\frac{x}{2}\right)x - x \exp\left(-\frac{x}{2}\right)(\exp x)' \exp\left(-\frac{x}{2}\right).$$

Т. к. L_x и R_x коммутируют, получим

$$\exp\left(\frac{D_x}{2}\right) = \exp\left(\frac{L_x}{2}\right) \exp\left(-\frac{R_x}{2}\right),$$

поэтому (2) может быть записано в виде

$$\left(\exp\left(\frac{D_x}{2}\right) - \exp\left(-\frac{D_x}{2}\right)\right)x' = D_x F_x x'. \tag{3}$$

Теперь возьмем кривую $x(t) = v + tw$ и вычислим предыдущее равенство в $t = 0$, поэтому $x'(0) = w$, что завершает доказательство леммы. ■

Теорема 3.4. *Пусть* $v \in \mathrm{Sym}$. *Тогда* $F_n = f(D_v)$ *на* Sym. *Следовательно, для* $w \in \mathrm{Sym}$, *получим*

$$\exp'(v)w = \exp\left(\frac{v}{2}\right) \cdot f(D_v)w \cdot \exp\left(\frac{v}{2}\right).$$

Доказательство.

Пусть $h_v = F_v - f(D_v)$. Тогда $h_v \colon \mathrm{Sym} \to \mathrm{Sym}$ — симметричное линейное отображение, и его образ содержится в подпространстве $E = \operatorname{Ker} D_v \cap \mathrm{Sym}$. Т. к. Sym — конечномерное, оно является прямой суммой E и его ортогонального дополнения $E^{\perp}$ в Sym. Поскольку h_v — симметричное, оно отображает $E^{\perp}$ в $E^{\perp}$, но h_v также отображает $E^{\perp}$ в E, поэтому $h_v = 0$ на $E^{\perp}$. Дополнительно E — коммутант v в Sym и, следовательно, $f(D_v) = \mathrm{id} = F_v$ на E, поэтому $h_v = 0$ на E. Следовательно, $h_v = 0$ на Sym, завершая таким образом доказательство теоремы. ■

Теорема 3.5. *Пусть* $v \in \mathrm{Sym}$, *тогда* D_v^2 *— неотрицательное и* $f(D_v) \geqslant I$.

Доказательство.

По лемме 3.2 для $w \in \mathrm{Sym}$ выполнено

$$\langle D_v^2 w, w\rangle_{\mathrm{tr}} = \mathrm{tr}(wv^2w - 2vwvw + v^2w^2) = 2\mathrm{tr}(v^2w^2 - (vw)^2).$$

поэтому неотрицательность D_v^2 вытекает из свойства Шварца tr-скалярного произведения. Теперь можем записать

$$f(t) = f_0(t^2),$$

где $f_0(t)$ — очевидный степенной ряд с положительными коэффициентами. Заметим, что $f_0(t) \geqslant 1$ для всех $t \geqslant 0$. Поэтому из спектрального свойства степенного ряда следует, что

$$f(D_v) = f_0(D_v^2) \geqslant I.$$

Это и завершает доказательство. ■

Теорема 3.6. *Экспоненциальное отображение* exp *не уменьшает* tr-*норму на* Sym, *т. е. для* $v, w \in \mathrm{Sym}$, *выбирая* $p = \exp(v)$, *имеем*

$$|w|_{\mathrm{tr}}^2 = \mathrm{tr}(w^2) \leqslant \mathrm{tr}((p^{-1}\exp'(v)w)^2) = |\exp'(v)w|_{p,\mathrm{tr}}^2.$$

Доказательство.

Правая часть вышеприведенного неравенства равна

$$\mathrm{tr}((p^{-1}\exp'(v)w)^2) = \mathrm{tr}\left(\left(\exp\left(\frac{-v}{2}\right)\cdot \exp(v)w \cdot \exp\left(\frac{-v}{2}\right)\right)^2\right) = \\ = \mathrm{tr}(F_v(w)^2) = |f(D_v)w|_{\mathrm{tr}}^2$$

по теореме 2.4.

Применение теоремы 3.5 завершает доказательство. ■

Следствие 3.7. *Для каждого* $v \in \mathrm{Sym}$ *отображения*

$$F_v \quad \textit{и} \quad \exp'(v)\colon \mathrm{Sym} \to \mathrm{Sym}$$

линейные автоморфизмы.

Доказательство.

Теорема 3.6 оказывает, что $\operatorname{Ker}\exp'(v) = 0$ и $\exp'(v)$ — линейный изоморфизм. Тогда следует утверждение для F_v, поскольку F_v составлено из $\exp'(v)$ и мультипликативных переносов с помощью обратимых элементов в Sym. Это завершает доказательство. ∎

Заметим, что теорема 3.6 завершает доказательство теоремы 2.1. ∎

Исторические примечания

Представление вышеизложенного материала в сущности следует пути, обратному историческому. Почти столетие потребовалось, чтобы определенные идеи получили их полную общность и простоту.

Исторически развитие этих идей началось в конце девятнадцатого века. Клингенберг [Kl 83/95] утверждает, что фон Мангольд в сущности доказал то, что сейчас называют теоремой Картана – Адамара для поверхностей [vM 81], на 15 лет раньше, чем это сделал Адамар [Ha 86]. В действительности фон Мангольд ссылается на предыдущие работы, написанные другими учеными, Адамар ссылается на фон Мангольдта, и Картан [Ca 28/46] ссылается на Адамара (Картан рассматривает произвольные римановы многообразия). Я не имел возможности прочитать оригинальные статьи.

Хельгасон [He 62] получил доказательство теоремы Картана о неподвижной точке, следуя идеям Картана, см. обновленную версию [He 78], глава 1, теорема 13.5, а именно: на римановых многообразиях неположительной кривизны компактная группа изометрий имеет неподвижную точку. Целью Картана было показать, что все максимальные компактные подгруппы полупростой группы Ли сопряженные. Мостоу [Mo 53] получил похожее утверждение, но в более ограниченном контексте. Они использовали центр масс вместо кругового центра.

Затем Брюа и Титс [BrT 72] сформулировали свою теорему о неподвижной точке (здесь сформулированную как теорема 1.2), установив условия закона полупараллелограмма. Серр использовал вариацию их доказательства и формулировку теоремы 1.1, чтобы достичь текущего окончательного результата с очень простым доказательством, включенным в эту книгу. Я не знаю, где Серр опубликовал это, но ссылаются на в точности ту же формулировку у Брауна [Bro 89], глава 6, теорема 2, § 5. Таким образом, линия размышлений, начатых столетием раньше, привела к фундаментальной простой теореме о метрических пространствах. Условие компактности сменилось условием ограниченности, и более сложное понятие кривизны сменилось законом полупараллелограмма.

Дополнительно центр масс, который возник в утверждении Картана (и других, следующих ему), заменен круговым центром, следуя Брюа – Титсу.

Брюа и Титс в действительности просто охарактеризовали связные римановы многообразия с неположительной кривизной законом полупараллелограмма [BrT 72]. С этого момента теория кривизны для метрических пространств, а не многообразий, развивалась отдельно, с расширенным изложением у Болмана [Ba 95], содержащим теорему 1.1. Болман ссылается на Брауна по теореме 1.1, см. [Ba 95], теорема 5.1 и предложение 5.10 главы 1.

Развитие теории симметричных пространств является заслугой Картана [Ca 28/46] и [Ca 51]. Мостоу привел очень элегантное изложение некоторых результатов Картана в [Mo 53], и доказательство, представленное в § 3, в сущности, взято из изложения Мостоу.

В стандартном базовом курсе анализа определяется *евклидова длина* кривой в открытом множестве U в $\mathbb{R}^N$ как

$$L(\alpha) = \int_a^b |\alpha'(t)|\, dt,$$

где норма — стандартная евклидова норма. В конце девятнадцатого столетия различные люди искали геометрию, отличную от евклидовой, и рассматривали переменные метрики, где метрика в каждой точке зависит от точки. Простые примеры переменной нормы уже можно встретить в некоторых курсах.

Например, пусть D — единичный круг в $\mathbb{R}^2$. Для каждой точки $x \in D$ пусть $r = |x|$ — евклидово расстояние от центра координат. Задан вектор $v \in \mathbb{R}^2$. Определим *метрику Лобачевского* или *Пуанкаре* как

$$|v|_x = \frac{1}{1-r^2}|v|_{\mathrm{euc}}, \qquad r = |x| = \sqrt{x_1^2 + x_2^2}.$$

Таким образом, $|v|_x$ зависит от расстояния от центра и $|v|_x \to \infty$, когда $|x| \to 1$. Зависимость от x на единичном круге по прежнему достаточно простая по сравнению с метрикой следа в евклидовом пространстве Sym.

Для систематического изложения того, как экспоненциальное свойство увеличения метрики связано с законом полупараллелограмма, см. [La 99], где результаты обсуждаются в контексте дифференциальной геометрии. Заметим, что из этого контекста можно также получить, что круг с метрикой Лобачевского – Пуанкаре является пространством Брюа – Титса и, в частности, удовлетворяет закону полупараллелограмма. Чтобы увидеть, что этот пример в сущности не отличается от приведенного примера с экспоненциальным отображением, потребуется общая теория экспоненциальных отображений, содержащая обе ситуации, которые априорно не выглядят похожими.

Литература

[Ba 95] W. Ballman, *Lectures on Spaces of Nonpositive Curvature*, Birkhäuser, 1995.

[BGS 85] W. Ballman, M. Gromov, and V. Schroeder, *Manifolds of nonpositive Curvative*, Birkhäuser, 1985.

[Bro 89] K. Brown, *Buildings*, Springer-Verlag, 1989.

[BrT 72] F. Bruhat and J. Tits, Groupes Réductifs sur un Corps Local I, *Pub. IHES*, **41** (1972), pp. 5–251.

[Ca 27a] E. Cartan, Sur une classe remarquable d'espaces de Riemann, *Bull. Soc. Math. France*, **54** (1927), pp. 114–134.

[Ca 27b] E. Cartan, sur certaines formes Riemanniennes remarquables des géometries á groupe fondamental simple, *Ann. Sci. Ecole Norm. Sup.* **44** (1927), pp. 345–467.

[Ca 28/46] E. Cartan, *Leçons sur la Géometrie des Espasec de Riemann*, Gauthiers-Villars, 1928; Second edition, 1946.

[Ca 51] E. Cartan, *Leçons sur la Géometrie des Espasec de Riemann* II, Gauthiers-Villars, 1951.

[Ha 86] J. Hadamard, Les surfaces á courbures opposées et leur lignes géodeiques, *J. Math. Pures Appl.*, (5) **4** (1896), pp. 27–73.

[He 62] S. Helgason, *Differential Geometry and Symmetric Spaces*, Academic Press, 1962.

[He 78] S. Helgason, *Differential Geometry, Lie Groups, and Symmetric Spaces*, Academic Press, 1978.

[Kl 83/95] W. Klingenberg, *Riemannian Geometry*, de Gruyter 1983; Second edition 1995.

[La 99] S. Lang, *Fundamentals of Differential Geometry*, Springer-Verlag, 1999.

[vM 81] H. von Mangoldt, Über diejenigen Punkte auf positv gekrümmten Flächen, welche die Eigenschaft haben, dass die von ihnen ausgehenden geodätischen Linien nie aufhören, kurzeste Linien zu sein, *J. reine angew. Math.* **91** (1881), pp. 23–52.

[Mo 53] D. Mostov, *Some New Decomposition Theorems for Semisimple Groups*, Memoirs AMS, 1953.

Гармонические и симметрические полиномы

Вернемся к более алгебраической теме. Курсы по линейной алгебре или по общей алгебре для студентов дают основы, но чаще всего нет времени, чтобы заниматься углубленным изучением, требующим объединения материала. Представленная тема иллюстрирует концепции из линейной алгебры в контексте, который вплоть до сегодняшнего времени изучался лишь аспирантами. Нет ничего сложного в любом из рассмотренных вопросов, но также, как пространство положительно определенных матриц дало возможность для изложения информации из сложных курсов анализа, кольцо полиномов и два важных подмножества — гармонические и симметрические полиномы — предоставляют возможность для иллюстрации концепций из линейной алгебры, и, конечно же, из математического анализа, связанных с частными производными. При работе с полиномами аналитическая часть в основном иллюзорна. Дифференцирование полинома является чисто алгебраическим понятием, мы можем определить производную полинома одной переменной x как

$$\frac{d}{dx}\sum_{k=0}^{d} x^k = \sum_{k=1}^{d} kx^{k-1}.$$

Аналогично для полинома нескольких переменных частные производные могут быть определены аналогичными чисто алгебраическими формулами. Никаких пределов не потребуется для основных теорем этой части.

Гармонические и симметрические полиномы

Мы будем рассматривать полиномы от нескольких переменных над вещественными числами. Некоторые основные факты мы будем полагать известными и приведем их во введении. Пусть переменные обозначаются $x_1, \ldots, x_n$. Множество всех полиномов от этих переменных с вещественными коэффициентами будет обозначаться

$$\mathbb{R}[x_1, \ldots, x_n].$$

Полином вида $x_1^{j_1} \cdots x_n^{j_n}$ будет называться ***простым мономом.*** Любой полином представляет собой линейную комбинацию

$$P(x_1, \ldots, x_n) = \sum c_{j_1, \ldots, j_n} x_1^{j_1} \cdots x_n^{j_n}$$

с вещественными коэффициентами $c_{(j)} \in \mathbb{R}$. Мы запишем простой моном в сокращенной форме

$$M_{(j)}(x) = x^{(j)} = x_1^{j_1} \cdots x_n^{j_n},$$

поэтому $P(x) = \sum c_{(j)} M_{(j)}(x)$. ***Степень*** $M_{(j)}(x)$ определяется как

$$\deg M_{(j)} = j_1 + \ldots + j_n.$$

Различные мономы $M_{(j)}$ данной степени d линейно независимы над $\mathbb{R}$. Полином $P(x)$, определенный, как указано выше, называется ***однородным***, если все мономы с ненулевыми коэффициентами имеют одинаковую степень d. Например, полином

$$3x_1^3 x_2 - 7x_1 x_2 x_3^2$$

однородный степени 4, но $3x_1^3 x_2 - x_1^2 x_2$ не однородный. Мы обозначим $\mathrm{Pol}(n, d)$ векторное пространство над $\mathbb{R}$, порождаемое мономами степени d от n переменных $x_1, \ldots, x_n$. Его размерность равна количеству простых мономов $M_{(j)}(x)$ степени d. Мы не будем использовать этот факт, но нетрудно доказать, что это число является биномиальным коэффициентом

$$\binom{n-1+d}{n-1}.$$

Другими словами

$$\dim \mathrm{Pol}(n, d) = \binom{n-1+d}{n-1}.$$

Мы можем взять производные от полиномов при помощи частных производных, которые мы обозначим через $\partial_1, \ldots, \partial_n$. При подстановке переменной в обозначение эти частные производные становятся обычными частными производными

$$\partial_i = \frac{\partial}{\partial x_i}, \qquad i = 1, \ldots, n.$$

Частная производная от полинома — это еще один полином. Если $P(x)$ — однородный степени $d \geqslant 1$, тогда $\partial_i P$ — однородный полином степени $d - 1$ или 0. Разумеется, если P имеет степень 0, то P — константа, и $\partial_i P = 0$ для всех i. Заметим, что можно записать дифференциальный оператор в виде

$$P(\partial_1, \ldots, \partial_n),$$

или сокращенно $P(\partial)$, т. е. мы можем подставить частные производные вместо переменных. Для любого полинома Q, если $P = \sum c_{(j)} M_{(j)}$, то

$$P(\partial_1, \ldots, \partial_n)Q = \sum c_{(j)} \partial_1^{j_1} \cdots \partial_n^{j_n} Q.$$

Положительно определенное скалярное произведение

Определим пример векторного пространства со скалярным произведением, используя $\mathrm{Pol}(n, d)$ и частные производные следующим образом. Для двух полиномов

$$P, Q \in \mathrm{Pol}(n, d)$$

определим *скалярное произведение*

$$\langle P, Q \rangle = (P(\partial)Q)(0).$$

Заметим, что билинейность получается сразу, т. е

$$\langle P_1 + P_2, Q \rangle = \langle P_1, Q \rangle + \langle P_2, Q \rangle,$$
$$\langle P, Q_1 + Q_2 \rangle = \langle P, Q_1 \rangle + \langle P, Q_2 \rangle.$$

Если $c \in \mathbb{R}$, то $\langle cP, Q \rangle = c\langle P, Q \rangle = \langle P, cQ \rangle$.

Более интересен тот факт, что скалярное произведение симметричное, т. е. $\langle P, Q\rangle = \langle Q, P\rangle$ для всех однородных полиномов $P, Q \in \mathrm{Pol}(n, d)$. Проверить это свойство достаточно для случая, когда P, Q — простые мономы. Тогда мы получим явное значение для скалярного произведения.

Пусть $P = M_{(j)}(x)$ и $Q = M_{(k)}(x)$ — мономы степени d. Если $(j) = (k)$, то

$$\langle M_{(j)}, M_{(j)}\rangle = j_1! \dots j_n!$$

Если $(j) \neq (k)$, то

$$\langle M_{(j)}, M_{(k)}\rangle = 0.$$

В частности, мономы образуют ортогональный базис для $\mathrm{Pol}(n, d)$.

Доказательство.

Предположим, что $(j) = (k)$. Это означает, что $j_1 = k_1, \dots$ $\dots, j_n = k_n$. Тогда

$$M_{(j)}(\partial)M_{(k)} = \partial_1^{j_1} \cdots \partial_n^{j_n} x_1^{j_1} \cdots x_n^{j_n}.$$

Вычисляя производные обычным способом, мы находим, что получается значение из утверждения $j_1! \dots j_n!$. Предположим, что $(j) \neq (k)$, тогда для некоторого i: $j_i \neq k_i$. Скажем, $j_1 \neq k_1$. Рассмотрим

$$\partial_1^{j_1} x_1^{k_1}.$$

Если $j_1 > k_1$, то мы взяли больше производных, чем степень x_1, и поэтому получим 0. Если $j_1 < k_1$, тогда $\partial_1^{j_1} x_1^{k_1}$ равно константе умноженной на $x_1^{k_1 - j_1}$, т. е. осталась положительная степень от x_1. Вычисляем при $(x) = 0$, поэтому $x_1 = 0$, и мы получим значение 0. Это доказывает ортогональность мономов и также показывает, в частности, что

$$\langle M_{(j)}, M_{(j)}\rangle > 0.$$

Мы приходим к заключению, что наше скалярное произведение симметрично, поскольку скалярное произведение двух мономов симметрично. Симметрия в общем случае следует из билинейности.

Мы также получаем положительную определенность из билинейности и из факта, что для $c_{(j)}, c_{(k)} \in \mathbb{R}$

$$\langle c_{(j)} M_{(j)}, c_{(k)} M_{(k)} \rangle = \begin{cases} 0, & \text{если } (j) \neq (k), \\ c_{(j)}^2 j_1! \dots j_n!, & \text{если } (j) = (k). \end{cases}$$

Поэтому если $P = \sum c_{(j)} M_{(j)} \neq 0$, то

$$\langle P, P \rangle = \sum c_{(j)}^2 j_1! \dots j_n! > 0.$$

Таким образом, мы получили довольно интересный пример положительно определенного скалярного произведения, смешивая вместе линейную алгебру и дифференцирование. ■

Мы упомянем еще одну формулу. Как и раньше, пусть $P \in \mathrm{Pol}(n, d)$ и $0 \leqslant m \leqslant d$.

Формула объединения. *Пусть* $R \in \mathrm{Pol}(n, m)$ *и* $Q \in \in \mathrm{Pol}(n, d - m)$, *тогда*

$$\langle R(\partial)P, Q \rangle = \langle P, RQ \rangle.$$

Доказательство в одну строку:

$$\langle P, RQ \rangle = (RQ)(\partial)P(0) = Q(\partial)R(\partial)P(0) = \langle R(\partial)P, Q \rangle,$$

потому что вычисление частных производных — коммутативная операция. Это и доказывает формулу. ■

Давайте рассмотрим множество всех полиномов $\mathbb{R}[x_1, \dots, x_n]$, которое мы обозначим S,

$$S = \mathbb{R}[x_1, \dots, x_n].$$

Используем сокращение

$$S^{(d)} = \mathrm{Pol}(n, d)$$

для пространства однородных элементов степени d (отличной от 0).

В этот момент можно выбрать два независимых направления. Можно рассмотреть гармонические полиномы, которые будут определены в дальнейшем, или симметрические полиномы совместно с обобщением гармонических полиномов. Мы разделили эти темы, но вы, возможно, захотите прочесть их одновременно.

Гармонические полиномы

Пусть
$$\Delta = \partial_1^2 + \ldots + \partial_n^2.$$

Из математического анализа функций нескольких переменных вам должно быть известно, что Δ называется *оператор Лапласа*. Функция f такая, что $\Delta f = 0$ называется *гармонической*. Нас интересуют гармонические полиномы, т.е. полиномы P такие, что $\Delta P = 0$. Заметим, что

$$\Delta\colon \mathrm{Pol}(n,\, d) \to \mathrm{Pol}(n,\, d-2)$$

линейное отображение из векторного пространства полиномов степени d в векторное пространство полиномов степени $d-2$.

Теорема 2. *Это линейное отображение является сюръекцией. Другими словами, для заданного полинома Q степени $d-2$ существует $P \in \mathrm{Pol}(n,\, d)$ такой, что $\Delta P = Q$.*

Теорема может быть доказана силовым методом. Мы приведем другое, более красивое доказательство, использующее линейную алгебру.

Определим $\mathrm{Har}(n,\, d)$ — множество всех однородных гармонических полиномов степени d или 0. Очевидно, что $\mathrm{Har}(n,\, d)$ — подпространство $\mathrm{Pol}(n,\, d)$, поэтому в терминах линейной алгебры

$$\mathrm{Har}(n,\, d) = \mathrm{Ker}\,\Delta.$$

Другими словами, $\mathrm{Har}(n,\, d)$ — ядро линейного отображения

$$\Delta\colon \mathrm{Pol}(n,\, d) \to \mathrm{Pol}(n,\, d-2).$$

Пусть
$$r^2 = r^2(x_1,\, \ldots,\, x_n) = x_1^2 + \ldots + x_n^2.$$

Поэтому r^2 — однородный полином степени 2. Заметим, что

$$r^2(\partial) = \Delta.$$

Мы доказали в общем случае правило перестановки

$$\langle R(\partial)P,\, Q\rangle = \langle P,\, RQ\rangle.$$

Применим его к специальному случаю $R = r^2$. Таким образом, получим

$$\langle \Delta P, Q \rangle = \langle P, r^2 Q \rangle$$

для всех $P \in \mathrm{Pol}(n, d)$, $Q \in \mathrm{Pol}(n, d-2)$.

Это позволяет нам доказать сюръективность Δ на $\mathrm{Pol}(n, d-2)$. Предположим, что образ Δ не равен всему $\mathrm{Pol}(n, d-2)$. Тогда существует полином $Q \in \mathrm{Pol}(n, d-2)$, $Q \neq 0$, который ортогонален образу Δ по теореме из линейной алгебры для положительно определенного скалярного произведения. Их этого следует, что

$$0 = \langle \Delta P, Q \rangle = \langle P, r^2 Q \rangle$$

для всех $P \in \mathrm{Pol}(n, d)$. Следовательно, $r^2 Q$ ортогонально всему $\mathrm{Pol}(n, d)$. Т. к. скалярное произведение положительно определено, мы должны получить $r^2 Q = 0$, следовательно $Q = 0$, противоречие, которое доказывает, что

$$\Delta \colon \mathrm{Pol}(n, d) \to \mathrm{Pol}(n, d-2)$$

сюръективно.

Таким образом, вы видите линейную алгебру и положительно определенные скалярные произведения, действующие в контексте, который обычно не рассматривается в стандартных курсах. Мы приведем еще один пример такого же характера.

Теорема 3. *Пространство однородных полиномов степени d допускает ортогональное разложение*

$$\mathrm{Pol}(n, d) = \mathrm{Har}(n, d) + r^2 \mathrm{Pol}(n, d-2).$$

Это означает, что $\mathrm{Pol}(n, d)$ *является суммой двух пространств из правой части, и эти два пространства* $\mathrm{Har}(n, d)$, $r^2\mathrm{Pol}(n, d-2)$ *ортогональны.*

Доказательство.

Общее правило перестановки показывает, что $\mathrm{Har}(n, d)$ и $r^2\mathrm{Pol}(n, d-2)$ ортогональны, поскольку для $P \in \mathrm{Har}(n, d)$ и $Q \in \mathrm{Pol}(n, d-2)$, получим

$$\langle P, r^2 Q \rangle = \langle \Delta P, Q \rangle = 0,$$

т. к. P — гармонический полином. Следовательно,

$$\dim[\operatorname{Har}(n, d) + r^2 \operatorname{Pol}(n, d-2)] = \\ = \dim \operatorname{Har}(n, d) + \dim r^2 \operatorname{Pol}(n, d-2).$$

Но $\dim r^2 \operatorname{Pol}(n, d-2) = \dim \operatorname{Pol}(n, d-2)$, т. к. умножение на r^2 является изоморфизмом

$$\operatorname{Pol}(n, d-2) \xrightarrow{\approx} r^2 \operatorname{Pol}(n, d-2).$$

Линейная алгебра также говорит нам, что для любого линейного отображения на векторном пространстве V размерность образа плюс размерность ядра равна размерности пространства. Здесь векторное пространство — $\operatorname{Pol}(n, d)$, ядро линейного отображения — подпространство гармонических полиномов и его образ $\operatorname{Pol}(n, d-2)$. Т. к.

$$\operatorname{Har}(n, d) + r^2 \operatorname{Pol}(n, d-2)$$

имеет ту же размерность, что и $\operatorname{Pol}(n, d)$, оно должно равняться $\operatorname{Pol}(n, d)$, что и доказывает теорему. ■

Симметрические полиномы

Пусть W — группа перестановок переменных. Если вас удивляет, почему мы используем букву W, то это из-за того, что в весьма экзотическом контексте Герман Вейль открыл нечто, называемое группой Вейля. Для обычной ситуации обычных полиномов эта группа Вейля является в точности, как мы сказали выше, группой перестановок переменных. Так что ничего страшного нет.

Если σ — перестановка из $\{1, \ldots, n\}$ и P — полином, то мы можем определить полином σP как полином

$$(\sigma P)(x_1, \ldots, x_n) = P(x_{\sigma(1)}, \ldots, x_{\sigma(n)}).$$

Поэтому W тоже действует на полиномах. Полином P такой, что $\sigma P = P$ для всех перестановок σ называется *симметрическим.* Заметим, что полином

$$r^2 = x_1^2 + \ldots + x_n^2$$

симметрический, и также полином $x_1^d + \ldots + x_n^d$ для любого положительного целого числа d. Обозначим через S^W пространство всех симметрических полиномов. Если P, Q симметрические, то такими же будут $P+Q$, PQ, и для любой константы $c \in \mathbb{R}$ полином cP тоже симметрический. Таким образом, множество симметрических полиномов замкнуто по сложению, умножению и умножению на константу. В действительности S^W содержит полиномы, равные константам.

Нас интересует другой пример симметрических полиномов. Пусть T — новая переменная, и мы рассмотрим полином

$$F(T) = (T - x_1)\ldots(T - x_n).$$

Если мы перемножим и приведем подобные, то получим

$$F(T) = T^n - I_1 T^{n-1} + I_2 T^{n-2} - \ldots + (-1)^n I_n,$$

где $I_1, \ldots, I_n$ — полиномы от $x_1, \ldots, x_n$. К примеру,

$$I_1 = I_1(x_1, \ldots, x_n) = x_1 + \ldots + x_n,$$

$$I_2 = I_2(x_1, \ldots, x_n) = x_1x_2 + x_1x_3 + \ldots + x_{n-1}x_n,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$I_n = I_n(x_1, \ldots, x_n) = x_1 \ldots x_n.$$

Явное вычисление полиномов $I_1, \ldots, I_n$ нам не понадобится. Можно увидеть априорно, что они симметрические в переменных $x_1, \ldots, x_n$, поскольку произведение

$$(T - x_1)\ldots(T - x_n)$$

не меняется от перестановки множителей $(T - x_i)$, $i = 1, \ldots, n$, поэтому коэффициенты этого полинома также не меняются. Заметим, что

$$\deg I_d = d.$$

Будем называть полиномы $I_1, \ldots, I_n$ *элементарными симметрическими полиномами.* Фундаментальный факт из алгебры дает нам

Теорема 4. *Элементарные симметрические полиномы порождают все симметрические полиномы. Более точно для данного симметрического полинома $P(x_1, \dots, x_n)$ существует полином F такой, что*

$$P(x_1, \dots, x_n) = F(I_1, \dots, I_n)$$

или, используя другие обозначения,

$$S^W = \mathbb{R}[I_1, \dots, I_n].$$

Доказательство не слишком сложное, и вы сможете найти его в некоторых учебниках для вузов, к примеру в [La 87], глава IV, теорема 8.1. Доказательство ведется индукцией по n, а также индукцией по d. Просто для развлечения докажите для себя случай $n = 2$, который показывает, что $x_1 + x_2$ и $x_1 x_2$ порождают все симметрические полиномы от двух переменных.

Мы хотим понять, что нужно для того, чтобы получить все полиномы из симметрических. Воспользуемся терминологией из линейной алгебры и векторных пространств, хотя S^W не замкнуто по вычислению частных. Пусть $P_1, \dots, P_N$ — элементы S. Будем говорить, что $P_1, \dots, P_N$ *линейно независимы* над S^W, если для выполнения соотношения

$$\sum_{j=1}^{N} F_j P_j = 0, \qquad F_j \in S^W$$

необходимо, чтобы $F_j = 0$ для всех j. Скажем, что $\{P_1, \dots, P_N\}$ — *базис* S над S^W, если $P_1, \dots, P_N$ линейно независимы над S^W и если каждый элемент S является линейной комбинацией $P_1, \dots, P_N$ с коэффициентами из S^W, другими словами, для данного $Q \in S$ существуют $F_1, \dots, F_N \in S^W$, такие, что

$$Q = F_1 P_1 + \dots + F_N P_N.$$

Мы утверждаем, что существует базис, и мы хотим явно представить такой базис. В действительности, мы представим два типа базисов, один более интересный, чем другой.

Существует один очень простой базис, который восходит к Кроннекеру и XIX столетию. Простым упражнением будет являться демонстрация того, что мономы

$$x_1^{r_1} \dots x_n^{r_n}, \qquad 0 \leqslant r_i \leqslant n - i$$

образуют базис. Заметим, что существует в точности $n!$ таких мономов. Мы не будем приводить доказательство. Давайте перейдем к более сложному и интересному базису, который связан с гармоническими функциями и имеет приложения в нескольких разделах математики, включая и алгебру и анализ.

Пусть S_+^W — подпространство симметрических полиномов с нулевым постоянным членом. Определим полином H — W*-гармонический* тогда и только тогда, когда

$$Q(\partial)H = 0 \quad \text{для всех} \quad Q \in S_+^W.$$

Заметим, что простой гармонический полином должен удовлетворять только одному уравнению $\Delta P = 0$. Здесь для W-гармонических полиномов мы требуем бесконечное количество уравнений, включающих все $Q \in S_+^W$. Можно выбраться из этой ситуации с конечным количеством уравнений, воспользовавшись генераторами $I_1, \dots, I_n$, но все же будет больше, чем одно уравнение, если $n \geqslant 2$.

Пусть:

Har_W — векторное пространство из W-гармонических полиномов;

$\mathrm{Har}_W^{(d)}$ — векторное пространство однородных элементов степени d в Har_W.

Тогда получаем разложение в прямую сумму

$$\mathrm{Har}_W = \sum_{d=0}^{\infty} \mathrm{Har}_W^{(d)}.$$

Другими словами, полином W-гармонический тогда и только тогда, когда все его однородные компоненты W-гармонические, и эти однородные компоненты взаимно ортогональны. Это сразу же следует из факта, что однородные полиномы различных степеней ортогональны, как мы уже видели.

Обозначим $S^W \mathrm{Har}_W$ множество всех полиномов вида

$$\sum_{j=1}^{m} P_j H_j, \qquad P_j \in S^W, \quad H_j \in \mathrm{Har}_W.$$

Следующий результат был доказан Шевалье в намного более общем контексте, чем то, что мы обсуждаем сейчас [Ch 55].

Теорема 5. *Пусть $S = R[x_1, \ldots, x_n]$ — кольцо полиномов над вещественными числами и S^W — подмножество симметрических полиномов, а именно те полиномы, которые не меняются при перестановках переменных. Пусть Har_W — векторное пространство W-гармонических полиномов. Тогда*

$$S = S^W \mathrm{Har}_W.$$

Более того, пространство Har_W конечномерно над $\mathbb{R}$, и его размерность

$$\dim_{\mathbb{R}} \mathrm{Har}_W = \#W.^2$$

Если $\{H_1, \ldots, H_N\}$ $(N = \#W)$ — базис Har_W над $\mathbb{R}$, тогда это также базис S над S^W.

Таким образом, мы получили очень ясную картину того, каким способом все полиномы порождаются линейно в терминах симметрических полиномов и W-гармонических полиномов. Доказательство будет проходить в несколько этапов.

Будем использовать обозначение

$$S_+^W S = S S_+^W,$$

имея в виду множество элементов

$$\sum_{j=1}^{m} P_j Q_j, \qquad P_j \in S_+^W, \quad Q_j \in S.$$

Заметим, что $S S_+^W S = S_+^W S S = S_+^W S$.

Прежде всего покажем, что мы можем разложить S в прямую сумму

$$S = S_+^W S + \mathrm{Har}_W. \tag{1}$$

[2] $\#W$ — количество перестановок. — *Прим. пер.*

Прямая сумма означает, что Har_W — ортогональное дополнение к $S_+^W S$. Вспомним, что если U — подпространство векторного пространства V, то ортогональное дополнение $U^\perp$ — множество $v \in V$ таких, что $\langle u, v\rangle = 0$ для всех $u \in U$. Поэтому (1) утверждает, что $\mathrm{Har}_W = = (S_+^W S)^\perp$. Это легко доказывается следующим образом. Пусть P — полином ортогональный к $S_+^W S$. Для $Q \in S_+^W$ получим

$$0 = \langle QS, P\rangle = \langle S, Q(\partial)P\rangle,$$

из чего следует, что $Q(\partial)P = 0$, следовательно $P \in \mathrm{Har}_W$. Поэтому

$$(S^W S)^\perp \subset \mathrm{Har}_W.$$

Обратное включение следует из обратимости шагов доказательства, поэтому (1) доказано. Т. к. однородные полиномы различных степеней ортогональны, мы получим однородную версию, для каждого положительного целого d, а именно, разложение в прямую сумму

$$S^{(d)} = \sum_{r=1}^{d} (S^W)^{(r)} S^{(d-r)} + \mathrm{Har}_W^{(d)}. \qquad (1_d)$$

Теперь мы можем повторять по индукции, чтобы получить первое утверждение теоремы, а именно

$$S = S^W \mathrm{Har}_W.$$

Затем нам понадобятся две леммы о $S_+^W S$, следуя Шевалье. Мы будем использовать стандартные обозначения. Если $Q_1, \ldots, Q_m$ — элементы S^W, определим идеал $(Q_1, \ldots, Q_m)$ в S^W как множество всех линейных комбинаций

$$P_1 Q_1 + \ldots + P_m Q_m, \qquad P_j \in S^W.$$

Лемма 5. *Пусть $Q_1, \ldots, Q_m \in S^W$ такие, что Q_1 не принадлежат идеалу $(Q_2, \ldots, Q_m)$ в S^W. Пусть $P_1, \ldots, P_m$ — однородные элементы, S такие, что $\sum_{\nu=1}^{m} P_\nu Q_\nu = 0$. Тогда $P_1 \in S_+^W S$.*

Доказательство.

В доказательстве мы будем использовать некоторые фундаментальные факты из алгебры. Если вы не знакомы с ними, то воспользуйтесь приложением к этому параграфу, где мы их обсудим и докажем.

Во-первых, мы воспользуемся *оператором проектирования* на W-инвариантах, а именно *оператором* A (от *усреднения*), определенного как

$$A(P) = \frac{1}{\#W} \sum_{w \in W} w(P).$$

Применим этот оператор к линейной комбинации $\sum P_\nu Q_\nu = 0$. Поскольку

$$w(PQ) = w(P)w(Q) \quad \text{и} \quad w(Q_\nu) = Q_\nu, \quad \text{т.к.} \quad Q_\nu \in S^W$$

получим

$$A(P_1)Q_1 + \ldots A(P_m)Q_m = 0.$$

Если $\deg P_1 = 0$, то P_1 — константа, тогда $P_1 = A(P_1)$, противоречие с гипотезой. Следовательно $\deg P_1 > 0$. Предположим, что лемма верна по индукции для всех отношений

$$\sum_{\nu=1}^{m} R_\nu Q_\nu = 0$$

с однородными R_ν и $\deg R_1 < \deg P_1$. Пусть τ — перестановка двух переменных, скажем, x_j и x_k. Тогда $\tau(P_\nu) - P_\nu$ делится на $x_j - x_k$, т.е.

$$\tau(P_\nu) - P_\nu = (x_j - x_k)R_\nu$$

для некоторого однородного полинома R_ν степени меньшей, чем $\deg P_\nu$ (см. приложение), и выполнено линейное отношение

$$R_1Q_1 + \ldots + R_mQ_m = 0.$$

По индукции $R_1 \in S_+^W S$ или, другими словами, $\tau(P_1) \equiv P_1 \mod S_+^W S$. Поскольку W образуется с помощью перестановок, следовательно,

$$w(P_1) \equiv P_1 \mod S_+^W S, \qquad w \in W.$$

Следовательно, $A(P_1) \equiv P_1 \mod S_+^W S$. Т.к. P_1 — однородный, строго положительной степени, то же самое верно для $A(P_1)$. Следовательно, $A(P_1) \in S_+^W S$, следовательно, наконец, $P_1 \in S_+^W S$, что и доказывает лемму. ■

Лемма 6. *Пусть $H_1, \ldots, H_m$ — однородные элементы S, линейно независимые по* $\mod S_+^W S$ *над константами. Тогда $H_1, \ldots, H_m$ линейно независимы над S^W.*

Доказательство.

Пусть $R_1H_1 + \ldots + R_mH_m = 0$ отношение линейной зависимости с $R_i \in S^W$, не все из них равны 0. Без потери общности мы можем предположить, что ни один из коэффициентов $R_1, \ldots, R_m$ не равен 0. Далее мы можем также предположить, что полиномы R_i однородные, и $\deg R_i + \deg H_i = d$ — константа, не зависящая от i.

Пусть $I_1, \ldots, I_n$ — элементарные симметрические полиномы. Мы уже упоминали, что $S^W = R[I_1, \ldots, I_n]$. Мономы $I_1^{j_1}, \ldots, I_n^{j_n}$ — элементы S, и $\deg I_1^{j_1}, \ldots, I_n^{j_n}$ относится к степени в S. Пусть $Q_0, Q_1, \ldots$ — упорядочение этих мономов по возрастанию степени, с $Q_0 = 1$. Для каждого $i = 1, \ldots, m$ имеем линейное выражение

$$R_i = \sum_{\nu \geqslant 0} c_{i\nu} Q_\nu$$

с постоянными коэффициентами c_i и $c_{i\nu} = 0$, если $\deg R_i \neq \neq \deg Q_\nu$. Приводя подобные, получаем

$$0 = \sum_{i=1}^m R_i H_i = \sum_{\nu \geqslant 0} P_\nu Q_\nu, \qquad P_\nu = \sum_{i=1}^m c_{i\nu} H_i,$$

и P_ν — однородные степени $d - \deg Q_\nu$. Мы доказали, что $P_\nu = 0$ для всех ν по индукции. Для $\nu = 0$ заметим, что $Q_0 \notin (Q_1, Q_2, \ldots)$, поэтому по лемме 1 $P_0 \in S_+^W S$, что противоречит предположению о линейной независимости $H_1, \ldots, H_m$. По индукции предположим, что $P_0, \ldots, P_{r-1} = 0$. Мы можем затем привести в точности те же аргументы, поскольку $Q_r \notin (Q_{r+1}, Q_{r+2}, \ldots)$. Это завершает доказательство леммы 2. ∎

Лемма 7. *$\mathbb{R}$-векторное пространство* Har_W *конечномерно, и в действительности его размерность $\leqslant \#W$, т. е. $\leqslant n!$.*

Доказательство.

Длина доказательства будет зависеть от того, насколько сильными фактами из алгебры вы собираетесь воспользоваться. Если хочется сохранить доказательство по возможности элементарным, то, возможно, наиболее быстрым является следующий способ доказательства. Мы воспользуемся уже упомянутым фактом, что $n!$ мономов вида

$$M_{(r)}(x) = x_1^{r_1} \dots x_n^{r_n}, \qquad 0 \leqslant r_i \leqslant n - i$$

образуют базис S над S^W. Также, как с векторными пространствами над полями, обычным способом можно показать, что два базиса S над S^W имеют одинаковое количество элементов. Далее, если элементы S — $H_1, \dots, H_m$ — линейно независимые над S^W, то $m \leqslant n!$, т. е. m не больше размерности S над S^W.

В частности, пусть $H_1, \dots, H_m$ — однородные элементы Har_W линейно независимые над $\mathbb{R}$. Согласно формуле (1), они линейно независимы $\bmod S_+^W S$ над константами, т. е. они удовлетворяют гипотезе леммы 2. По лемме 2, они линейно независимы над S^W. По предыдущим замечаниям мы приходим к заключению, что $m \leqslant \#W$, поэтому Har_W — конечномерно над $\mathbb{R}$, и его размерность меньше либо равна $\#W$. Это доказывает лемму 3, но посмотрите комментарии по теории полей в приложении к этому параграфу. ∎

Сейчас мы можем завершить доказательство теоремы 2. Пусть $H_1, \dots, H_m$ — базис Har_W над $\mathbb{R}$. Первое утверждение теоремы 2, которое мы доказали,

$$S = S^W \mathrm{Har}_W$$

показывает, что

$$S = S^W H_1 + \dots + S^W H_m,$$

и поэтому $H_1, \dots, H_m$ — линейные образующие S над S^W. Лемма 2 показывает, что $H_1, \dots, H_m$ также линейно независимы над S^W, поэтому они образуют базис S над S^W. Как уже упоминалось в доказательстве леммы 3, мы тогда получаем $m = \#W$. Это завершает доказательство теоремы 2. ∎

Приложение

В предыдущем доказательстве теоремы 2 мы использовали некоторые простые алгебраические факты, которые мы сейчас собираемся обсудить в случае, если вы не встречались с ними раньше.

***W*-усредняющая проекция.** Вы должны были познакомиться с проекциями в курсе линейной алгебры. Пусть V — векторное пространство. Пусть $A\colon V \to V$ — линейное отображение. Пусть U — подпространство V. Мы говорим, что A — *проекция* на U, если образ A равен U и если A оставляет элементы U на своих местах, т.е. $Au = u$ для всех $u \in U$. Если V имеет положительно определенное скалярное произведение и является конечномерным и если U — подпространство, то всегда существует ортогональная проекция. Но при применении леммы 1 мы находились в другой ситуации. Векторное пространство — множество всех полиномов S. Подпространство S^W, где W — группа перестановок. Каждое w может рассматриваться как линейный автоморфизм S, также удовлетворяющий мультипликативному свойству

$$w(PQ) = w(P)w(Q) \quad \text{для всех} \quad P, Q \in S.$$

Линейные отображения S в самого себя образуют векторное пространство. Мы определяем

$$A = \frac{1}{\#W} \sum_{w \in W} w.$$

Тогда $AQ = Q$ для всех $Q \in W$, поскольку $w(Q) = Q$, поэтому сумма, применяемая к Q, дает $(\#W)Q$, и деленная на $\#W$ наконец дает само Q. Далее, для любого $P \in S$ элемент $A(P)$ находится в S^W, т.е. $A(P)$ — симметрический полином, поскольку если $w_1 \in W$, то отображение $w \to w_1 w$ является перестановкой W, поэтому

$$w_1(A(P)) = \frac{1}{\#W} w_1 \sum_{w \in W} w(P) = \frac{1}{\#W} \sum_{w \in W} w_1 w(P) = A(P),$$

поэтому A — проекция на S^W.

Свойство делимости. В доказательстве леммы 1 мы использовали следующий факт:

Пусть $P \in S$ — полином от n переменных $x_1, \ldots, x_n$ и пусть τ — перестановка двух переменных, скажем, τ меняет местами x_j и x_k. Тогда $\tau(P) - P$ делится на $x_j - x_k$.

Мы приведем полное доказательство, которое является очень простым. Достаточно доказать утверждение, когда P — моном. После перенумерования переменных мы можем предположить без потери общности, что $j = 1$ и $k = 2$, поэтому моном может быть записан в виде

$$P(x_1, \dots, x_n) = x_1^{j_1} x_2^{j_2} Q(x_3, \dots, x_n).$$

Поскольку перестановка оставляет $x_3, \dots, x_n$ неподвижными, достаточно доказать утверждение для двух переменных, скажем для x и y, т.е. достаточно доказать:

Пусть x, y — две переменные. Пусть i, k — целые числа, бóльшие либо равные 0. Тогда

$$x^j y^k - x^k y^j$$

делится на $x - y$.

Доказательство.

Если j или $k = 0$, тогда обычное биномиальное разложение показывает, что $x^j - y^j$ и $x^k - y^k$ делятся на $x - y$. Предположим, что $j, k \geqslant 1$. Мы можем тогда разложить на множители

$$x^j y^k - x^k y^j = xy(x^{j-1} y^{k-1} - x^{k-1} y^{j-1}).$$

Затем мы можем завершить доказательство индукцией. ■

Идеалы. Третьим понятием, которым мы воспользовались, были идеалы. Пусть J — подмножество S. В лемме 1 мы использовали $J = S_+^W S$. Будем говорить, что J — идеал, если J замкнуто по сложению и умножению на элементы S. Поэтому $S_+^W S$ — идеал. Пусть $P, Q \in S$. Определим, что P *сравнимо* с Q по $\bmod J$, в символьной записи:

$$P \equiv Q \bmod J,$$

что означает, что $P - Q \in J$. Сразу получаем, что сравнение по $\bmod J$ — отношение эквивалентности, т.е. для всех P, Q, $R \in S$

$$\begin{aligned} &P \equiv P \quad \bmod J, \\ &P \equiv Q \quad \text{влечет } Q \equiv P, \\ &P \equiv Q \quad \text{и } Q \equiv R \text{ влечет } P \equiv R. \end{aligned}$$

Более того, если $P \equiv Q \bmod J$, то $RP \equiv RQ \bmod J$.

Пусть $H_1, \ldots, H_m \in S$. Мы говорим, что $H_1, \ldots, H_m$ *линейно независимы по* $\mathrm{mod} J$, имея в виду, что отношение

$$c_1H_1 + \ldots c_mH_m \equiv 0 \bmod J, \qquad c_i \in \mathbb{R}$$

влечет, что $c_i = 0$ для всех $i = 1, \ldots, m$.

Заметим, что линейная зависимость использовалась по отношению к двум типам коэффициентов: линейная зависимость над скалярами из $\mathbb{R}$ и линейная зависимость элементов S над S^W. Контекст делает ясным, какое понятие используется. В лемме 2 мы показали, что когда $J = S^W_+ S$, то одно их этих условий следует из другого.

Замечания по теории полей. Использование мономов $x_1^{r_1} \ldots x_n^{r_n}$ слишком специальный случай, чтобы довольствоваться им. Вместо этого можно использовать аргументы из теории полей, применяя их к очень общей ситуации, как в последующем изложении. Каждый элемент P из S является корнем полинома степени $\#W$ с коэффициентами в S^W, а именно, полинома

$$F(T) = \prod_{w \in W} (T - wP).$$

Действительно, этот полином содержит $T - P$ как множитель (используя $w = 1$), поэтому $F(P) = 0$. Пусть $c_1, \ldots, c_n$ — различные скаляры, и пусть

$$Y = c_1x_1 + \ldots + c_nx_n.$$

Тогда все элементы wY (с $w \in W$) различны и существует $\#W$ таких элементов. Пусть $\mathrm{Rat} = \mathbb{R}(x_1, \ldots, x_n)$ — поле рациональных функций от $x_1, \ldots, x_n$ (частные полиномов) и $K = \mathbb{R}(I_1, \ldots, I_N)$ — рациональные функции от $I_1, \ldots, I_N$. Тогда Rat — векторное пространство над K и можно применить обычную линейную алгебру. Довольно просто показать, что размерность Rat над K не больше $\#W$ из-за полиномиального уравнения $F(Y) = 0$, упомянутого выше. Дальнейшие аргументы из теории полей показывают, что эта размерность должна в точности равняться $\#W$. Если вы желаете посмотреть детальный вывод, то смотрите мою книгу «Алгебра для вузов» [La 87], глава VII, теорема 4.9 (теорема Артина). Представленная ситуация с симметрическими полиномами дает пример этой теоремы.

Собственные функции и характеры

В этом параграфе мы будем иметь дело с дифференциальными операторами и получим примеры собственных функций для них. Мы по-прежнему будем использовать переменные $x_1, \dots, x_n$ и частные производные

$$\partial_i = \frac{\partial}{\partial x_i}, \qquad i = 1, \dots, n.$$

Если $P \in S = \mathbb{R}[x_1, \dots, x_n]$, тогда, как и раньше, мы можем образовать дифференциальный оператор

$$P(\partial) = P(\partial_1, \dots, \partial_n) \in \mathbb{R}[\partial_1, \dots, \partial_n].$$

Иногда $P(\partial)$ называют *дифференциальным оператором с постоянными коэффициентами.* Определим W*-гармоническую функцию* f как бесконечно дифференцируемую функцию на $\mathbb{R}^n$, такую, что $P(\partial)f = 0$ для всех $P \in S_+^W$.

Теорема 6. *W-гармонические функции — это полиномиальные W-гармонические функции, т. е. просто элементы* Har_W.

Доказательство.

Пусть $I_1, \dots, I_n$, как и раньше, элементарные симметрические полиномы от переменных $x_1, \dots, x_n$. Тогда для каждого i, x_i — корень симметрического полинома

$$0 = x_i^n - I_1 x_i^{n-1} + \ldots + (-1)^n I_n.$$

Следовательно, $x_i^n \in S_+^W S$ для $i = 1, \dots, n$, и поэтому

$$\partial_i^n \in S(\partial) S_+^W(\partial),$$

таким образом, ∂_i^n обнуляет любую W-гармоническую функцию. Поэтому, если f — W-гармоническая функция, то $\partial_i^n f = 0$ для всех i, следовательно, f — полином. ■

Пусть V — векторное пространство над $\mathbb{R}$. Пусть L: $V \to V$ — линейное отображение. *Собственным вектором* v для L мы будем называть элемент V, для которого существует число $\lambda \in \mathbb{R}$ такое, что $L(v) = \lambda v$. В наших приложениях V состоит из функций, и поэтому вместо названия собственный вектор будем использовать название *собственная функция* для L.

В частности, мы рассматриваем S как векторное пространство над $\mathbb{R}$. Каждый полиномиальный дифференциальный оператор $P(\partial)$ с $P \in S$ тогда является линейным отображением S в себя. Мы хотим анализировать собственные функции (собственные полиномы) для всех таких операторов $P(\partial)$. По определению пусть $S(\partial)$ — множество всех полиномиальных дифференциальных операторов $P(\partial)$, $P \in S$. *Собственной функцией* f для $S(\partial)$ мы будем называть ненулевую функцию f, такую, что для каждого $D \in S(\partial)$ существует вещественное число $\lambda(D)$, такое, что

$$Df = \lambda(D)f.$$

Заметим, что для такой функции f отображение $D \to \lambda(D)$ — линейное отображение из $S(\partial)$ в $\mathbb{R}$, которое удовлетворяет мультипликативному свойству

$$\lambda(D_1D_2) = \lambda(D_1)\lambda(D_2).$$

Более обобщенно, линейное отображение λ: $S \to \mathbb{R}$, которое удовлетворяет мультипликативному правилу

$$\lambda(PQ) = \lambda(P)\lambda(Q) \quad \text{для всех} \quad P, Q \in S$$

будет называться *характером* S. Аналогично, линейное отображение

$$\lambda\colon S(\partial) \to \mathbb{R},$$

удовлетворяющее мультипликативному правилу $\lambda(D_1D_2) = \lambda(D_1)\lambda(D_2)$, будет называться *характером* $S(\partial)$.

Примеры характеров. Для каждой n-ки $s = (s_1, \ldots, s_n)$ вещественных чисел, отображение подстановки

$$P \mapsto P(s) \quad \text{или} \quad P(x_1, \ldots, x_n) \mapsto P(s_1, \ldots, s_n)$$

является характером S, который мы обозначим λ_s. Других характеров не существует, поскольку если $\lambda\colon S \to \mathbb{R}$ — характер, обозначим $s_i = \lambda(x_i)$ для $i = 1, \ldots, n$. Тогда по свойствам линейности и мультипликативности характеров мы получим

$$\lambda = \lambda_s.$$

Примеры собственных функций. Пусть снова $s = (s_1, \ldots, s_n)$ — n-ка вещественных чисел. Определим функцию

$$f_s\colon \mathbb{R}^n \to \mathbb{R}, \qquad f_s(x_1, \ldots, x_n) = e^{s_1 x_1} \ldots e^{s_n x_n}.$$

Тогда

$$(\partial_i f_s)(x) = s_i f_s(x) \quad \text{или} \quad \partial_i f_s = s_i f_s.$$

Следовательно, для любого полинома $P \in S$ получим

$$(P(\partial) f_s)(x) = P(s) f_s(x).$$

Следовательно, f_s — собственная функция S с собственным характером λ_s.

Интересной задачей является нахождение всех собственных функций S^W. Эта задача была решена в очень общем контексте Штейнбергом (Steinberg) [St 64]. Мы приведем решение в частном случае, поскольку общий случай потребует более обширного инструментария. Но основная идея доказательства та же самая и в частном случае и в общем случае.

Пусть λ — характер S. Пусть $w \in W$. Тогда мы можем определить действие w на λ, то есть $w\lambda$ по формуле

$$(w\lambda)(P) = \lambda(w^{-1}P).$$

Из этого следует, что $w\lambda$ также характер S. Если $\lambda = \lambda_s$, то λ отвечает n-ке $s = (s_1, \ldots, s_n)$, тогда $w\lambda_s$ отвечает перестановке чисел $(s_1, \ldots, s_n)$, что в действительности соответствует

$$(s_{w(1)}, \ldots, s_{w(n)}).$$

Нам потребуется частный случай теоремы, доказанной в очень общем случае Е. Артином.

Теорема. *Пусть $\lambda_1, \ldots, \lambda_m$ — различные ненулевые характеры S. Тогда они линейно независимы над $\mathbb{R}$.*

Доказательство.

Мы приведем доказательство Артина, верное в намного более общем случае. Доказательство индукцией по m. Предположим что $m = 1$. Тогда не может быть соотношения

$$c_1\lambda_1 = 0, \quad c_1 \in \mathbb{R}, \quad c_1 \neq 0,$$

иначе, пусть P такое, что $\lambda_1(P) \neq 0$. Тогда $c_1\lambda_1(P) = 0$ влечет

$$\lambda_1(P) = 0,$$

противоречие. Теперь предположим, что теорема верна по индукции для $m-1$ различных ненулевых характеров. Предположим, что мы имеем соотношение

$$c_1\lambda_1 + \ldots + c_m\lambda_m = 0, \tag{1}$$

где не все коэффициенты c_j равны 0. Если один коэффициент равен 0, то мы имеем соотношение между не более чем $m-1$ различными ненулевыми характерами, что невозможно по индукции. Поэтому $c_j \neq 0$ для $j = 1, \ldots, m$ и $m \geqslant 2$. По предположению, существует элемент $P \in S$ такой, что $\lambda_1(P) \neq \lambda_2(P)$. Затем мы используем школьное правило уничтожения слагаемых, чтобы получить соотношение между менее чем m характерами.

По формуле (1) имеем

$$c_1\lambda_1(PQ) + \ldots + c_m\lambda_m(PQ) = 0 \quad \text{для всех} \quad Q \in S.$$

По мультипликативному свойству получаем

$$c_1\lambda_1(P)\lambda_1(Q) + c_2\lambda_2(P)\lambda_2(Q) + \ldots + c_m\lambda_m(P)\lambda_m(Q) = 0 \tag{2}$$

для всех $Q \in S$. Умножаем (1) на $\lambda_1(P)$ и вычисляем в Q, чтобы получить

$$c_1\lambda_1(P)\lambda_1(Q) + c_2\lambda_1(P)\lambda_2(Q) + \ldots + c_m\lambda_1(P)\lambda_m(Q) = 0. \tag{3}$$

Вычитание (3) из (2) уничтожает первое слагаемое и приводит к

$$c_2(\lambda_2(P)-\lambda_1(P))\lambda_2(Q)+\ldots+c_m(\lambda_m(P)-\lambda_1(P))\lambda_m(Q)=0 \tag{4}$$

для всех $Q \in S$. Так как $\lambda_2(P)-\lambda_1(P) \neq 0$ и $c_2 \neq 0$, мы видим, что (4) является линейным соотношением с ненулевыми коэффициентами между не более чем $m-1$ различными ненулевыми характерами, что противоречит предположению индукции и завершает доказательство. ■

Пусть λ — характер S. Путь λ_W — его ограничение на S^W. Пусть

$$\mathrm{Ei}(S^W(\partial), \lambda_W)$$

множество аналитических функций в $\mathbb{R}^n$, которые являются собственными функциями $S^W(\partial)$ с характером λ_W или 0. Под *аналитической функцией* мы понимаем функцию, чей ряд Тейлора сходится к самой функции. Свойство собственной функции означает, что функция F удовлетворяет соотношению

$$P(\partial)F = \lambda_W(P)F \quad \text{для всех} \quad P \in S^W.$$

Заметьте, что $\mathrm{Ei}(S^W(\partial), \lambda_W)$ — векторное пространство над $\mathbb{R}$. Фактом является то, что такая аналитическая функция должна иметь вид Qf, где Q — полином и f — характер, но мы собираемся доказать этот факт другим способом. Назовем

$$\mathrm{Ei}(S^W(\partial), \lambda_W)$$

λ_W-собственным пространством $S^W(\partial)$. Штейнберг описал это собственное пространство в общем случае [St 64]. Мы опишем его и дадим доказательство в частном случае, где характер λ имеет вид λ_s с n-кой $s=(s_1, \ldots, s_n)$ такой, что числа $s_1, \ldots, s_n$ различны. В этом случае характер $\lambda=\lambda_s$ называется *регулярным*.

Теорема 7. *Пусть λ — регулярный характер. Тогда*

$$\dim \mathrm{Ei}(S^W(\partial), \lambda_W) = \#W.$$

Пусть $\lambda=\lambda_s$, и пусть $f=f_s$ — экспоненциальная функция

$$f(x)=e^{s_1x_1}\ldots e^{s_nx_n}.$$

Тогда функции wf, $w \in W$ образуют базис в собственном пространстве. Другими словами, функции, полученные из f перестановкой переменных, образуют базис в собственном пространстве.

Доказательство.

Т. к. координаты $s_1, \ldots, s_n$ различные, мы получаем, что функции wf, полученные из f перестановкой переменных, — различные, и они являются характерами, поэтому они линейно независимы по теореме Артина. Теперь мы должны доказать, что размерность $\mathrm{Ei}(S^W(\partial), \lambda_W)$ не больше $\#W$. Это завершит доказательство теоремы Штейнберга.

Пусть $\{H_1, \ldots, H_n\}$ — базис Har_W. Отобразим собственное пространство в $\mathbb{R}^n$, используя отображение

$$f \mapsto ((H_1(\partial)f)(0), \ldots, (H_m(\partial)f)(0)).$$

Отображение линейное, и достаточно доказать, что оно инъективно. Если образ f равен $(0, \ldots, 0)$, то $(H(\partial)f)(0) = 0$ для всех $H \in \mathrm{Har}_W$. По теореме 2, $S = S^W \mathrm{Har}_W$. Пусть $P \in S^W$, тогда для некоторого скаляра μ получаем

$$H(\partial)P(\partial)f = \mu H(\partial)f,$$

поэтому для всех $Q \in S$ — $(Q(\partial)f)(0) = 0$. Следовательно, ряд Тейлора функции f в 0 тождественно равен нулю. Т. к. f — вещественно аналитическая, то $f = 0$, из чего получаем инъективность и ограничение, равное $\#W$, на размерность собственного пространства. Это завершает доказательство теоремы 2. ■

Замечание. Доказательство Штейнберга в общем случае следует тому же шаблону, но вместо экспоненциальных функций в роли собственных функций приходится использовать те же функции умножения на определенные полиномы, которые можно явно вычислить. Обсуждение становится несколько более техническим, и сейчас самое время закончить.

Литература

[Ch 55] C. Chevalley, Invariants of finite groups generated reflections, *Amer. J. Math.* **77** (1955), pp. 778–782.

[La 87] S. Lang, *Undergraduate Algebra,* Springer-Verlag, 1987.

[La 93] S. Lang, *Algebra,* Third edition, Addition Wesley, 1993.

[St 64] R. Steinberg, Differential equations invariant under finite reflection groups, *Trans. Amer. Math. Soc.* **112** (1964), pp. 392–400.