УДК 51(06)
С88

Издание осуществлено при поддержке РФФИ
(издательский проект № 01-01-14108).

Студенческие чтения НМУ. Вып. 2 / Под общей редакцией
С88 В. Прасолова. — М.: МЦНМО, 2001. — 144 с.: ил.

ISBN 5-94057-009-7

В книге представлены лекции, прочитанные в Независимом московском университете в 1999–2000 г., предназначенные для широкой аудитории. Их цель — рассказать о некоторых областях математики и описать новые идеи.

Для студентов, аспирантов и преподавателей математических специальностей.

ISBN 5-94057-009-7

Майлз Рид

Кольца и алгебраические многообразия

Лекция 19 февраля 1999 года

Пусть k — алгебраически замкнутое поле (например, $k = \mathbb{C}$). Мы будем рассматривать только кольца вида $R = k[x_1, \ldots, x_n]/I$, где I — простой идеал. Иными словами, будем рассматривать конечнопорожденные алгебры без делителей нуля над алгебраически замкнутым полем.

Кольцу R можно сопоставить его *спектр* $\operatorname{Spec} R$ — совокупность всех простых идеалов. В рассматриваемой ситуации простые идеалы — это то же самое, что максимальные идеалы.

Спектр нас интересует не как множество, а как алгебраическое многообразие V в аффинном пространстве $\mathbb{A}^n$ (если $k = \mathbb{C}$, то $\mathbb{A}^n = \mathbb{C}^n$). Алгебраическое многообразие V — совокупность всех точек пространства $\mathbb{A}^n$, в которых обращаются в нуль все многочлены из идеала I. Отождествление $\operatorname{Spec} R$ и V соответствует выбору координат $x_1, \ldots, x_n$.

Наоборот, если задано алгебраическое многообразие V, то ему соответствует кольцо $k[V]$ — кольцо полиномиальных функций на V; это кольцо совпадает с R.

Любой математик слышал, что алгебраическое многообразие — это множество нулей каких-то полиномов. Здесь есть два подхода:

1) заданы уравнения и нужно выяснить, каково множество точек;

2) задан геометрический объект и его нужно задать как алгебраическое многообразие (погрузить в аффинное пространство и задать уравнениями).

Пример. Пусть в аффинном пространстве $\mathbb{C}^2$ действует конечная группа линейных преобразований $G \subset \mathrm{GL}(2, \mathbb{C})$. На факторпространстве $\mathbb{C}^2/G$ мы хотим задать структуру алгебраического многообразия.

Например, каково кольцо функций? Правдоподобный кандидат таков: в кольце многочленов $k[x, y]$ возьмем кольцо инвариантных функций $k[x, y]^G$. Это кольцо имеет вид $k[x, y]^G = k[u_1, \ldots, u_n]/I$. Действительно, мы выбираем инвариантный многочлен u_1, затем u_2; так делаем до тех пор,

Майлз Рид (Miles Reid), профессор Уорикского университета (Великобритания).

пока новых (независимых) многочленов больше не будет. Идеал I соответствует соотношениям между выбранными инвариантными многочленами.

Если группа G порождена отображением $x, y \mapsto -x, -y$, то все инвариантные многочлены выражаются через $x^2 = u$, $xy = v$, $y^2 = w$. Эти многочлены связаны соотношением $uw = v^2$. В результате возникает отображение $\mathbb{C}^2 \to Q \subset \mathbb{C}^3$, где Q — квадрика, заданная уравнением $uw = v^2$. Это отображение отождествляет все точки каждой орбиты и ничего другого не отождествляется.

Этот пример можно обобщить. Пусть $\varepsilon = \exp(2\pi i/r)$. Рассмотрим группу, порожденную отображением $x, y \mapsto \varepsilon x, \varepsilon^{-1} y$. В этом случае инвариантными многочленами будут $x^r = u$, $xy = v$, $y^r = w$. Они связаны соотношением $uw = v^r$. В результате возникает отображение $\mathbb{C}^2 \to X \subset \mathbb{C}^3$, где X задано уравнением $uw = v^r$. Это уравнение задает особенность типа A_{r-1}.

Можно также рассмотреть группу, порожденную отображением $x, y \mapsto$ $\mapsto \varepsilon x, \varepsilon^3 y$, где $\varepsilon^7 = 1$. В этом случае инвариантными многочленами будут $x^7 = u_1$, $x^4 y = u_2$, $xy^2 = u_3$, $y^7 = v$. Связывающие их соотношения можно записать в виде $\operatorname{rk}\begin{pmatrix} u_1 & u_2 & u_3^2 \\ u_2 & u_3 & v \end{pmatrix} \leqslant 1$.

Упражнение. Доказать, что указанные многочлены порождают все кольцо инвариантных многочленов, а указанные соотношения порождают весь идеал соотношений.

Теперь перейдем к определению проективного многообразия. В этом случае R — градуированное кольцо, т. е. $R = \bigoplus_{n \geqslant 0} R_n$, где $R_0 = k$ и для однородных полиномов предполагается, что $x_n y_m \subset R_{n+m}$. Будем также предполагать, что кольцо конечнопорожденное и без делителей нуля. Таким образом, $R = [x_1, \ldots, x_n]/I$, где образующая x_i имеет вес a_i, не обязательно равный 1, а I — простой однородный (относительно этих весов) идеал.

Градуированному кольцу R соответствует $\operatorname{Proj} R$ — совокупность однородных простых идеалов $P \subset R$, за исключением тривиального простого идеала $m_0 = \bigoplus_{n>0} R_n$. В рассматриваемой ситуации однородные простые идеалы — это то же самое, что максимальные однородные идеалы.

Проективное многообразие $X = \operatorname{Proj} R$ является объединением аффинных многообразий. А именно, $X = \bigcup X_f$, где аффинное многообразие X_f $(0 \notin X_f \subset R_n)$ имеет вид $X_f = \left(R\left[\frac{1}{f}\right]\right)_0 = \left\{\frac{g}{f^k}\right\}$; (индекс 0 означает, что берутся элементы степени 0) здесь $\deg g = nk$.

Если мы выберем координаты, то получится, что X вложено во взвешенное проективное пространство $\mathbb{P}(a_0, a_1, \ldots, a_n)$, которое определяется следующим образом. Введем на множестве $\mathbb{C}^{n+1} \setminus \{0\}$ отношение эквивалентности

$$(x_0, x_1, \ldots, x_n) \sim (\lambda^{a_0} x_0, \lambda^{a_1} x_1, \ldots, x_n \lambda^{a_n})$$

и рассмотрим факторпространство по этому отношению эквивалентности.

Если все a_i равны 1, то получаем обычное проективное пространство.

Пример. Взвешенное проективное пространство $\mathbb{P}(1, 2, 3)$ — это проективная плоскость $\mathbb{P}^2$ с двумя рогами (рис. 1). Эти рога являются факторособенностями $\frac{1}{2}(1, 1)$ и $\frac{1}{3}(1, 2)$. (Это именно те факторособенности, с которыми мы только что встречались.)

Отношение эквивалентности в пространстве $\mathbb{C}^3$ в данном случае имеет вид $(x, y, z) \sim (\lambda x, \lambda^2 y, \lambda^3 z)$. Здесь появляются исключительные орбиты. Рассмотрим, например, ограничение на ось y, т. е. рассмотрим орбиту точки $(0, 1, 0)$. При действии элемента $\lambda = -1$ точка $(0, 1, 0)$ остается неподвижной. Это означает, что у орбиты есть нетривиальный стабилизатор. А там, где увеличивается стабилизатор, появляются факторособенности.

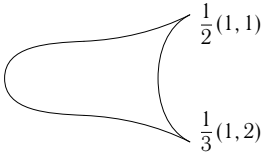


Рис. 1

Почему же факторособенность здесь равна половине? На рассматриваемой орбите единственная ненулевая функция равна (с точностью до пропорциональности) y. Поэтому в знаменателе будет только y:

$$\left(R\left[\frac{1}{f}\right]\right)_0 = \left(k[x, y, z]\frac{1}{y}\right)_0 = \left\{\frac{g}{y^k}\right\}.$$

Здесь y имеет степень 2, поэтому многочлен g имеет не произвольную степень, а лишь четную.

По-другому можно сказать так. Мы берем $\eta = \sqrt{y}$. Тогда локальными координатами в окрестности точки $(0, 1, 0)$ будут x/y и z/y^3. Это будут координаты не на самом многообразии, а на его циклическом накрытии.

Ситуация в окрестности точки $(0, 0, 1)$ рассматривается аналогично.

Взвешенное проективное пространство $\mathbb{P}(1, 2, 3)$ можно вложить в обычное проективное пространство, но большой размерности. А именно, его можно вложить в $\mathbb{P}^6$. Для этого нужно взять $k[x, y, z]^{[6]}$ — многочлены, степень которых делится на 6. Базисными полиномами будут

$$\begin{array}{cccc} x^6 & x^4y & x^2y^2 & y^3 \\ x^3z & xyz & & \\ x^2 & & & \end{array}$$

Девять соотношений среди этих мономов задают поверхность дель Пеццо $S_6 \subset \mathbb{P}^6$.

Применения

Кольцу $R = k[X]$ (все полиномиальные функции на X) сопоставляется аффинное алгебраическое многообразие $X = \operatorname{Spec} R$. Для проективных многообразий нужно зафиксировать обильный дивизор; тогда получим $X = \operatorname{Proj} R$. Но это сложно. Желающие могут обратиться ко второй главе Хартсхорна.

Мы будем рассматривать только один простой случай: E — эллиптическая кривая с заданной точкой $P \in E$. Эллиптическую кривую можно здесь понимать либо как неособую плоскую кубическую кривую, либо как компактную риманову поверхность рода 1. Задача заключается в том, чтобы погрузить ее в проективное пространство и найти задающее ее уравнение.

Пусть $\mathscr{L}(nP)$ — пространства, которые участвуют в теореме Римана—Роха, т. е. пространства глобальных мероморфных функций с полюсами только в точке P и кратности не выше n. Согласно теореме Римана—Роха

$$\dim \mathscr{L}(nP) = \begin{cases} n & \text{при } n > 0; \\ 1 & \text{при } n = 0. \end{cases}$$

Рассмотрим пространство $R(E, P) = \bigoplus_{n \geqslant 0} \mathscr{L}(nP)$. (Так можно поступить для произвольного алгебраического многообразия с заданным обильным дивизором.)

Пространства R_0 и R_1 соответствуют одним и тем же функциям. Действительно, если для эллиптической функции допускается только один простой (некратный) полюс, то такая эллиптическая функция постоянна. Поэтому в R_0 и R_1 входят только постоянные функции. Но мы рассматриваем прямую сумму, поэтому одна и та же функция соответствует разным элементам кольца $R(E, P)$, когда мы ее рассматриваем как элемент пространств R_0 и R_1. Для удобства введем обозначения $R_0 = k$ и $R_1 = k \cdot x$ (здесь x соответствует функции, тождественно равной 1, когда мы рассматриваем ее как элемент пространства R_1).

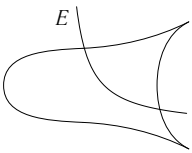


Рис. 2

Теорема. *$R(E, P) = k[x, y, z]/(f_6)$, где $\dim x = 1$, $\dim y = 2$ и $\dim z = 3$. Идеал (f_6) порожден однородным соотношением вида $z^2 + \ldots = y^3 + \ldots$ (Любое такое соотношение можно привести к виду $z^2 = y^3 + a_4(x)y + b_6(x)$.)*

Эта теорема означает, что E вкладывается в $\mathbb{P}(1, 2, 3)$ в качестве гиперплоского сечения. Член z^2 показывает, что при этом вложении эллиптическая кривая не пересекает особые точки взвешенного проективного пространства $\mathbb{P}(1, 2, 3)$ (рис. 2).

Доказательство теоремы.

Пространство $\mathscr{L}(0)$ порождено элементом 1;

пространство $\mathscr{L}(P)$ порождено элементом x;

пространство $\mathscr{L}(2P)$ порождено элементами x^2 и y;

пространство $\mathscr{L}(3P)$ порождено элементами x^3, xy и z.

Вплоть до этого момента количество мономов совпадало в размерностью пространства. Но в пространстве $\mathscr{L}(4P)$ мономов уже будет больше, чем размерность пространства. Поэтому возникает соотношение между мономами.

Линейная независимость мономов следует из того, что отображение $E \to \mathbb{P}^6$, заданное x^2 и y, не постоянное (т. е. x^2 и y алгебраически независимы). □

Процесс можно алгоритмизировать. Мы рассматриваем градуированное кольцо R, для которого $\dim R_n = l(n)$, где $l(n) = n$ при $n > 0$ и $l(0) = 1$. Градуированному кольцу можно сопоставить ряд Пуанкаре $L(t) = \sum l(n)t^n$. Для конечнопорожденного кольца ряд Пуанкаре — рациональная функция. Вид этой рациональной функции показывает размерности образующих и количество соотношений и их размерность. Например, в нашем случае

$$L(l) = \frac{1 - t^6}{(1 - t)(1 - t^2)(1 - t^3)}. \tag{1}$$

Это означает, что есть три образующие размерностей 1, 2 и 3, а также одно соотношение размерности 6.

Равенство (1) можно доказать, например, так. По определению $L(t) = 1 + t + 2t^2 + 3t^3 + \ldots$ Поэтому

$$(1 - t)L(t) = 1 + t^2 + t^3 + t^4 + \ldots,$$

а значит,

$$(1 - t^2)(1 - t)L(t) = 1 + t^3 = \frac{1 - t^6}{1 - t^3}.$$

Специалисты по коммутативной алгебре знают, что это весьма общая конструкция. Если есть градуированный модуль M и в кольце есть элемент v, который не является делителем нуля, то умножение на v индуцирует точную последовательность

$$0 \to M \xrightarrow{v} M/v.$$

Здесь M/v — модуль над меньшим кольцом. А именно, над кольцом, где подставляется $v = 0$. Умножение ряда, которые мы рассмотрели выше, в точности соответствует этой редукции.

А. Б. Каток

Биллиардный стол как игровая площадка для математика

Лекция 10 марта 1999 года

Название этой лекции можно понимать двояко. В буквальном, несколько шутливом смысле: математики резвятся, запуская биллиардные шары на столах различной формы и наблюдая (а также пытаясь предсказать) что получится. В более серьезном значении выражение «игровая площадка» следует понимать как «испытательный полигон»: различные вопросы, гипотезы, методы решения и т. д. в теории динамических систем «испытываются» на различных типах биллиардных задач. Я надеюсь убедительно продемонстрировать, что по крайней мере вторая интерпретация заслуживает серьезного внимания.

О биллиардах написано довольно много и в научных статьях, и в монографиях, и в учебниках, и в научно-популярной литературе. Короткие брошюры Г. А. Гальперина и А. Н. Землякова [4] и Г. А. Гальперина и Н. И. Чернова [5] написаны весьма доступно и освещают широкий круг вопросов. Введение в проблематику, связанную с биллиардами, для более подготовленного читателя содержится в главе 6 книги [9]. Следующий уровень представлен очень хорошо написанной книгой С. Табачникова [14], выход которой на русском языке, к сожалению, задерживается. Книга автора и Б. Хасселблатта [8] содержит достаточно детальное современное изложение теории выпуклых биллиардов и закручивающих отображений. Серьезное, но вполне доступное изложение современного состояния теории параболических биллиардов содержится в обзорной статье Х. Мезера и С. Табачникова, которая выйдет в свет (на английском языке) весной 2002 года [11]. Сборник [12] содержит богатый материал по гиперболическим биллиардам и смежным вопросам. Ссылки более специального характера даются по ходу изложения.

1. Эллиптические, параболические и гиперболические явления в динамике

Задача о движении биллиардного шара формулируется очень просто. Имеется замкнутая кривая $\Gamma \subset \mathbb{R}^2$. Внутри области D, ограниченной этой кривой, равномерно движется точка по отрезкам прямых, а когда точка встречается с кривой, она отражается от кривой по закону «угол падения равен углу отражения». Задача состоит в том, чтобы понять характер этого движения за большое время.

Мы имеем дело с динамической системой, которая, вообще говоря, не всюду определена. Например, если в области с кусочно гладкой границей точка попадает в угол, то непонятно, как продолжать траекторию. Есть и более тонкие эффекты: при некоторых начальных условиях возможна ситуация, когда за конечное время происходит бесконечное число соударений и движение не может быть продолжено. Но это эффекты патологические; можно говорить, что имеется динамическая система.

Решение задачи о движении шара зависит от области. Одна из причин, по которым интересна эта задача, заключается в том, что формальное описание движения очень просто и остается только содержательная часть. Вторая, более серьезная причина уже была упомянута. Она связана с тем, что если попытаться каким-то образом расклассифицировать задачи теории динамических систем, то, с некоторым огрублением, их можно разделить на эллиптические, параболические и гиперболические (рис. 1). Таким образом, биллиардный стол — это полигон, на котором можно испытывать методы, гипотезы, вопросы, возникающие в разных областях теории динамических систем.

Ничего нового в использовании этих слов для выражения некоторой трихотомии нет. Соответствующая классификация в теории дифференциальных уравнений в частных производных хорошо известна. Но для динамических систем подобная классификация систематически, по-видимому, не была проведена.

В случае биллиардов эллиптические эффекты возникают, например, для эллипса. Это совпадение не совсем случайное, однако оно непосредственно не распространяется на биллиарды внутри параболы и гиперболы. Более общая ситуация, в которой возникают эллиптические эффекты, такова: кривая гладкая (достаточно высокого класса гладкости), выпуклая, и ее кривизна нигде не обращается в нуль. Изучение бильярдной задачи внутри таких областей дает хорошее поле для демонстрации проблематики, а также результатов, связанных с эллиптическим поведением динамических систем.

В параболической ситуации область — обычный многоугольник. Для простоты можно даже взять прямоугольный треугольник, углы которого

отличны от 30° и 45°. Прямоугольный треугольник с углом $\pi/8$ уже дает пример динамической системы с параболическим поведением.

Гиперболическая ситуация хорошо представляется тремя примерами (см. рис. 1): квадратом с вырезанным кружком, «стадионом» и кардиоидой.

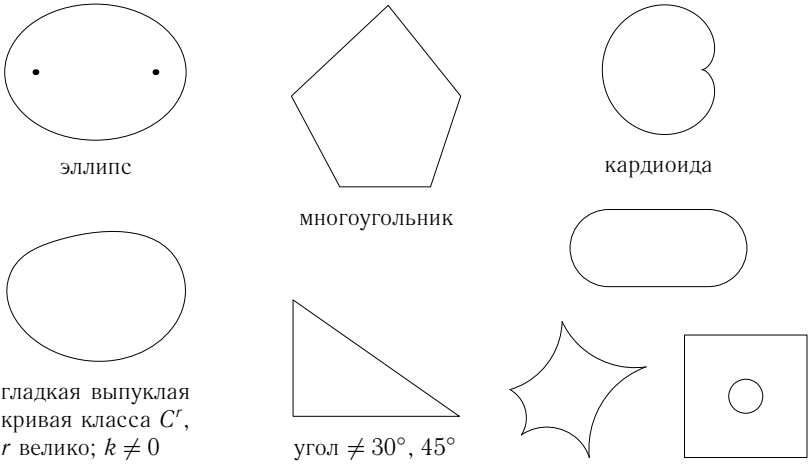


Рис. 1. Эллиптические, параболические и гиперболические бильярды

Идея, по крайней мере, о дихотомии, которая существует в теории динамических систем, за последние годы уже укоренилась. Одна из наиболее замечательных книг по теории динамических систем, написанных во второй половине XX века, — книга Юргена Мозера «Stable and random motion in dinamical systems». «Stable» — это эллиптические эффекты, «random» — гиперболические. Параболические эффекты в книге Мозера не обсуждаются.

Чтобы дать некоторое представление о характере трихотомии, которая здесь возникает, поясню, откуда взялись эти названия. Для линейных отображений соответствующая трихотомия хорошо известна. Для линейного отображения $L\colon \mathbb{R}^n \to \mathbb{R}^n$ возможны три основных вида поведения:

(1) Устойчивое поведение. Оно возникает тогда, когда все собственные значения λ_i по модулю равны 1 и нет нетривиальных жордановых клеток: $\operatorname{Sp} L \subset S^1$. В этой ситуации все орбиты возвращающиеся и устойчивые. Это эллиптическое поведение.

(2) По-прежнему $|\lambda_i| = 1$, но есть нетривиальные жордановы клетки. У жордановой клетки есть собственный вектор, поэтому есть и устойчивые орбиты. Однако в этой ситуации типично полиномиальное разбегание орбит. Это параболическое поведение.

(3) Гиперболическое поведение: $\mathrm{Sp}\, L \cap S^1 = \varnothing$. В этой ситуации любые две орбиты экспоненциально разбегаются либо в положительном, либо в отрицательном направлении.

Возможны также комбинации этих трех парадигм. Например, весьма важно то, что мы называем частично гиперболической ситуацией, когда в спектре есть гиперболическая компонента и что-то еще. В динамике это очень важная парадигма.

Было бы очень наивно пытаться строить концепцию нелинейной дифференциальной динамики на основании только этих трех моделей. Чем вообще занимается нелинейная дифференциальная динамика? Она занимается анализом асимптотического поведения гладких систем, для которых есть понятие локального инфинитезимального поведения системы, а с другой стороны, в силу компактности фазового пространства, есть феномен возвращения орбит сколь угодно близко к начальному положению. Грубо говоря, деление нелинейной динамики на эллиптическую, параболическую и гиперболическую соответствует ситуациям, когда линейное поведение, которое более или менее аппроксимируется этими тремя типами, сочетается с нетривиальным характером возвращения.

Такой подход игнорирует очень существенную часть проблематики теории динамических систем, например, такие вещи, как анализ систем Морса—Смейла или эффекты, связанные с бифуркациями. Речь идет о ситуациях, когда возвращение простое, а интересные феномены относятся, например, к тому, как фазовое пространство разделяется на бассейны притяжения к нескольким имеющимся притягивающим точкам или предельным циклам. Это все игнорируется. Мы сейчас говорим только о той части динамики, которая относится к рекуррентному поведению. Нерекуррентное поведение нами сейчас более или менее игнорируется.

Чтобы правильно проинтерпретировать интересующие нас феномены, нужно понять, что такое линеаризация динамической системы. Пусть задано отображение $f\colon M \to M$, действующее на фазовом пространстве. Мы считаем, что фазовое пространство является гладким объектом, поэтому можно говорить о действии на касательные векторы. Для любой точки $x \in M$ имеется линейное отображение $Df_x\colon T_xM \to T_{f(x)}M$. Такое отображение само по себе интересно только в случае неподвижной точки. Но в общем случае в динамике можно рассмотреть итерации Df_x^n. Введя риманову метрику, можно говорить об асимптотической скорости роста длины векторов. Риманова метрика вводится неоднозначно, но на компактном многообразии любые две метрики отличаются в пределах мультипликативной константы, поэтому скорость роста векторов определена корректно.

Эллиптическое поведение возникает тогда, когда в линеаризованной системе роста длины векторов либо совсем нет, либо он медленнее, чем линейный (сублинейный рост).

Жорданова клетка минимального порядка 2 соответствует уже линейному росту. Параболическое поведение — это субэкспоненциальный рост (обычно полиномиальный).

Гиперболическая парадигма понята лучше всего. Она соответствует ситуации, в которой система расщепляется и в некоторых направлениях происходит экспоненциальный рост, а в других направлениях экспоненциальное убывание. Когда время обращается, эти направления меняются местами.

Бывают смешанные ситуации. Например, можно взять прямое произведение двух систем разного типа. Но как метаутверждение можно сказать, что гиперболическая парадигма доминирующая: если есть нетривиальный гиперболический эффект и что-то еще, то обычно поведение системы может быть понято на основании ее гиперболической части. Это неверно в буквальном смысле. Например, это неверно для прямого произведения динамических систем. Но для типичной динамической системы гиперболическое поведение «забивает» все остальное.

Можно также сделать такое интересное замечание. Когда размерность фазового пространства мала, смешанное поведение невозможно. Например, когда размерность фазового пространства равна 2, частично гиперболическое поведение невозможно, потому что для гиперболического поведения нужно хотя бы одно расширяющее и хотя бы одно сжимающее направление. По той же самой причине в малых размерностях чаще встречается эллиптическое или параболическое поведение.

Самым чистым примером эллиптического поведения является ситуация, когда имеется гладкая изометрия. В этом случае никакого роста не происходит. Для гладких изометрий понять динамику довольно легко. Если на компактном фазовом пространстве есть гладкая изометрия, то фазовое пространство разбивается на инвариантные торы, и на каждом торе возникает параллельный перенос (или вращение, если применяются мультипликативные обозначения). В частности, если само многообразие не является тором, то такое движение не транзитивно.

Это, конечно, частный случай того, что хорошо известно в гамильтоновой механике, а именно, это соответствует вполне интегрируемым гамильтоновым системам. Это хороший пример взаимодействия парадигм, потому что если на вполне интегрируемую систему посмотреть наивно, то ее нужно отнести к параболической парадигме. Действительно, линейная часть вполне интегрируемой гамильтоновой системы параболическая, потому что в направлении, трансверсальном инвариантным торам, происходит закручивание. С другой стороны, пространство разбивается на инвариантные торы и на каждом торе анализ производится с помощью эллиптических методов.

Эта ситуация типична. Именно по этой причине эллиптическая парадигма важна. Довольно редко бывает, что глобальное поведение на всем фазовом пространстве характеризуется отсутствием роста. Но довольно часто внутри фазового пространства есть какие-то элементы, где поведение может быть описано с помощью эллиптической парадигмы.

Гиперболическая ситуация изучена лучше всего. В некотором смысле, это единственная универсальная парадигма сложного поведения в динамике. Она может быть хорошо понята с помощью цепей Маркова и простых стохастических моделей. С точки зрения приложений динамики, если гиперболическое поведение установлено, то дальше можно применять довольно мощный разработанный аппарат, который позволяет изучать поведение нелинейных систем. Все это возникает благодаря взаимодействию определенного поведения линеаризованной системы с более или менее a priori существующим возвращением. В линейных системах гиперболичность сопровождается просто разбеганием системы в бесконечность. Но если разбегаться некуда, если обязательно нужно возвращаться, то возникают упомянутые выше и хорошо понятые типы сложного поведения.

Параболическое поведение, в отличие от эллиптического и гиперболического, во-первых, неустойчиво, а во-вторых, оно характеризуется отсутствием стандартных моделей. В эллиптической ситуации есть универсальная модель — вращение на торе (или какие-то его следы), а в гиперболической ситуации есть марковская модель, которая все описывает. В параболической ситуации, по-видимому, нельзя даже сказать, что есть какой-то набор моделей, к которым все более или менее сводится. Тем не менее, есть довольно характерные явления, которые встречаются в конкретных классах систем. Одно из этих явлений состоит в том, что часто эффект умеренного растяжения можно подменить эффектом разрезания. Например, если есть система, которая локально выглядит как изометрия, но при этом испытывает разрывы, то такая система относится к параболической парадигме.

Хорошо известный пример — перекладывание отрезков. Мы разрезаем отрезок на части и перекладываем их в соответствии с заранее заданной перестановкой. Локально эта система выглядит как эллиптическая, но есть эффект разрезания. Довольно легко сообразить, что эту систему нужно рассматривать как параболическую: при итерациях число отрезков растет линейно. Эта линейность не является результатом закручивания, она является результатом разрезания. Но эффект примерно тот же самый.

Таким образом, параболическое поведение часто связано с существованием умеренных особенностей в системах. Не случайно на картинке для иллюстрации параболического поведения был нарисован биллиард в многоугольнике.

2. Биллиарды в гладких выпуклых областях

Джордж Д. Биркгоф был первым кто стал систематически рассматривать биллиарды как модели для задач классической механики. Биркгоф рассматривал биллиарды только в гладких выпуклых областях. Он, конечно, не думал о биллиардах в многоугольниках, а тем более, в невыпуклых областях.

Прежде всего можно провести редукцию к биллиардному отображению. Первоначальная динамическая система для биллиарда — это система с непрерывным временем. Но то, что происходит внутри биллиардного стола, легко восстановить, зная то, что происходит в моменты отражения. Поэтому достаточно рассмотреть так называемое биллиардное отображение. Фазовое пространство биллиардного отображения устроено следующим образом. Выходящий после отражения вектор v характеризуется циклической координатой $\varphi \in S^1$, задающей положение точки на кривой Γ, и углом $\theta \in [0, 2\pi)$ между касательным вектором и вектором v (рис. 2).

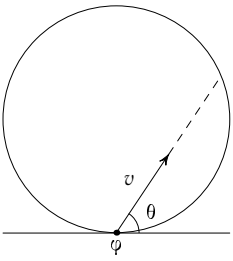


Рис. 2. Координаты вектора после отражения

Фазовое пространство биллиардного отображения представляет собой цилиндр. После отражения мы получаем новую точку φ_1 и новый выходящий вектор, которому соответствует угол θ_1. Отображение $T(\varphi_0, \theta_0) = (\varphi_1, \theta_1)$ и является биллиардным отображением. Оно отображает открытый цилиндр в себя; по непрерывности это отображение можно продолжить на замкнутый цилиндр. Точки, для которых $\theta = 0$, неподвижные (мы предполагаем, что кривая Γ не содержит отрезков).

Упражнение. Покажите что наличие прямолинейных отрезков у границы стола влечет разрывность биллиардного отображения.

Упражнение. Найдите условия при которых биллиардное отображение дифференцируемо (один раз или бесконечно много раз) на границе цилиндра.

Биллиардное отображение обладает двумя важными качественными свойствами.

1) Сохранение площади. Сохраняется элемент площади

$$dA = \sin\theta\, d\theta\, d\varphi = d\alpha\, d\varphi, \quad \text{где} \quad \alpha = \cos\theta.$$

(Чтобы ввести координаты, в которых сохраняется площадь, нужно вместо θ взять $\cos\theta = \alpha$).

2) Закручивание. Фиксируем координату φ_0 и будем изменять координату θ. Координата φ образа будет монотонно изменяться, пока она не

обойдет всю окружность и вернется обратно (рис. 3). Образ вертикали закручивается.

Эти два свойства позволяют сделать вывод о наличие эллиптического поведения. Проблематика, которая возникает в связи с эллиптическим поведением, разбивается на две части:

1) каустики,

2) орбиты Биркгофа и множества Обри—Мазера.

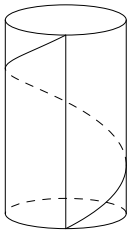


Рис. 3. Закручивание

Я начну со второй части. Мы хотим найти периодические орбиты биллиардной системы. Периодические орбиты бывают разные. Отличаются они не только периодом, но и некоторой комбинаторикой. Например, две орбиты периода 5 на рис. 4 имеют разную комбинаторику. В первом случае делается один оборот, а во втором два. Эти орбиты регулярные: порядок точек на орбите сохраняется; он такой же, как при вращении. Именно такие (регулярные) орбиты называют орбитами Биркгофа. Это название связано с тем, что он доказал замечательную и сравнительно простую теорему о существовании регулярных орбит. По-видимому, именно с этой теоремы началось применение вариационных методов в динамике.

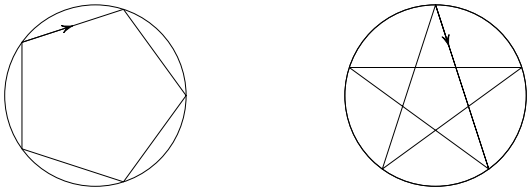


Рис. 4. Орбиты одинакового периода и разной комбинаторики

Теорема (Биркгоф). *Для любых взаимно простых чисел p и q существуют по крайней мере две периодические орбиты типа* (p, q).

Набросок доказательства. При доказательстве используется только выпуклость и гладкость. Рассмотрим всевозможные вписанные многоугольники с требуемыми комбинаторными свойствами. Будем называть такие многоугольники состояниями. Состояния образуют конечномерное пространство. На пространстве состояний есть функционал длины. Если мы разрешим вершинам многоугольников совпадать, то получим компактное пространство. Поэтому у функционала длины есть максимумы.

Любая экстремальная точка функционала длины является биллиардной орбитой (если только эта точка не граничная). Это локальное утверждение.

Легко проверить, что производная длины обращается в нуль тогда и только тогда, когда углы равны. Линейная часть изменения функционала зависит как раз от разности углов.

Легко доказать, что максимум не может достигаться на границе, т. е. вершины многоугольника не могут совпадать.

Итак, самый длинный многоугольник является требуемой периодической орбитой. Но это пока еще только самая простая часть теоремы. Нужно еще найти вторую периодическую орбиту. Это можно сделать следующим образом. Циклические перенумерации вершин найденной орбиты дают q максимумов. Продеформируем один из этих максимумов в другой. Если мы идем от одного максимума к другому, то мы должны опуститься. Будем при этом стараться терять как можно меньше высоты. В таком случае надо идти через перевал (рис. 5), потому что если бы в самой нижней точке мы не оказались на перевале, то можно слегка изменить траекторию и уменьшить потерю высоты. Перевал — тоже критическая точка, т. е. искомая периодическая орбита.

Если высоту мы вообще не теряем, то в таком случае есть целое семейство периодических орбит. □

Рис. 5. Перевал

В этом доказательстве четко видно, как можно поменять одну трудность на другую. Трудность этого рассуждения состоит в том, чтобы удержаться в стороне от границы. Этого легко добиться, если просто отказаться от границы и рассматривать все состояния. Очевидно, что рассматриваемая функция ограничена: любое звено не больше диаметра кривой. Можно отказаться от условия упорядоченности точек, а потом доказать, что глобальный максимум обязательно достигается на правильно упорядоченной орбите. Если мы, например, рассматриваем глобально максимальные орбиты, которые при полном обходе делают два оборота, то они это делают в правильном порядке. А можно вместо этого доказывать, что границы можно избежать внутри упорядоченного семейства.

Теорема Биркгофа важна тем, что мы сразу находим бесконечно много периодических орбит.

Дальше начинается интересная история о том, как Биркгоф упустил важное открытие.

Биркгоф приводит свое вариационное рассуждение и говорит, что точно так же можно доказать чисто топологически так называемую последнюю геометрическую теорему Пуанкаре: «Если на цилиндре в разные стороны крутятся основания с сохранением площади, то такой диффеоморфизм имеет по крайней мере две неподвижные точки». Более того, если

на верхнем и нижнем основаниях углы поворота разные, то для любого рационального угла поворота можно найти соответствующую периодическую орбиту, даже без условия закручивания.

Биркгоф чрезвычайно гордился тем, что доказал последнюю геометрическую теорему Пуанкаре. Но при этом он упустил совершенно замечательный вывод из своего собственного элементарного доказательства. Этот вывод состоит в следующем. Давайте посмотрим, что происходит при переходе к пределу $p_n/q_n \to \alpha$, где α — иррациональное число. Обычно в динамике такие трюки не проходят, потому что асимптотическое поведение неустойчиво относительно начальных данных, и никакого предельного перехода осуществить нельзя. Но здесь, именно из-за того, что мы имеем дело с эллиптической ситуацией, возникает простой, но удивительный феномен. Если мы рассмотрим на цилиндре биркгофовскую орбиту, то она состоит из конечного числа точек. Если число q_n велико, то точек тоже будет много и они будут сильно конденсироваться. Довольно просто доказывается, что эти точки всегда лежат на липшицевом графике (т. е. на графике функции, удовлетворяющей условию Липшица). При этом константа Липшица фиксирована, она не зависит от длины орбиты. Множество липшицевых функций с заданной константой Липшица компактно, поэтому можно перейти к пределу. Немного иначе можно сказать так: возьмем конечные орбиты и рассмотрим их предел в топологии Хаусдорфа. В топологии Хаусдорфа замкнутые подмножества компактного множества образуют компактное множество, поэтому предел существует — это неудивительно. Но предел — инвариантное множество, которое является подмножеством липшицева графика, потому что в топологии Хаусдорфа подмножества липшицевых графиков образуют замкнутое множество.

Мы пока не знаем, какая у полученного графика геометрия, но зато знаем, какая у него динамика. Динамика у него такая же, как у поворота на угол α, ничего другого быть не может. Действительно, порядок, в котором переставляются точки при повороте на угол α, однозначно определяется порядками, в которых переставляются точки при поворотах на углы p_n/q_n, аппроксимирующие угол α. Поэтому на любом конечном отрезке в пределе комбинаторика окажется такой, как надо, потому что на любом конечном отрезке в пределе комбинаторика стабилизируется и будет такой же, как при повороте на угол α.

Итак, возникает замкнутое инвариантное множество на окружности (ибо предельный липшицев график топологически есть окружность). На этом множестве есть динамика, которая сохраняет порядок и в точности воспроизводит поворот на угол α. Со времен Пуанкаре хорошо известно, в каких ситуациях такое возможно: либо инвариантное множество — вся окружность, орбиты на ней плотны и преобразование сопряжено повороту окружности (это, конечно, эллиптическое поведение, по крайней мере, в

топологическом смысле), либо на окружности есть инвариантное канторово подмножество, которое появляется в так называемом контрпримере Данжуа (см., например, главы 11 и 12 в [8]). Контрпример Данжуа устроен следующим образом. Возьмем на окружности точку и раздуем ее в интервал. Тогда ее образ и прообраз тоже придется раздуть в интервалы (рис. 6) и т. д. Чтобы была сходимость, нужно, чтобы эти интервалы становились все короче и короче. Это довольно легко сделать топологически. В результате получится преобразование окружности, у которого есть инвариантное канторово множество и которое полусопряжено повороту (существует непрерывное отображение, переводящее его в поворот, но эти интервалы сжимаются в точки).

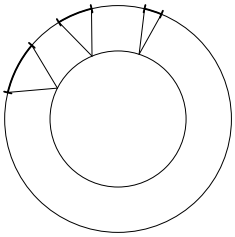


Рис. 6. Контрпример Данжуа

Для преобразований окружности такое поведение является экзотикой, потому что по теореме Данжуа в классе гладкости C^2 такого не бывает; это возможно только в C^1. Но для закручивающих отображений это вполне нормальное явление (если, конечно, не случилось, что мы получили целую окружность). Таким образом, возникает интересная альтернатива. Когда возникает накапливание биркгофовских орбит на инвариантное множество (это как раз и есть множество Обри—Мазера), то это инвариантное множество либо канторово (возможно, с некоторыми добавками), либо целая окружность. Случай целой окружности называют каустикой. Одно из двух справедливо всегда и для любого числа вращения. При этом само канторово множество единственно, т. е. если из инвариантного множества выбросить блуждающую часть, соответствующую отдельным блуждающим точкам, то оставшееся канторово множество единственно. Но это не препятствует существованию других канторовых множеств, которые имеют то же самое число вращения и на которых соблюдается порядок, но этот порядок, хотя он и совместим с циклическим порядком, не будет совместим с порядком на этом множестве. Канторово множество, построенное как предел орбит Биркгофа максимальной длины особое; его называют минимальным. Оно есть множество минимальной энергии.

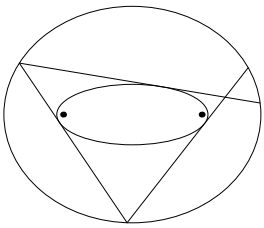


Рис. 7. Софокусные эллипсы

Следующий вопрос таков: случается ли так, что получается целая окружность? Ответ на этот вопрос иллюстрирует рис. 7. Возьмем большой эллипс в качестве биллиардного стола и рассмотрим орбиту, которая касается софокусного с ним внутреннего эллипса. Оказывается, что эта

орбита всегда будет его касаться. То же самое верно и для софокусных гипербол. Гипербола, правда, состоит из двух ветвей, но если какая-то орбита касается одной из ветвей гиперболы, то она будет и дальше касаться гиперболы, а ветви, которых касается орбита, будут при этом чередоваться.

Это картинка в конфигурационном пространстве. А что же будет в фазовом пространстве? Картинка в фазовом пространстве тоже хорошо известна; она очень похожа на картинку для маятника (рис. 8; на этом рисунке цилиндр развернут). А именно, здесь есть две орбиты периода 2, соответствующие большому и малому диаметру эллипса. «Восьмерка» соответствует орбитам, проходящим через фокусы эллипса (если орбита проходит через фокус, то она и дальше будет поочередно проходить через фокусы). То, что расположено вне этой восьмерки, состоит из орбит, которые касаются эллипсов. А то, что внутри, — из орбит, касающихся гипербол.

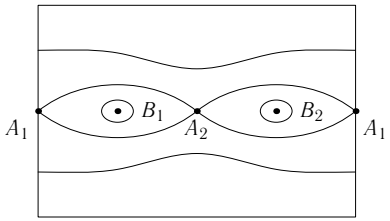


Рис. 8. Траектории в фазовом пространстве

Какие же из этих орбит соответствуют орбитам Биркгофа и Обри—Мазера, а какие не соответствуют? Иными словами, какие из этих орбит получаются с помощью конструкций Биркгофа и Обри—Мазера, а какие не получаются? Орбиты с рациональными числами вращения, касающиеся эллипсов, получаются конструкцией Биркгофа, а остальные орбиты, касающиеся эллипсов, получаются конструкцией Обри—Мазера. А гиперболы не получаются такими конструкциями. Действительно, для числа вращения $1/2$ будет одна минимальная орбита и одна минимаксная.

Интересно понять, что происходит при обратном переходе к пределу, когда мы переходим от иррациональных чисел к рациональным. В рассматриваемой ситуации ответ достаточно прост. Мы получаем инвариантную окружность, но она не вся состоит из биркгофовских орбит. Она состоит из биркгофовских орбит и асимптотических кривых. Это достаточно общий феномен, за исключением того, что не всегда получается вся окружность.

Мы рассмотрели биллиардные столы весьма специального вида. Весьма знаменит следующий вопрос, ответ на который до сих пор не получен: «Какие же еще бывают биллиардные столы, для которых хотя бы окрестность верха или низа цилиндра расслаивается на инвариантные кривые?» Иными словами, когда система вполне интегрируема? Предполагается что это случается только для эллипса.

Гораздо более фундаментален такой вопрос: когда сохраняются хотя бы какие-то кривые? Весьма замечательно, что необходимые и достаточные

условия существования хотя бы одной инвариантной кривой весьма просты. Мы, конечно, имеем в виду инвариантную кривую, обходящую вокруг цилиндра. Только такая кривая и может получиться как предел биркгофовских орбит. Несложно доказать, что если есть инвариантная кривая, на которой сохраняется порядок и имеется число вращения α, то такая кривая единственна, если число α иррационально, и эта кривая есть предел биркгофовских орбит.

Если мы интересуемся тем, что получается как предел биркгофовских орбит — кривая или канторово множество, то естественно спросить, когда именно получается кривая. Будем предполагать, что кривая, ограничивающая стол, достаточно гладкая. Например, класса C^∞ (достаточно потребовать, чтобы кривая была класса C^6). В таком случае теорема, доказанная Владимиром Федоровичем Лазуткиным (1941—2001) [10], утверждает, что инвариантная окружность существует, когда кривизна края стола нигде не обращается в нуль. На самом деле в такой ситуации инвариантных окружностей бесконечно много.

Доказательство Лазуткина является адаптацией для этого случая знаменитой теоремы Колмогорова о возмущениях вполне интегрируемых гамильтоновых систем. Формально теорема Колмогорова эту ситуацию не покрывает, потому что здесь мы имеем дело с поведением вырождающейся системы. Нужно должным образом изменить координаты, чтобы можно было что-то применять. Ненулевая кривизна нужна как раз для того, чтобы эту замену координат можно было сделать.

Если кривизна обращается в нуль, то инвариантных кривых нет. Этот гораздо более простой факт был доказан Джоном Мазером. На самом деле можно доказать более сильное утверждение. А именно, если на границе есть плоская точка, то никакое множество Обри—Мазера не может через эту точку проходить. А на инвариантной окружности должны быть не только точки, соответствующие орбитам Биркгофа, но и точки, соответствующие орбитам Обри—Мазера. (См., например, §13.5 в [8].)

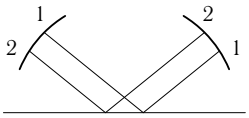


Рис. 9. Изменение порядка точек

Это рассуждение достаточно простое. При отражении от прямой меняется порядок точек (рис. 9). Если сначала идет точка 1, а потом точка 2, то после отражения сначала будет идти точка 2, а потом точка 1. На рисунке линии параллельные, но такой же эффект будет и в том случае, когда углы разные. Таким образом, при отражении от прямой не может сохраниться порядок точек. Инфинитезимально то же самое происходит и при отражении от кривой в точке с нулевой кривизной.

Здесь возникают некоторые интересные геометрические эффекты. Рассмотрим обратную задачу: как построить биллиардный стол, для которого

есть каустики? Для этого есть конструкция, которая хорошо известна для эллипса. Можно взять эллипс и набросить на него шнурок, длина которого больше длины эллипса. Затем нужно натянуть этот шнурок и провести кривую (рис. 10). В результате получится софокусный эллипс. Для большего эллипса меньший будет каустикой.

Та же самая конструкция работает и для произвольной кривой. Если вы возьмете произвольную кривую и возьмете шнурок, который длиннее этой кривой, а затем натянете шнурок и проведете кривую, то для полученной кривой исходная кривая будет каустикой.

Иногда из негладкой внутренней кривой получается гладкий биллиардный стол. Например, если в качестве внутренней кривой взять астроиду, то в результате получится гладкий стол, для которого астроида будет каустикой (рис. 11).

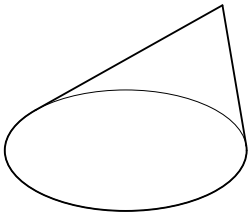


Рис. 10. Построение софокусного эллипса

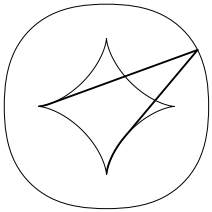


Рис. 11. Биллиардный стол для астроиды

Довольно понятно, почему эллиптические биллиарды можно рассматривать как испытательный полигон. Во-первых, они дают примеры закручивающих отображений. Биллиарды дают некоторую геометрическую интуицию, которую можно развить и потом использовать для произвольных закручивающих отображений, а закручивающие отображения покрывают много интересных случаев помимо биллиардов. А во-вторых, биллиарды дают пример одной стандартной трудности в динамике. Как учесть лагранжеву структуру? Картинка на цилиндре, где есть координата и импульс, — это гамильтонова картинка, это картинка в фазовом пространстве. А биллиард — это лагранжева система, она задается функционалом действия. И в динамике есть много трудностей, связанных с тем, как динамические эффекты связаны с лагранжевой структурой. Лагранжева структура неинвариантна, она связана с разделением на координаты и импульс.

Например, такой вопрос. Может ли у биллиарда быть открытое множество периодических орбит? Для гамильтонова закручивающего отображения пример строится очень просто. Нужно из цилиндра вырезать маленький кружок, сделать рациональный поворот, а потом вклеить кружок

обратно (рис. 12). Гамильтоновых препятствий никаких нет. Тем не менее, есть основания подозревать, что для биллиардов ничего подобного сделать нельзя. И это не праздный вопрос, потому что, например, оценки остаточных членов в асимптотиках Вейля для собственных функций оператора Лапласа зависят от предположения, что в биллиарде множество периодических орбит имеет меру нуль. Доказано это только для орбит периода 3.

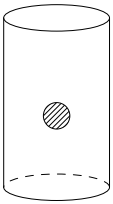


Рис. 12. Вырезанный кружок

Мы закончим обсуждение эллиптических эффектов описанием естественного «мостика» к параболическому случаю.

Рассмотрим выпуклый многоугольник P обладающий тем свойством что группа, порожденная отражениями от его сторон, порождает «замощение» плоскости. Другими словами, образы P под действием элементов этой группы покрывают плоскость и если два таких образа пересекаются, то они совпадают. Таких многоугольников совсем немного: прямоугольники, правильные треугольники, прямоугольные треугольники с углами 45 и 30 градусов. Группа, порожденная отражениями от сторон такого многоугольника, содержит нормальную подгруппу конечного индекса состоящую из параллельных переносов. В четырех случаях индексы равны 4, 6, 8 и 12 соответственно. Взяв представителей смежных классов подгруппы параллельных переносов и подействовав ими на первоначальный многоугольник получаем фундаментальную область для подгруппы параллельных переносов, которая является тором. Проделаем частичную развертку биллиардного потока с помощью выбранной фундаментальной области, т. е. вместо отражения траектории будем отражать многоугольник. Некоторые пары параллельных сторон будут теперь отождествляться параллельными переносами и биллиардный поток будет таким образом представлен как свободное движение частицы на (плоском) торе: каждый касательный вектор движется своем направлении с единичной скоростью. Это вполне интегрируемая система: начальный угол является первым интегралом, фазовое пространство расслаивается на инвариантные торы, и на каждом торе действует поток изометрий. Каждый такой поток это стандартная эллиптическая система.

3. Параболическое поведение: биллиарды в многоугольниках

Простейший параболический биллиардный стол — это прямоугольный треугольник с углом $\pi/8$. Когда траектория встречает стенку, вместо того чтобы отражать траекторию, будем отражать треугольник. В данном

конкретном случае все довольно быстро закончится. Если взять 16 копий треугольника и сделать из них 8-угольник (рис. 13), то дальнейшее движение превратится в параллельный поток на этом 8-угольнике, причем противоположные стороны будут отождествлены. Полученный объект — риманова поверхность (в данном случае рода 2) с квадратичным дифференциалом. При склейке вершин 8-угольника получится угол 6π. Чтобы разрешить эту особенность, нужно взять кубический корень. Тогда можно будет получить риманову поверхность с полем направлений. Поле направлений имеет одну особую точку, которая является седлом с шестью сепаратрисами. Это поле направлений можно реализовать с помощью квадратичного дифференциала.

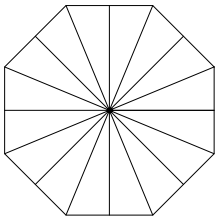


Рис. 13. Простейший параболический биллиард

У этого потока есть первый интеграл — угол (в восьмиугольнике направление движения не меняется). Этот первый интеграл имеет особенности.

Упражнение. Проанализируйте подобным образом биллиарды в правильном шестиугольнике и «гномоне».

Подобная конструкция проходит всегда, когда углы треугольника соизмеримы с π. В таком случае из конечного числа экземпляров биллиардного стола можно построить риманову поверхность с квадратичным дифференциалом. Там есть первый интеграл, и то, что получается, можно исследовать, используя весьма мощные методы из теории Тейхмюллера. Результатом является весьма хорошее понимание того, что происходит. Здесь встречаются типично параболические эффекты. Например, на всех инвариантных многообразиях (в данном случае — при фиксированном значении угла), кроме счетного числа, система получается топологически транзитивной; и также на почти всех инвариантных многообразиях система строго эргодична, т. е. инвариантная мера единственна. А в исключительных случаях, когда инвариантная мера не единственна, количество инвариантных мер не превосходит рода поверхности. Это типично параболические эффекты: инвариантная мера не всегда единственна, но обычно нетривиальных инвариантных мер конечное число.

Таким образом, биллиардная система в многоугольнике с углами соизмеримыми с π, $\pi p_i/q_i$, где p_i и q_i — взаимно простые целые числа, порождает однопараметрическое семейство потоков на некоторой поверхности, род которой определяется геометрией многоугольника и арифметическими свойствами чисел p_i/q_i. Не следует предаваться иллюзии, что структура этих потоков достаточно простая. Например, род поверхности (а, следовательно, в типичных случаях и число неподвижных точек потока) пропорционален наименьшему общему кратному знаменателей q_i.

Тем не менее, эти однопараметрические семейства обладают усложненными вариантами некоторых свойств семейства линейных потоков на торе (которые, как было объяснено выше, соответствуют биллиардам в прямоугольниках и некоторых простых треугольниках). Как я уже отметил, для почти всех значений первого интеграла, поток имеет единственную инвариантную меру (точечные меры, соответствующие положениям равновесия, не принимаются в расчет). Однако, в отличие от случая потоков на торе, множество исключительных значений параметра несчетно. Вспомним, что на торе имеется простая дихотомия между углами наклона с рациональными тангенсами, когда все орбиты замыкаются, и углами с иррациональными тангенсами, когда инвариантная мера единственна и, следовательно, любая орбита равномерно распределена по мере Лебега. В случае семейств потоков, порожденных квадратичными дифференциалами на поверхностях рода больше единицы (в частности, для семейств потоков, возникающих из биллиардов в рациональных многоугольниках), ситуация более сложная. По-прежнему имеется счетное число «рациональных» значений параметра, при которых все траектории замыкаются. Заметим, что в отличие от тора, возникает несколько разных гомотопических типов замкнутых орбит. Число таких типов можно оценить из простого соображения, что орбиты из разных семейств не пересекаются, и, следовательно, их число не превосходит рода поверхности. Кроме этого, существует множество значений параметра нулевой меры, но положительной размерности Хаусдорфа, при которых поток квазиминимален (т. е. любая полутраектория, которая не упирается в неподвижную точку, плотна), однако, существует более одной неатомической инвариантной меры.

При более глубоком рассмотрении оказывается, что это различие является отражением дихотомии между *диофантовыми* иррациональными числами или векторами, для которых скорость рациональной аппроксимации не очень высокая, и *лиувиллевыми* числами или векторами, для которых возникает «аномально хорошая» аппроксимация. В случае линейных потоков на торе, при диофантовых углах наклона временные средние для достаточно гладких функций сходятся очень быстро. Более того, диофантовы потоки отличаются большой устойчивостью: замены времени и даже сохраняющие число вращения малые нелинейные возмущения таких потоков, могут быть «выпрямлены». При лиувиллевых углах наклона временные средние могут вести себя весьма нерегулярно: то они подходят очень близко к интегралу, то проявляют сравнительно большие отклонения, так что скорость сходимости по одним последовательностям времен очень быстрая, а по другим весьма медленная. В соответствие с этим, даже гладкие замены времени существенно меняют долговременную динамику: например, собственные функции, даже измеримые, могут исчезнуть, и поток становится слабо перемешивающим.

Для потоков, возникающих из квадратичных дифференциалов, и биллиардов в рациональных многоугольниках, значения параметров, при которых имеется больше одной инвариантной меры, соответствует углам наклона с иррациональными лиувиллевыми тангенсами. Поэтому не удивительно, что возникают подобные явления, только в более яркой форме: вместо медленной сходимости средних к интегралу по мере Лебега, сходимости нет вообще. С другой стороны, для множества значений параметра полной меры, которые соответствуют углам наклона с диофантовыми тангенсами, наблюдаются сходные, хотя и значительно более сложные, явления устойчивости. Они были открыты и исследованы в течение последних пяти лет молодым математиком Джиованни Форни; его работы представляют одно из самых ярких современных достижений в теории динамических систем. Центральное наблюдение Форни состоит в том, что, хотя инвариантная мера и единственна, имеются также инвариантные распределения (обобщенные функции), т. е. инвариантные непрерывные линейные функционалы, определенные на меньших пространствах функций, чем все непрерывные функции. Для функций заданного класса гладкости, пространство инвариантных распределений конечномерно, но размерность стремится к бесконечности с ростом класса гладкости. Сочетание строгой эргодичности (единственности инвариантной *меры*) с существованием бесконечного множества независимых инвариантных *распределений* весьма характерно для динамических систем с параболическим поведением. Простейший пример, где полное исследование проводится с помощью элементарного анализа Фурье, это аффинное отображение двумерного тора

$$(x, y) \mapsto (x + \alpha, x + y) \pmod 1,$$

где α — иррациональное число. Более интересный пример, который исследуется с помощью теории бесконечномерных унитарных представлений группы $\mathrm{SL}(2, \mathbb{R})$, это орициклический поток на поверхности постоянной отрицательной кривизны.

Возвращаясь к потокам на поверхностях, отметим, что, согласно результатам Форни, инвариантные распределения определяют скорость сходимости временных средних. Грубо говоря, есть некоторая типичная степенная скорость; если первая группа инвариантных распределений обращается в нуль, эта скорость повышается, и так происходит несколько раз, пока не достигается максимально возможная скорость убывания средних, обратно пропорциональная времени. Обращение в нуль достаточного числа инвариантных распределений также гарантирует выпрямляемость потока, полученного заменой времени.

Даже в случае многоугольников с рациональными углами описание биллиарда не сводится полностью к рассмотрению по отдельности потоков на инвариантных многообразиях. Возьмем, например, вопрос о росте

числа периодических траекторий длины не более T как функции T. Конечно, периодические орбиты возникают в виде семейств, которые состоят из «параллельных» орбит одинаковой длины. Поэтому нужно считать число $P(T)$ таких семейств. В случае биллиарда в прямоугольнике (который, как мы несколько раз отмечали, сводится к геодезическому потоку, т. е. свободному движению частицы на плоском торе) эта задача сводится после соответствующей перенормировки к подсчету числа точек с целыми координатами в круге радиуса T с центром в начале координат. Таким образом

$$\lim_{T\to\infty} \frac{P(T)}{\pi T^2} = 1.$$

Для общих рациональных биллиардов рост $P(T)$ также оказывается квадратичным, т. е.

$$0 < \liminf_{T\to\infty} \frac{P(T)}{T^2} \leqslant \limsup_{T\to\infty} \frac{P(T)}{T^2} < \infty.$$

Кроме того, известно, что периодические орбиты плотны в фазовом пространстве. Вопрос о существовании предела $\dfrac{P(T)}{T^2}$ при $T \to \infty$ для произвольного рационального прямоугольника пока остается открытым. Положительный ответ получен, с одной стороны, для некоторых специальных многоугольников, которые приводят к квадратичным дифференциалам на поверхностях с большим количеством симметрий (поверхности Вича), а с другой стороны, для квадратичных дифференциалов общего положения. Вполне возможно, что существуют многоугольники с патологическим поведением функции $P(T)$. Отметим, что наш первый нетривиальный пример биллиарда в прямоугольном треугольнике с углом $\pi/8$ и гипотенузой 1 приводит к поверхности Вича, и для него $\lim\limits_{T\to\infty} \dfrac{P(T)}{T^2} = ?$.

Про биллиарды в многоугольниках, у которых не все углы соизмеримы с π известно удивительно мало. Такие биллиарды являются хорошими примерами параболических систем достаточно общего вида. Приходится констатировать, что доступные нам методы анализа недостаточны для серьезного исследования таких систем. Действительно, успехи в исследовании параболических систем связаны с двумя специальными ситуациями:

(1) потоки на поверхностях, которые обсуждались выше, и где размерность фазового пространства очень мала (в дополнение к размерности, соответствующей орбитам, есть только одна трансверсальная размерность), и

(2) потоки на однородных пространствах, где локально имеется большая симметрия.

Два основных открытых вопроса, относящихся к произвольным биллиардам, это описание глобальной сложности поведения траекторий, и асимптотическое поведение типичных траекторий то отношению к мере Лебега.

Начнем со второго вопроса. Здесь известно очень много, и в то же время очень мало. Если фиксировать тип биллиардного стола (например, выпуклые многоугольники с заданным числом сторон), то углы служат естественными параметрами в пространстве таких биллиардов. Биллиарды с углами, соизмеримыми с π, про которые, как было выше объяснено, известно довольно много, образуют плотное множество в таком пространстве. Отправляясь от эргодичности рациональных биллиардов на большинстве инвариантных многообразий, и принимая во внимание тот факт, что при больших знаменателях каждое такое многообразие почти равномерно заполняет фазовое пространство, можно показать с помощью довольно стандартных категорных рассуждений, что для всюду плотного G_δ в пространстве параметров биллиард эргодичен во всем фазовом пространстве. Однако, это топологически значимое множество биллиардов является весьма тощим с метрической точки зрения: не только его мера Лебега, но даже размерность Хаусдорфа этого множества в пространстве параметров равна нулю. Это множество напоминает множество чисел, допускающих рациональную аппроксимацию с исключительно высокой скоростью, наподобие тройной экспоненты. Предполагается, что для типичных диофантовых значений вектора углов биллиард эргодичен. На сегодняшний день серьезных подходов к этой задаче не известно. Неизвестны также более тонкие статистические свойства, такие как перемешивание, для каких бы то ни было иррациональных биллиардов, включая описанную выше лиувиллевскую ситуацию, для которой доказана эргодичность. Неизвестна также структура сингулярных инвариантных мер для иррациональных биллиардов.

Конечно, частный случай этого последнего вопроса, это описание периодических траекторий, т. к. каждая такая траектория порождает сингулярную эргодическую инвариантную меру. С одной стороны не известно, имеется ли для произвольного многоугольника хотя бы одна периодическая биллиардная траектория. Как отмечалось выше, для рациональных многоугольников таких траекторий бесконечно много, и они плотны в фазовом пространстве. Однако, предельного перехода к иррациональным многоугольникам, даже специального вида, осуществить не удается. Проблема здесь состоит в том, что при увеличении знаменателя инвариантные многообразия становятся поверхностями очень высокого рода, и периодические орбиты имеют очень сложный гомотопический тип, и следовательно являются очень длинными. Есть, однако, некоторые специальные ситуации, когда возникают периодические орбиты с простой комбинаторикой, которые сохраняются при малом возмущении углов. Классический пример это орбита периода 3, образованная основаниями высот в произвольном остроугольном треугольнике. Эта орбита, конечно, допускает вариационное описание. Однако, в отличие от орбит Биркгофа в выпуклых биллиардах, треугольник, образованный основаниями высот, имеет минимальный

периметр среди всех вписанных треугольников. Максимальный же, и минимаксный треугольники вырождаются в удвоенную максимальную высоту. Орбита периода 3, построенная таким образом, окружена семейством параллельных орбит периода 6 (см. рис. 14). Заметим, что это единственные периодические орбиты, существование которых известно для всех остроугольных треугольников. Вопрос о плотности, или хотя бы существовании бесконечного числа параллельных семейств периодических орбит остается открытым.

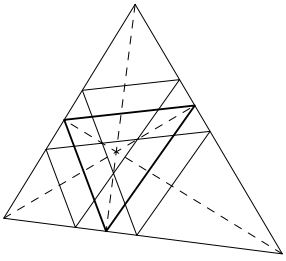


Рис. 14. орбиты периода 3 и 6 в остроугольном треугольнике

Для произвольного прямоугольного треугольника существование периодических орбит было доказано всего несколько лет назад. К сожалению, вид этих орбит несколько разочаровывает. Это траектории, которые отражаются перпендикулярно от одной из сторон, и после конечного числа отражений возвращаются на эту же сторону тоже в перпендикулярном направлении. Очевидно, такая орбита отскакивает обратно и повторяет свой путь в противоположном направлении. Это как раз примеры орбит с устойчивой комбинаторикой. Оказывается, что для почти любого начального положения, орбита, перпендикулярная к стороне, возвращается на эту сторону в перпендикулярном направлении и, следовательно, оказывается периодической. Это является довольно простым следствием сохранения меры и того факта, что возможные направления орбиты образуют единственную траекторию бесконечной диэдрической группы, порожденной отражениями от двух не перпендикулярных сторон треугольника. Это соображение обобщается на некоторые многоугольники, «близкие» к рациональным, т. е. такие, у которых значения углов по модулю π лежат в одномерном пространстве над рациональными числами. Для произвольного тупоугольного треугольника это соображение не применимо и существование даже одной периодической орбиты неизвестно.

Существование периодических орбит тесно связано с вопросом о глобальной сложности поведения траекторий. Рост числа различимых траекторий со временем можно оценивать разными способами. Наиболее естественный способ связан с кодированием. Каждой траектории ставится в соответствие последовательность символов, в соответствии с отражениями от сторон многоугольника, так что каждая сторона обозначается своим символом. Конечно, при этом естественным образом кодируются биллиардные отображения, т. е. отображения возвращения биллиардного потока на границу. Для того, чтобы получить полную информацию о потоке, нужно еще указать время между последовательными отражениями. Рост слож-

ности для биллиардного отображения (соответственно, потока) задается функцией $S(N)$ (соответственно, $\mathscr{S}(T)$) равной числу различных кодов длины N (соответственно, числу различных кодов, возникающих для отрезков траекторий длины T). Очевидно, что каждое семейство параллельных периодических орбит порождает бесконечный периодический код, и почти столь же очевидно, что и наоборот, каждому бесконечному периодическому коду соответствует семейство параллельных периодических орбит. Эти орбиты могут замыкаться либо после одного периода, либо после двух (как орбиты периода 6, параллельные треугольнику Фаньяно в остроугольном треугольнике).

В случае многоугольников с рациональными относительно π углами, обе функции допускают квадратичную оценку:

$$0 < \liminf_{N\to\infty} \frac{S(N)}{N^2} \leqslant \limsup_{N\to\infty} \frac{S(N)}{N^2} < \infty$$

и

$$0 < \liminf_{T\to\infty} \frac{\mathscr{S}(T)}{T^2} \leqslant \limsup_{T\to\infty} \frac{\mathscr{S}(T)}{T^2} < \infty.$$

Заметим, что в этом случае положительная доля всех допустимых кодов реализуется периодическими траекториями.

Альтернативный способ описания сложности состоит в подсчете числа способов, которыми коды могут изменяться. Очевидно, что код меняется, когда траектория попадает в угол. Очевидно также, что имеется только конечное число отрезков траекторий ограниченной длины, которые попадают в углы и в положительном, и в отрицательном направлении. По довольно очевидным причинам такие сингулярные траектории называют *обобщенными диагоналями* многоугольника. Определим $D(N)$ (соответственно, $\mathscr{D}(T)$) как число обобщенных диагоналей из $\leqslant N$ звеньев (соответственно, число обобщенных диагоналей длины $\leqslant T$). Как и выше, эти величины допускают квадратичную оценку для рациональных многоугольников.

Естественно предположить, что для произвольных многоугольников рост траекторий не должен быть намного более быстрым, чем для рациональных, т. к. локальная геометрическая структура биллиардного потока одна и та же в рациональном и в общем случае. Однако, единственный известный факт в этом направлении состоит из гораздо более слабых субэкспоненциальных оценок:

$$\lim_{N\to\infty} \frac{\log S(N)}{N} = \lim_{N\to\infty} \frac{\log D(N)}{N} = \lim_{T\to\infty} \frac{\log \mathscr{S}(T)}{T} = \lim_{T\to\infty} \frac{\log \mathscr{D}(T)}{T} = 0.$$

4. Гиперболическое поведение: биллиарды Синая, Бунимовича, Войтковского и прочих авторов

Как мы уже отмечали, гиперболическое поведение довольно широко распространено и позволяет установить основные элементы стохастического или «хаотического» поведения. Преобладание гиперболического поведения естественно по аналогии с линейными системами. Действительно, случайно выбранная матрица, скорее всего не имеет собственных значений по модулю равных единице. Даже если заранее ограничиться матрицами с единичным определителем, это по-прежнему верно для матриц размера 3 на 3 или больше. Хотя эту аналогию и не удается распространить буквально на нелинейные системы, она по крайней мере показывает важность гиперболической парадигмы.

Самые ранние примеры гиперболического поведения биллиардов были найдены Я. Г. Синаем [13]. Простейшие примеры биллиарда типа Синая это, во-первых, квадрат с вырезанным круговым отверстием, а во-вторых, выпуклый многоугольник, стороны которого заменены дугами, выпуклыми внутрь (см. рис. 1). С точки зрения строгого математического анализа, второй пример оказывается несколько более простым, чем первый. Гиперболическое поведение в биллиардах Синая связано с явлением рассеяния света, хорошо известным из геометрической оптики: параллельный или расходящийся поток света при отражении от выпуклого зеркала становится более расходящимся. Не очень сложные вычисления показывают, что если отражение происходит достаточно регулярно, то угловой размер пучка растет экспоненциально. Это и обеспечивает гиперболичность линеаризованной системы. При анализе рассеивающих биллиардов возникают две технические трудности.

Во-первых, нужно добиться достаточной регулярности отражений от выпуклых внутрь частей границы. Сразу видно, почему второй пример в этом отношении лучше первого: в нем время между двумя последовательными отражениями ограниченно. В первом же примере, имеются периодические траектории, параллельные сторонам квадрата, которые вообще не отражаются от препятствия. Такие траектории, конечно, образуют множество нулевой меры, но траектории, образующие с ними очень маленький угол, впервые встречают препятствие только через очень большое время. Это явление называется бесконечным горизонтом; соответственно, ограниченность времени между отражениями, соответствует конечному горизонту. Бесконечный горизонт влечет неравномерность гиперболических оценок по фазовому пространству. Хотя это приводит к существенным техническим усложнениям в доказательствах эргодичности, перемешивания, и других

стохастических свойств, это также подтверждает роль биллиардов, как важного испытательного полигона для различных методов и средств анализа динамики. Действительно, неравномерная гиперболичность гораздо более распространена, чем равномерная. Например, глобальное равномерное гиперболическое поведение для классических консервативных систем, налагает ограничение на топологию фазового пространства. В то же время, неравномерная гиперболичность совместима с любой топологией. Этот факт, хотя и был предсказан довольно давно, был в полной общности установлен только недавно Д. И. Долгопятом и Я. Б. Песиным [6].

Вторая трудность при анализе рассеивающих биллиардов — это наличие особенностей (разрывов и неограниченности производных) в системе. Этим они отличаются от биллиардов в гладких выпуклых областях, рассмотренных выше, где биллиардное отображение является гладким. Особенности возникают в точках касания траекторий с выпуклыми внутрь частями границы. Они конечно, также возникают, когда траектория попадает в угол. Особенности второго типа возникают и в параболических биллиардах, и в случае рассеивающих биллиардов приводят только к небольшим осложнениям. Такие особенности приводят к разрывам первого рода для функций, представляющих динамику: возникает поверхность разрыва, и функции являются гладкими с обеих сторон этой поверхности. Таким образом, дифференциал вдоль биллиардной траектории, не попадающей в саму точку разрыва, ведет себя вполне регулярно. Для траекторий, касающихся границы изнутри, производные около этих траекторий неограниченны, так что разрывы носят более серьезный характер. Отметим, что упругие столкновения, и более сложные эффекты подобного рода естественно возникают во многих важных задачах классической механики, например, в задаче n тел. Влияние таких явлений на долговременное поведение траекторий является одной из центральных проблем механики. Здесь также биллиарды, и особенно их многомерные аналоги, играют роль важного испытательного полигона.

Рассеивающие биллиарды весьма существенны для математического обоснования моделей статистической физики. Это важная и интересная тема, которой мы, однако, не будем здесь касаться. С точки зрения геометрии, рассеивающие биллиарды обладают некоторыми дефектами, например, неизбежным наличием особенностей на границе. Правда, если рассматривать биллиарды не в областях на плоскости, а в областях на плоском торе, этого недостатка можно избежать. Например, биллиард на торе с вырезанным кругом — это классический пример биллиарда Синая. Тем не менее, интересно выяснить, каким образом гиперболические поведение может возникать другими способами, чем в результате рассеивания выпуклыми внутрь частями границы. Первый ответ на этот вопрос дается довольно знаменитым примером «стадиона», т. е. двух полуокружностей,

соединенных отрезками общих касательных (см. рис. 15). Это пример так называемых биллиардов Бунимовича [2], где гиперболическое поведение возникает как результат последовательных фокусировок пучков орбит. С точки зрения конфигурационного пространства эта картина драматически отличается от случая рассеивающих биллиардов; однако, в фазовом пространстве, где принимаются во внимание и координаты и скорости, возникает равномерный экспоненциальный рост.

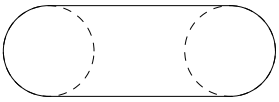


Рис. 15. Стадион

Биллиарды Бунимовича были изобретены очень интересным образом. В начале 70-х годов Л. А. Бунимович, который был тогда аспирантом Синая, работал над расширением класса биллиардов с экспоненциальным разбеганием и стохастическим поведением орбит. Он обнаружил, что если добавить к рассеивающему биллиарду небольшие круглые «лузы», то биллиард на полученном таким образом столе, в котором выпуклые участки чередуются с вогнутыми, обладает экспоненциальным разбеганием траекторий. На самом деле, Бунимович открыл важный новый механизм гиперболичности. Однако, он сам сначала рассматривал свою работу лишь как небольшое обобщение результатов о рассеивающих биллиардах. Во время доклада Бунимовича на семинаре в МИАНе, которым руководили Д. В. Аносов и автор, возник естественный вопрос о механизме гиперболичности, и в частности о том, необходимо ли наличие каких бы то ни было рассеивающих компонент. Я обратил внимание докладчика, что из его рассуждений такая необходимость как будто бы не вытекала, и предложил стадион как модель для проверки этой гипотезы. Остальные геометрические условия Бунимовича были выполнены, по крайней мере, если полные круги не пересекались (см. рис. 15). Немного подумав, Бунимович сказал, что его рассуждения должны проходить в этом случае, и в следующем варианте своей работы он сформулировал условия, которые не требовали наличия рассеивающих компонент. Более того, оказалось, что первоначальные геометрические условия можно ослабить, так что, например, в случае стадиона, окружности достаточно раздвинуть на произвольно малое расстояние.

Среди биллиардов Бунимовича много других интересных и весьма простых форм, однако все они имеют общее свойство, что граница, помимо рассеивающих участков, может включать только отрезки прямых и дуги окружностей. Естественный вопрос, насколько существенно это условие, занимал специалистов в течении, примерно, десяти лет. Техническая трудность заключается в следующем. Гиперболичность устанавливается с помощью системы конусов в касательных пространствах к точкам фазового пространства, которые переходят в себя под действием динамики. Для простоты и геометрической наглядности лучше думать о биллиардном

отображении, а не о потоке. В этом случае фазовое пространство двумерно, и конуса, о которых идет речь, это внутренности двух противоположных углов, образованных парой прямых, пересекающихся в начале координат. Система конусов, которая инвариантна и в рассеивающих биллиардах, и в биллиардах Бунимовича, одна и та же. Геометрически, эти конуса определяются, как множества инфинитезимальных расходящихся кусков траекторий. Для получения гиперболичности необходимо, чтобы конус вместе со своей границей отображался строго внутрь соответствующего конуса в образе. Это, конечно, выполняется в случае рассеивающих биллиардов уже при одном отражении. В случае же плоских и круговых зеркал, конус переходит в себя, но одна из его сторон остается инвариантной. Это типично параболический эффект, т. к. именно так действуют унипотентные матрицы. Возьмем, например, матрицу $\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}$. Конус, о котором идет речь, определяется условием $x_1x_2 > 0$, т. е. это объединение первого и третьего квадранта на плоскости. Его образ — это конус $|x_1| > |x_2|$, $x_1x_2 > 0$ (см. рис. 16). При дальнейших итерациях образ становится все

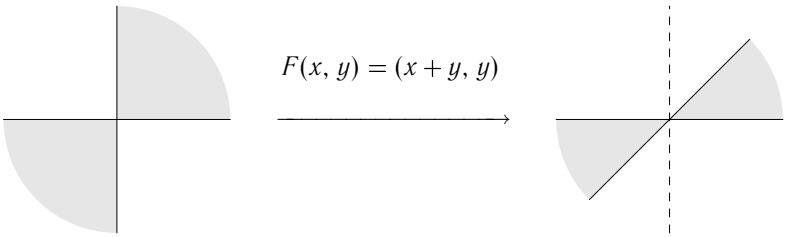


Рис. 16. Действие параболического преобразования на конус

более узким, но он по-прежнему «прилипает» к горизонтальной оси. Для того, чтобы получить гиперболичность, Бунимович использует геометрическое условие, которое обеспечивает строгую инвариантность конусов после отражения от *различных* круговых участков границы (как в случае стадиона). Так как траектория, отражающаяся от кругового участка под очень маленьким углом, продолжает это делать много раз, казалось, что явный вид итерации при отражении от круговых участков (интегрируемость биллиарда в круге), играл существенную роль. Так я объяснял для себя жесткие ограничения в условиях Бунимовича.

Оказалось, однако, что эту трудность можно преодолеть. Биллиарды с выпуклыми участками границы могут быть гиперболическими по целому ряду причин. Как только точка зрения, основанная на использовании систем инвариантных конусов. была осознана, задача нахождения новых классов гиперболических биллиардов стала задачей на сообразительность.

Отметим, что Бунимович использовал другую технику, которая формально эквивалентна системе инвариантных конусов, но значительно менее наглядна. Пионерами использования метода инвариантных конусов в динамике были В. М. Алексеев (1932—1980) и Юрген Мозер (1928—1999). Существенным шагом было введение этого метода в неравномерной гиперболической ситуации. Автор использовал этот метод для построения примеров гладких систем со стохастическим поведением на различных многообразиях. Однако, основная заслуга здесь принадлежит Мачею Войтковскому. И опять биллиарды оказались идеальным испытательным полигоном. Осознав ключевую роль систем конусов, Войтковский понял, что задачу можно решать в обратном порядке: подбирать классы биллиардных столов под заданную систему конусов. Препринт его ключевой работы на эту темы [3] назывался «Principles for the design of billiards with nonvanishing Lyapunov exponents», в несколько вольном переводе «Принципы проектирования биллиардов с ненулевыми показателями Ляпунова». Также как квадрат или тор с вырезанным кружком является квинтэссенцией феномена, открытого Синаем, а стадион символизирует биллиарды Бунимовича, характерный пример биллиардов Войтковского дается кардиоидой (см. рис. 1). Принципиальная важность результатов Войтковского для теории биллиардов состоит в том, что он открыл классы гиперболических примеров, которые открыты в C^2 топологии и таким [11] образом это свойство не зависит от небольших погрешностей зеркал.

Как я уже отмечал, построение новых классов гиперболических биллиардов стало возможно с использованием метода инвариантных конусов. Как пример гибкости этого метода упомянем следующий результат Виктора Доннэ [7]: любой достаточно малый кусок выпуклой кривой является частью границы кусочно-гладкого выпуклого гиперболического биллиарда. Отметим также, что использование метода инвариантных конусов позволило получить много замечательных примеров классических динамических систем с неравномерным гиперболическим поведением.

Важные нерешенные задачи связаны с существованием гиперболических биллиардов с гладкой (по крайней мере дважды дифференцируемой) границей. Отметим что граница стадиона дифференцируема, но кривизна (и, следовательно, вторая производная) разрывна. Неизвестны дважды дифференцируемые примеры даже с невыпуклой или неодносвязной границей.

Что же дает гиперболичность? Она позволяет показать, что детерминистская динамическая система во многих отношениях ведет себя как последовательность независимых случайных величин. В некотором смысле это утверждение верно буквально: при выполнении некоторых (часто проверяемых без большого труда) условий, в дополнение к (даже неравномерной) гиперболичности фазовое пространство сохраняющей ко-

нечный объем системы, можно разделить на конечное число множеств $A_1, \ldots, A_n$ положительной меры, так что, во-первых, каждая точка фазового пространства кодируется последовательностью попаданий в эти множества в положительные и отрицательные моменты времени, а во-вторых, эти множества полностью независимы по отношению к динамике F, т. е.

$$\operatorname{vol}\Big(\bigcap_{k=0}^{n} F^k(A_{i_k})\Big) = \prod_{k=0}^{n} \operatorname{vol} A_{i_k}.$$

Хотя эти множества носят экзотический характер, но из этого свойства, которое, естественно, называется свойством Бернулли, следует много важных свойств: сходимость временных средних к пространственному (эргодичность), убывание корреляции (перемешивание), асимптотическая независимость будущего от прошлого (K-свойство, или свойство Колмогорова).

Литература

[1] Дж. Д. Биркгоф. Динамические Системы. М.—Л.: ОГИЗ, Гостехиздат, 1941.

[2] L. A. Bunimovich. On the ergodic properties of nowhere dispersing billiards // Comm. Mathematics. Phys. **65** (1979). № 3. P. 295—312.

[3] M. Wojkowski. Invariant families of cones and Lyapunov exponents // Erg. Theory and Dynam. Syst. **5** (1985). P. 145—161.

[4] Г. А. Гальперин, А. Н. Земляков. Математические бильярды. М.: Наука, 1990. (Библиотечка «Квант»; вып. 77.)

[5] Г. А. Гальперин, Н. И. Чернов. Биллиарды и хаос. М.: Знание, 1991.

[6] D. Dolgopyat, Ya. Pesin. Every compact manifold carries a completely hyperbolic diffeomorphism. To appear in Erg. Theory and Dynam. Syst.

[7] V. J. Donnay. Using integrability to produce chaos: billiards with positive entropy // Comm. Math. Phys. **141** (1991). № 2. P. 225—257.

[8] А. Б. Каток, Б. Хасселблатт. Введение в современную теорию динамических систем. М.: Факториал, 1999.

[9] И. П. Корнфельд, Я. Г. Синай, С. В. Фомин. Эргодическая теория. М.: Наука, 1980.

[10] В. Ф. Лазуткин. Существование каустик для биллиардной задачи в выпуклой области // Изв. АН СССР **37** (1973), № 1, С. 186—216.

[11] H. Masur, S. Tabachnikov. Rational billiards and flat structures. To appear in Handbook in Dynamical Systems **1A**, Elsevier.

[12] Hard ball systems and the Lorentz gas / Edited by D. Szász. Springer-Verlag, Berlin, 2000. (Encyclopaedia of Mathematical Sciences, **101**. Mathematical Physics, II.)

[13] Я. Г. Синай. Динамические системы с упругими отражениями. Эргодические свойства рассеивающих бильярдов // Успехи Мат. Наук **25** (1970), вып. 2. С. 141—192.

[14] S. Tabachnikov. Billiards // Panoramas et Synthèses **1** (1995).

А. Н. Рудаков

Числа Фибоначчи и простота числа $2^{127}-1$

Лекция 3 апреля 1999 года

1. Введение

Число $M = 2^{127} - 1$ долгое время было в списке рекордов, оно являлось самым большим известным простым числом с 1877 г. по 1951 г. Простота $2^{127} - 1$ была установлена Э. Лукасом (É. Lucas). Им был найден замечательный способ доказательства простоты, потребовавший для $M = 2^{127} - 1$ около ста часов вычислений, но никаких делений на меньшие простые числа. Я собираюсь изложить математическую часть алгоритма Лукаса, обсудив заодно некоторые изящные результаты из конечной арифметики. Сами вычисления мы проводить не будем.

Мне лично этот сюжет кажется очень хорошим примером того, что для построения хорошего алгоритма нужна хорошая теория. Впрочем, существуют изложения результата Лукаса, привлекающие значительно меньше «теории», чем мое изложение здесь (см. [R] и [B]).

Подробное историческое исследование работ Э. Лукаса и нахождения простых чисел см. в [W].

2. Числа Фибоначчи и основная теорема

Как известно, последовательность чисел Фибоначчи получается следующим образом: мы определяем $u_1 = 1$, $u_2 = 1$ и находим каждое следующее число по формуле $u_{n+1} = u_n + u_{n-1}$. У чисел Фибоначчи

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, \ldots$$

немало замечательных свойств. Например, сначала идут два нечетных числа, потом четное, а потом опять два нечетных и т. д. Это легко увидеть, если рассматривать числа Фибоначчи по модулю 2. Если мы знаем u_{n-1} и u_n по модулю 2, то u_{n+1} будет их «суммой по модулю 2». Следовательно,

мы имеем последовательность:

$$u_3 \equiv 1+1 \equiv 0 \pmod 2$$
$$u_4 \equiv 0+1 \equiv 1 \pmod 2$$
$$u_5 \equiv 1+0 \equiv 1 \pmod 2$$
$$\dots\dots\dots$$

или 1, 1, 0, 1, 1, 0, 1, 1, 0, ... Это и означает, что каждое третье число четно, а числа перед ним и после него нечетны и т. д.

Можно заметить и что каждое пятое число делится на 5. Для этого надо только вычислить числа Фибоначчи по модулю 5. Это будет последовательность чисел

$$1, 1, 2, 3, 0, 3, 3, 6, 9 \equiv -1, 0, -1, -1, -2, -3, 0, -3, -3, \dots$$

После 20-го члена все начнет повторяться, и регулярно, через четыре места на пятом идут нули.

Задача 1. *Покажите, что каждое четвертое число Фибоначчи делится на* 3.

Задача 2. *Покажите, что если* m *делит* u_k, *то* m *делит* u_{2k}, u_{3k}, u_{4k}, ...

Можно получить также формулу для чисел Фибоначчи. Она очень известна, но позвольте мне напомнить, как мы рассуждаем. Если мы отвлечемся от «начальных данных», $u_1 = 1$, $u_2 = 1$, а рассмотрим только уравнение перехода

$$x_{n+1} = x_n + x_{n-1}, \tag{1}$$

то, конечно, есть много последовательностей, удовлетворяющих этому уравнению. Одна из них, называемая иногда *числами Лукаса*, это:

$$v_1 = 1,\ v_2 = 3,\ v_3 = 4,\ \dots,\ v_{n+1} = v_n + v_{n-1}.$$

Есть и еще, при этом если $\{a_n\}$ и $\{b_n\}$ — две такие последовательности, то можно построить третью, взяв их линейную комбинацию с некоторыми коэффициентами, например, $c_n = 2a_n + 3b_n$. Тут стоят коэффициенты 2 и 3, но они могут быть любыми. В частности, если

$$\alpha = \frac{1+\sqrt{5}}{2} \quad \text{и} \quad \beta = \frac{1-\sqrt{5}}{2},$$

т. е. α и β — корни уравнения $x^2 = x + 1$, то последовательности $a_n = \alpha^n$ и $b_n = \beta^n$ удовлетворяют уравнению перехода (1), а значит, и любая их линейная комбинация обладает этим свойством. Так как $\alpha+\beta = 1$, $\alpha^2+\beta^2 = 3$, то сумма этих последовательностей дает числа Лукаса

$$\alpha^2 + \beta^2 = v_n. \tag{2}$$

Для чисел Фибоначчи надо более искусно подобрать коэффициенты. В результате получится

$$u_n = \frac{\alpha^n - \beta^n}{\alpha - \beta}. \tag{3}$$

В частности, из этого следует, что $u_{2n} = u_n \cdot v_n$.

Мне бы хотелось сейчас сформулировать нашу основную теорему, которая есть по существу теорема Лукаса (1886), хотя она не была сформулирована им в такой форме. Современное изложение исторических деталей есть в [W].

Теорема 1. *Пусть q — простое число вида $4k+3$ и $M = 2^q - 1$. Тогда M простое если и только если $v_{\frac{M+1}{2}} \equiv 0 \pmod{M}$.*

Этот результат является основой алгоритма, позволяющего установить простоту числа $2^{127}-1$, однако надо еще добавить «быстрый» способ вычисления $v_{\frac{M+1}{2}}$. Мы это обсудим позже.

3. Комплексные числа в конечной арифметике

Давайте немножко изменим способ выражения: вместо того чтобы говорить a сравнимо с b по модулю m, $a \equiv b \pmod{m}$, будем говорить a равно b в «арифметике по модулю m», $a =_{(m)} b$. Формально это ничего не меняет, чуть-чуть другие слова, но можно начать представлять себе, что есть некие числа «арифметики по модулю m», которые просто обозначаются целыми числами, а сами по себе есть нечто другое. Например, 6 и -1 это два обозначения для одного и того же числа «арифметики по модулю 7».

При таком подходе почти сразу возникает вопрос, а нельзя ли увеличить область чисел, рассмотрев, например, «комплексные числа». Ведь комплексные числа — это пары действительных, а пары можно рассматривать и здесь. Давайте рассмотрим «комплексные числа по модулю 7». Определим такое комплексное число z как пару $z = (a, b)$, где a и b — «числа по модулю 7». Сложение задается обычным образом:

$$(a_1, b_1) + (a_2, b_2) = (a_1 + a_2,\ b_1 + b_2);$$

умножение тоже:

$$(a_1, b_1) \cdot (a_2, b_2) = (a_1 a_2 - b_1 b_2,\ a_1 b_2 + a_2 b_1).$$

Довольно просто убедиться, что это хорошее определение: есть нуль, единица, ассоциативность, коммутативность... Можно вычислить и обратный

элемент: если $z = (a, b)$, то

$$z^{-1} = \Big(\frac{a}{a^2+b^2}, \frac{-b}{a^2+b^2}\Big).$$

Однако может оказаться, что $a^2 + b^2 =_{(7)} 0$ и обратный элемент не определен.

Мы можем это явно проверить, так как 7 — очень небольшое число. Всевозможные элементы по модулю 7 легко выписываются, это:

$$0,\ 1,\ 2,\ 3,\ 4,\ 5,\ 6.$$

Квадратами будут 0, 1, 4, 2, и все! Суммы двух квадратов получаются такие:

$$0,\ 1,\ 4,\ 2;\quad 1,\ 2,\ 5,\ 3;\quad 4,\ 5,\ 3,\ 6;\quad 2,\ 3,\ 6,\ 4.$$

То есть 0 встречается только один раз как $0 = 0^2 + 0^2$, все остальные суммы ненулевые. Значит, у ненулевого комплексного числа есть обратный элемент. Получилась хорошая арифметика со всеми четырьмя операциями, или то, что иначе называют полем, а точнее, квадратичным расширением простого поля из 7 элементов.

Кстати, 7 нельзя заменить на 5, поскольку $1^2 + 2^2 =_{(5)} 0$ в арифметике по модулю 5. Проблема, собственно, в том, какие числа в арифметике по модулю p надо считать отрицательными.

Вспомним, для построения обычных комплексных чисел мы берем -1, т. е. отрицательное число, для которого не существует квадратного корня, и «добавляем» этот квадратный корень формально, т. е. пишем $z = a + bi$, где $i^2 = -1$. Далее правила операций возникают сами собой, из раскрытия скобок:

$$(a_1 + b_1 i) + (a_2 + b_2 i) = (a_1 + a_2) + (b_1 + b_2)i,$$

$$(a_1 + b_1 i) \cdot (a_2 + b_2 i) = (a_1 a_2 + b_1 b_2 (-1)) + (a_1 b_2 + a_2 b_1)i.$$

Можно взять и другое отрицательное число, например -2, и рассмотреть комплексные числа в виде $z = a + bj$, где $j^2 = -2$. Получится все то же самое, в частности, операции тоже возникают сами собой, из раскрытия скобок:

$$(a_1 + b_1 j) + (a_2 + b_2 j) = (a_1 + a_2) + (b_1 + b_2)j,$$

$$(a_1 + b_1 j) \cdot (a_2 + b_2 j) = (a_1 a_2 + b_1 b_2 (-2)) + (a_1 b_2 + a_2 b_1)j.$$

Отличие возникает только в одном месте, где приходится считать j^2. Формулой для обратного элемента будет

$$(a + bj)^{-1} = \frac{a}{a^2+2b^2} + \frac{-b}{a^2+2b^2}\, j,$$

и так как $a^2+2b^2 \neq 0$ как только $(a, b) \neq (0, 0)$, то никаких проблем не возникает.

Задача 3. *Проверьте, что можно построить квадратичное расширение простого поля из* 5 *элементов, рассматривая числа* $a+bj$, *где* $j^2=-2$. *Все четыре действия арифметики будут корректно определены.*

Давайте будем использовать такое определение: скажем, что элемент a является отрицательным в «арифметике по модулю p», где p — простое число, если уравнение $x^2 =_{(p)} a$ не имеет решений, и положительным в противном случае (если при этом $a \neq_{(p)} 0$). Например, по модулю 5 числа 1 и 4 положительные, а 2 и 3 — отрицательные. Так как $-1 =_{(5)} 4$, то число -1 тоже положительное, так уж получается. Зато по модулю 7 числа 1, 2, 4 положительные, а -1, -2 и -4, или 6, 5 и 3, отрицательные. То, что нам естественно назвать «знаком элемента», исторически называется символом Лежандра $\left(\frac{a}{p}\right)$. По определению

$$\left(\frac{a}{p}\right) = \begin{cases} +1, & \text{если } a \text{ положительно по модулю } p; \\ -1, & \text{если } a \text{ отрицательно по модулю } p; \\ 0, & \text{если } a =_{(p)} 0. \end{cases}$$

Можно проверить, что для нечетного простого числа p ровно половина $\left(\text{т. е. } \frac{p-1}{2}\right)$ ненулевых чисел по модулю p положительна и ровно половина отрицательна, и что произведение двух отрицательных всегда положительно.

Задача 4. *Докажите, что если* $\left(\frac{a}{p}\right)=-1$ *и* $\left(\frac{b}{p}\right)=-1$, *то* $\left(\frac{ab}{p}\right)=+1$.

Задача 5. *Проверьте, что если* t *отрицательно по модулю* p, *то числа* $a+bj$, *где* $j^2=t$, *определяют квадратичное расширение простого поля из* p *элементов* (*где все четыре действия арифметики корректно определены*).

Главное приложение вышесказанного для нас в следующем. Пусть p — простое число и число 5 отрицательно по модулю p. Тогда числа $\alpha=\frac{1+\sqrt{5}}{2}$ и $\beta=\frac{1-\sqrt{5}}{2}$ определены как комплексные числа по модулю p (как элементы квадратичного расширения) и формулы (2) и (3) для чисел Лукаса и Фибоначчи сохраняют смысл в комплексных числах по модулю p.

4. Комплексное сопряжение для чисел по модулю p

Существенной составляющей структуры обычных комплексных чисел является операция комплексного сопряжения: если $z = a + bi$, то $\bar{z} = a - bi$. Мы знаем, что сопряженное суммы есть сумма сопряженных и то же для произведения:

$$\overline{(z_1 + z_2)} = \bar{z}_1 + \bar{z}_2, \quad \overline{(z_1 \cdot z_2)} = \bar{z}_1 \cdot \bar{z}_2.$$

Отсюда легко заключить, что если α есть комплексный корень уравнения с действительными коэффициентами:

$$x^2 + ax + b = 0,$$

то $\bar{\alpha}$ тоже будет корнем этого уравнения.

Мы можем определить сопряжение и в квадратичном расширении поля из p элементов формулой

$$\overline{(a + b \cdot j)} := a - b \cdot j.$$

Ясно, что сопряженное суммы есть сумма сопряженных, сопряженное произведения есть произведение сопряженных. Кроме того, имеет место следующая замечательная формула.

Пусть p — простое число и t отрицательно по модулю p, т. е. $\left(\frac{t}{p}\right) = -1$. Построим комплексные числа как числа вида $a + b \cdot j$, где a и b рассматриваются по модулю p и $j^2 = t$.

Предложение 1. *В этих условиях если $z = a + bj$ и $\bar{z} = a - bj$, то*

$$\bar{z} = z^p. \tag{1}$$

В частности, $z^{p+1} = z\bar{z} = a^2 - tb^2$, т. е. $(p+1)$-я степень «комплексного» числа обязательно будет «действительным» числом.

Для доказательства формулы (1) давайте вспомним, что для наших чисел

$$(x + y)^p =_{(p)} x^p + y^p.$$

Это следует из того, что коэффициенты бинома Ньютона, $\binom{p}{i} = \frac{p!}{i!(p-i)!}$, являются целыми числами, делящимися на p при $0 < i < p$. Тогда мы можем написать

$$(a + b \cdot j)^p =_{(p)} a^p + b^p \cdot j^p.$$

Используя малую теорему Ферма, мы заключаем, что $a^p =_{(p)} a$, $b^p =_{(p)} b$. Остается вычислить j^p. Конечно,

$$j^p = j^{p-1} \cdot j = t^{(p-1)/2} \cdot j.$$

Нам нужно показать, что для отрицательного элемента t выполняется равенство $t^{(p-1)/2} =_{(p)} -1$. Отметим, что число $(p-1)/2$ целое, и если s является положительным элементом, то $s = a^2$ и

$$s^{(p-1)/2} =_{(p)} a^{p-1} =_{(p)} 1,$$

где последнее равенство следует из теоремы Ферма. Тем самым положительные элементы предоставляют нам $(p-1)/2$ корней полиномиального уравнения

$$x^{(p-1)/2} = 1$$

в поле «элементов по модулю p». Полиномиальное уравнение, по теореме Безу, не может иметь больше корней, чем его степень, и тем самым для отрицательного элемента t имеем

$$t^{(p-1)/2} \neq_{(p)} 1.$$

В то же время $t^{p-1} =_{(p)} 1$, и так как

$$t^{p-1} - 1 =_{(p)} (t^{(p-1)/2} - 1)(t^{(p-1)/2} + 1),$$

то остается единственная возможность: $t^{(p-1)/2} =_{(p)} -1$. Это завершает доказательство формулы.

Следствие. *Пусть p простое и 5 отрицательно по модулю p. Тогда для $\alpha = \dfrac{1+\sqrt{5}}{2}$ и $\beta = \dfrac{1-\sqrt{5}}{2}$ имеем*:

1) $\alpha^p =_{(p)} \beta$, $\beta^p =_{(p)} \alpha$;
2) $\alpha^{p+1} =_{(p)} \beta^{p+1} =_{(p)} \alpha \cdot \beta =_{(p)} -1$.

Мы можем применить этот результат к числам Фибоначчи и Лукаса. В этих условиях получаем:

$$u_{p+1} = \frac{\alpha^{p+1} - \beta^{p+1}}{\alpha - \beta} \equiv 0 \pmod{p};$$
$$v_p = \alpha^p + \beta^p \equiv \alpha + \beta \equiv 1 \pmod{p}.$$

Чтобы пользоваться этими сравнениями, нам надо уметь определять, для каких p «5» будет положительным и для каких отрицательным. Сейчас мы попробуем в этом разобраться.

5. Квадратный корень из 5 по модулю p

Свойство, которое я хочу сейчас сформулировать, легко следует из более общих и довольно глубоких результатов о символе Лежандра, которые объединяются под названием квадратичного закона взаимности. Нам нужен только частный случай этого общего закона, обнаруженного Эйлером

и Лежандром, доказанного Гауссом, и являющегося одной из жемчужин «элементарной» теории чисел.

Предложение 2.

$$\left(\frac{5}{p}\right) = \begin{cases} +1, & \text{если } p \equiv \pm 1 \pmod 5; \\ -1, & \text{если } p \equiv \pm 2 \pmod 5. \end{cases}$$

Сначала две общих леммы.

Лемма 1 (Лежандр).

$$a^{(p-1)/2} \equiv \left(\frac{a}{p}\right) \pmod p.$$

Фактически это значит, что для положительных a, $(p-1)/2$-я степень равна $+1$, а для отрицательных равна -1 по модулю p. Это мы разобрали в предыдущем пункте.

Заметим, что любое ненулевое число по модулю p равно с точностью до знака одному из чисел $1, 2, \ldots, (p-1)/2$. Если обозначить через $\mathscr{P}$ множество этих чисел:

$$\mathscr{P} = \{1, 2, \ldots, (p-1)/2\},$$

то для любого ненулевого x по модулю p либо $x \in \mathscr{P}$, либо $-x \in \mathscr{P}$. Фиксируем p и некоторое $a \neq_{(p)} 0$.

Лемма 2 (Гаусс). *Пусть для $k = 1, 2, \ldots, (p-1)/2$ число ε_k равно $+1$ или -1 и выбрано так, что $a \cdot k \cdot \varepsilon_k \in \mathscr{P}$ по модулю p. Тогда*

$$\left(\frac{a}{p}\right) = \prod_{k=1}^{(p-1)/2} \varepsilon_k.$$

Действительно, во-первых заметим, что если числа k' и k'' различны, то произведения $a \cdot k' \cdot \varepsilon_{k'}$ и $a \cdot k'' \cdot \varepsilon_{k''}$ тоже будут различны. Они могли бы совпадать только если $a \cdot k' =_{(p)} a \cdot k''$ или $a \cdot k' =_{(p)} -a \cdot k''$, но и первое и второе невозможно. Тогда, когда k пробегает все множество $\mathscr{P}$, то и произведение $a \cdot k \cdot \varepsilon_k$ пробегает все $\mathscr{P}$. Пусть K есть произведение всех элементов из $\mathscr{P}$. Мы имеем:

$$K = \prod_{k=1}^{(p-1)/2} a \cdot k \cdot \varepsilon_k =_{(p)} a^{(p-1)/2} \cdot K \cdot \prod_{k=1}^{(p-1)/2} \varepsilon_k.$$

Сокращая на K, получаем, что $1 =_{(p)} a^{(p-1)/2} \cdot \prod_{k=1}^{(p-1)/2} \varepsilon_k$, что с учетом леммы Лежандра доказывает лемму Гаусса.

Замечание 1. Можно сказать, что мы здесь моделируем одно из известных доказательств малой теоремы Ферма.

Теперь мы можем обратиться к нашему предложению. У нас будет $a=5$. Для нечетного p

$$p \equiv \pm 1 \pmod 5 \iff p = 10n+1 \quad \text{или} \quad p = 10n+9,$$
$$p \equiv \pm 2 \pmod 5 \iff p = 10n+3 \quad \text{или} \quad p = 10n+7.$$

Давайте применим лемму Гаусса для $p = 10n+1$. Здесь $(p-1)/2 = 5n$, и нам нужны $k = 1, 2, \ldots, 5n$.

При $k = 1, 2, \ldots, n$:

$$5k = 5, 10, \ldots, 5n \quad \text{и} \quad \varepsilon_k = +1.$$

При $k = n+1, \ldots, 2n$:

$$5k = 5n+1, \ldots, 10n \quad \text{и} \quad \varepsilon_k = -1.$$

При $k = 2n+1, \ldots, 3n$:

$$5k = (10n+1)+4, \ldots, (10n+1)+5(n-1)+4 \quad \text{и} \quad \varepsilon_k = +1.$$

При $k = 3n+1, \ldots, 4n$ аналогичным образом $\varepsilon_k = -1$, и при $k = 4n+1, \ldots, 5n$ снова $\varepsilon_k = +1$.

Тем самым -1 встречается $2n$ раз и $\prod \varepsilon_k = +1$. То есть если $p = 10n+1$, то $\left(\dfrac{5}{p}\right) = +1$.

Для случая $p = 10n+3$ можно рассуждать совершенно аналогично. Здесь $(p-1)/2 = 5n+1$.

При $k = 1, \ldots, n \quad \varepsilon_k = +1$.

При $k = n+1, \ldots, 2n \quad \varepsilon_k = -1$.

При $k = 2n+1, \ldots, 3n \quad \varepsilon_k = +1$.

При $k = 3n+1, \ldots, 4n \quad \varepsilon_k = -1$.

Теперь при $k = 4n+1$, $5k = 20n+5 = (10n+3)+(10n+2)$, что дает $\varepsilon_k = -1$.

При $k = 4n+2, \ldots, 5n+1 \quad \varepsilon_k = +1$.

В результате мы получаем на один элемент «-1» больше, всего $2n+1$ «минус единиц», а значит, $\left(\dfrac{5}{p}\right) = -1$ в этом случае. Мы предоставляем читателям самостоятельно проверить два оставшихся случая и будем считать предложение 2 доказанным.

6. Доказательство основной теоремы

Возвращаясь к доказательству основной теоремы, сформулированной в конце раздела 2, заметим прежде всего, что мы можем вычислить значение $M \pmod 5$. Мы знаем, что $2^4 \equiv 1 \pmod 5$ и что

$$M = 2^q - 1 = 2^{4k+3} - 1 \equiv 2^3 - 1 \equiv 2 \pmod 5.$$

Запомним это: в условиях теоремы $M \equiv 2 \pmod 5$.

Предположим теперь, что M — простое число. Тогда $\left(\frac{5}{M}\right) = -1$ и мы можем применить лемму и следствие из пункта 4. В частности,

$$\alpha^{M+1} \equiv \beta^{M+1} \equiv -1 \pmod{M},$$

а значит, $v_{M+1} \equiv -2 \pmod{M}$.

Пусть $N = 2^{q-1} = \frac{M+1}{2}$, т. е. $M + 1 = 2N$. Заметим, что

$$(v_N)^2 = (\alpha^N + \beta^N)^2 = \alpha^{2N} + \beta^{2N} + 2(\alpha\beta)^N = v_{2N} + 2 \cdot (-1)^N. \qquad (1)$$

Мы знаем, что N четно, значит,

$$(v_N)^2 = v_{2N} + 2 \equiv -2 + 2 \equiv 0 \pmod{M}.$$

Тем самым мы получили утверждение теоремы в этом случае.

Обратно, пусть известно, что $v_N \equiv 0 \pmod{M}$. Надо доказать, что число M простое. Во всяком случае, мы можем утверждать (поскольку $M \equiv 2 \pmod 5$), что не все простые делители p числа M имеют вид $p \equiv \pm 1 \pmod 5$; найдется простой делитель p числа M, для которого $p \equiv \pm 2 \pmod 5$, и тем самым $\left(\frac{5}{p}\right) = -1$, поэтому число 5 отрицательно по модулю p и мы можем использовать результаты пункта 4. В частности, $\alpha^{p+1} =_{(p)} \beta^{p+1} =_{(p)} -1$. Раз p делит M и $v_M \equiv 0 \pmod{M}$, значит,

$$v_N = \alpha^N + \beta^N =_{(p)} 0.$$

Пусть $\varepsilon = \alpha/\beta$. Тогда мы получим, с одной стороны,

$$\varepsilon^N =_{(p)} -1, \qquad (2)$$

а с другой стороны, $\varepsilon^{p+1} =_{(p)} 1$. Из этих двух равенств следует, что $p + 1 = 2N$, т. е. $p = M$.

Действительно, из равенства (2) следует, что $\varepsilon^{2N} = \varepsilon^{2^q} =_{(p)} +1$.

Лемма 3. *Пусть* $\varepsilon^a = 1$ *и* $\varepsilon^b = 1$. *Используя деление с остатком, запишем* $a = b \cdot c + r$. *Тогда* $\varepsilon^r = 1$.

Действительно, $1 = \varepsilon^a = (\varepsilon^b)^c \cdot \varepsilon^r = 1 \cdot \varepsilon^r = \varepsilon^r$. Пусть d минимальное положительное число, для которого $\varepsilon^d =_{(p)} 1$. Тогда, ввиду леммы, d делит $2N$ (значит $d = 2^s$) и d делит $p+1$. Если $s = q$, то $p + 1 = 2^q$. Иначе $s < q$, тогда d делит $N = 2^{q-1}$, значит, $\varepsilon^N = (\varepsilon^d)^{N/d} =_{(p)} 1$, что противоречит равенству (2). Теорема доказана.

Таким образом, простота числа $M = 2^q - 1$ зависит от значения числа $v_{2^{q-1}}$ по модулю M.

Замечательным образом мы можем использовать формулу (1) для вычисления чисел v_{2^i}. Обозначим $r_i = v_{2^i}$. Тогда $r_0 = v_1 = 1$. Теперь из формулы (1) имеем: $r_1 = r_0^2 + 2 = 3$ (здесь N нечетно). При $i \geqslant 1$ мы применяем

формулу (1) с четным N:

$$r_{i+1} = r_i^2 - 2, \quad r_1 = 3.$$

И наш основной результат принимает следующий вид.

Теорема 2. *Если q — простое число вида $4k+3$, то число $M = 2^q - 1$ простое если и только если $r_{q-1} \equiv 0 \pmod{M}$.*

7. Организация вычислений. Примеры

Вычисление r_i удобно производить в двоичной записи: $r_1 = 11$ (в двоичной системе).

Для r_2 получаем:

$$\begin{array}{rr} & \times\ 1\ 1 \\ & 1\ 1 \\ \hline & 1\ 1 \\ + & 1\ 1 \\ \cdots\cdots & \\ - & 1\ 0 \\ \hline r_2 = & 1\ 1\ 1 \end{array}$$

Таким образом, $r_2 = 111$, т. е. 7 в десятичной системе.

Для r_3:

$$\begin{array}{rr} & \times\ 1\ 1\ 1 \\ & 1\ 1\ 1 \\ \hline & 1\ 1\ 1 \\ & 1\ 1\ 1 \\ 1\ 1\ 1 & \\ \cdots\cdots\cdots & \\ - & 1\ 0 \\ \hline r_3 = 1\ 0\ 1\ 1\ 1\ 1 & \end{array}$$

Итак, $r_3 = 101111$, т. е. 47 в десятичной системе.

Здесь мы уже видим один случай нашей теоремы: при $q = 3$, $M = 7$ — простое, и как раз $r_2 = 7 \equiv 0 \pmod 7$.

Следующим q будет $q = 7$, $M = 2^7 - 1 = 127$. Конечно, можно использовать обычные деления для анализа простоты числа 127, но посмотрим, как работает алгоритм.

Нам надо посчитать r_4, r_5 и $r_6 \pmod{127}$. Приятно заметить, что мы можем производить редукцию по модулю 127 уже в процессе вычислений, и она соответствует «сдвигу двоичной записи» на 7 единиц:

$$2^7 \equiv 1 \pmod{2^7 - 1}, \quad \text{значит,} \quad 2^{7+k} \equiv 2^k \pmod{2^7 - 1}.$$

Тем самым r_4 по модулю 127 можно считать следующим образом:

$$\begin{array}{rr}
 & :\ 1\,0\,1\,1\,1\,1 \\
 & :1\,0\,1\,1\,1\,1 \\
+ & 1:0\,1\,1\,1\,1 \\
 & 1\,0:1\,1\,1\,1 \\
1\,0\,1\,1:1\,1 & \\
\hline
- & 1\,0
\end{array}$$

После переноса получаем

$$\begin{array}{r}
:\ 1\,0\,1\,1\,1\,1 \\
:1\,0\,1\,1\,1\,1\,0 \\
:0\,1\,1\,1\,1\,0\,\mathbf{1} \\
:1\,1\,1\,1\,0\,\mathbf{1\,0} \\
:1\,1\,0\,\mathbf{1\,0\,1\,1} \\
\hline
- \qquad\qquad 1\,0
\end{array}$$

Теперь мы должны считать «циклически» — перенося любую вылезающую влево за 7 разрядов единичку направо. Получаем:

$$\begin{array}{rr}
 & 1\,0\,1\,1\,0\,1 \\
 & 1\,0\,1\,1\,1\,1\,0 \\
+ & 0\,1\,1\,1\,1\,0\,1 \\
 & 1\,1\,1\,1\,0\,1\,0 \\
 & 1\,1\,0\,1\,0\,1\,1 \\
\hline
 & 0\,1\,1\,0\,0\,0\,0
\end{array}$$

Таким образом, $r_4 \equiv 0110000 \pmod{127}$.

Теперь для r_5:

$$\begin{array}{rr}
+ & 0\,0\,0\,0\,1\,1\,0 \\
 & 0\,0\,0\,1\,1\,0\,0 \\
\hline
 & 0\,0\,1\,0\,0\,0\,0
\end{array}$$

То есть $r_5 \equiv 2^4 \pmod{127}$. Теперь $r_6 \equiv 2^8 - 2 \equiv 2 - 2 \equiv 0 \pmod{127}$. Тем самым 127 — простое число.

Аналогичным образом было посчитано, что число $M = 2^{127} - 1$ простое. Только тут надо было осуществлять циклические сложения двоичных чисел длины 127. Как объясняет Вильямс [W], Лукас сделал себе шахматную доску и записывал числа по линиям этой доски, расставляя ладьи на местах единиц и оставляя пустыми клетки нулей. Циклические сложения можно тогда осуществлять как «игру», следуя нескольким простым правилам. Потребовалось примерно 100 часов такой игры, чтобы вычислить r_{127} по модулю $2^{127} - 1$.

Литература

[B] J. M. Bruce. A really trivial proof of the Lucas—Lehmer test // Amer. Math. Monthly **100** (1993), P. 370—371.

[R] M. I. Rosen. A proof of the Lucas—Lehmer test // Amer. Math. Monthly **95** (1988), P. 855—856.

[W] H. C. Williams. Édouard Lucas and primality testing, Canadian Math. Soc. Monographs **22**, John Wiley & Sons, N. Y., 1998.

Стивен Смейл

О проблемах вычислительной сложности

Лекция 20 мая 1999 года

Мы сейчас обсудим одну задачу, которая в элементарной форме иллюстрирует основные трудности теории вычислительной сложности. Для полинома $f \in \mathbb{Z}[t]$ определим число $\tau(f)$ следующим образом. Рассмотрим последовательность $(1, t, u_1, \ldots, u_m = f)$, в которой каждый последующий член получается из некоторых двух предшествующих: $u_k = u_i \circ u_j$, $i, j < k$; под операцией $\circ$ здесь подразумевается одна из трех арифметических операций (сложение, вычитание, умножение). Инвариант $\tau(f)$ равен наименьшему возможному m.

Имеется следующая гипотеза Шуба—Смейла: *количество различных целых корней многочлена f не превосходит $\tau(f)^c$, где c — некоторая абсолютная константа.*

Пример. Последовательность $1, t, t^2, t^{2^2}, \ldots, t^{2^k}, t^{2^k} - 1$ показывает, что $\tau(t^{2^k} - 1) \leqslant k + 1$. Но при этом количество различных корней многочлена $t^{2^k} - 1$ равно 2^k. Поэтому для различных комплексных корней аналогичная гипотеза неверна.

Аналогичный пример можно построить с помощью многочленов Чебышева. Многочлены Чебышева вычисляются с помощью простой рекуррентной формулы. Они тоже дают пример многочленов высокой степени с малым τ. При этом все корни многочленов Чебышева вещественны и попарно различны. Для различных вещественных корней аналогичная гипотеза тоже неверна.

Теорема 1 (Шуб—Смейл). *Из гипотезы Шуба—Смейла следует, что* $\mathrm{P} \neq \mathrm{NP}/_{\mathbb{C}}$.

Теперь нужно объяснить, что означает $\mathrm{P} \neq \mathrm{NP}/_{\mathbb{C}}$.

Прежде всего отметим, что у алгебраистов нетривиальные проблемы обычно начинаются с диофантовых уравнений, соответствующих

Стивен Смейл (Stephen Smale), профессор Калифорнийского Университета (США).

алгебраическим кривым, т. е. двум переменным. У нас проблемы начинаются уже в случае одной переменной.

Если забыть про $\mathbb{C}$, то проблема $\mathrm{P} \neq \mathrm{NP}$ — это одна из ключевых проблем компьютерной математики. Вместе с гипотезой Пуанкаре и гипотезой о нулях дзета-функции Римана она является одной из важнейших проблем математики — это подарок от computer science.

Рассмотрим многочлены $f_1(z_1, \ldots, z_n), \ldots, f_k(z_1, \ldots, z_n)$ над $\mathbb{C}$. Спрашивается, имеют ли эти многочлены общий нуль? Это — задача распознавания свойства: в качестве условия задаются многочлены $f_1, \ldots, f_k$ (точнее говоря, задается несколько комплексных чисел — коэффициентов многочленов), а результатом работы должен быть один из двух ответов: «да» (есть общий нуль) или «нет» (общего нуля нет).

Теорема Гильберта о нулях дает следующий ответ: общего нуля не существует тогда и только тогда, когда существуют такие многочлены $g_1, \ldots, g_k$, что $\sum g_i f_i = 1$.

Теорема Гильберта о нулях — критерий, но не метод. Она не дает никакого алгоритма. Но примерно 10 лет назад Браунвелл[1)] показал, что в теореме Гильберта о нулях можно считать, что

$$\deg g_i \leqslant \max(3, \max \deg f_i)^n,$$

причем этот результат неулучшаем.

Теорема Браунвелла дает алгоритм: все сводится к решению системы линейных уравнений для коэффициентов многочленов g_i.

Займемся теперь вопросом о скорости этого алгоритма: сколько арифметических операций нужно выполнить, чтобы ответить на поставленный вопрос. Назовем *размером* входных данных количество коэффициентов многочленов f_i, а *временем* работы алгоритма назовем количество арифметических операций. Будем называть данный алгоритм *алгоритмом с полиномиальным временем*, если

$$\text{время} \leqslant (\text{размер})^C, \tag{1}$$

где C — некоторая константа.

Алгоритмы с полиномиальным временем — это как раз те алгоритмы, которые имеет смысл практически реализовывать на вычислительной машине. Если, скажем, время зависит от размера экспоненциально, то при увеличении размера входных данных время быстро выходит за разумные пределы. Алгоритм Браунвелла является алгоритмом с экспоненциальным временем. Экспоненциальная верхняя оценка для этого алгоритма легко выводится, например, из гауссова метода исключения для решения системы линейных уравнений.

[1)] W. Brownawell. Bounds for the degrees in the Nullstellensatz // Annals of Math. **126** (1987), № 3. P. 577–591.

Гипотеза такова: задача $HN/_{\mathbb{C}}$ (существует ли общий нуль системы полиномиальных уравнений над $\mathbb{C}$) трудноразрешима, т. е. не существует алгоритма с полиномиальным временем для решения этой задачи.

Здесь имеется в виду алгоритм не в смысле машины Тьюринга, а алгоритм над $\mathbb{C}$ в следующем смысле. Алгоритм — это ориентированный граф с одной вершиной, в которую не ведет ни одного ребра (*входом*). Граф может иметь циклы. Он задает работу вычислительной машины следующим образом. На вход подается бесконечная в обе стороны последовательность комплексных чисел $(\dots, 0, z_1, \dots, z_n, 0, \dots)$, среди которых только $z_1, \dots, z_n$ отличны от нуля, никаких ограничений на величину n не предполагается, так что такая модель вычислительной машины может работать со сколь угодно длинными последовательностями чисел. Вершины этого графа относятся к одному из трех типов:

- *Выходы*. Из них не ведет ни одного ребра. По достижении такой вершины работа заканчивается.

- *Вычислительный узел*. В вычислительный узел входит одно ребро и из него выходит тоже одно ребро. В вычислительном узле производится арифметическая операция с какими-то членами последовательности и один из членов последовательности заменяется на результат вычислений. Кроме того, можно все члены последовательности умножить на одно и то же число или произвести сдвиг последовательности.

- *Узел ветвления*. В узел ветвления входит одно ребро, а выходят из него два ребра, на которых стоят пометки «да» и «нет». В узле ветвления выясняется, верно ли что $z_i = 0$. Если $z_i = 0$, то мы идем дальше по ребру с пометкой «да», а если $z_i \neq 0$, то мы идем дальше по ребру с пометкой «нет». (Для вычислений над $\mathbb{R}$ вместо этого можно ввести проверку типа $x_i > 0$ или $x_i \geqslant 0$.)

На выходе алгоритма тоже получается последовательность чисел. В интересующем нас алгоритме $HN/_{\mathbb{C}}$ на выходе только один ненулевой элемент, который может принимать ровно два значения, соответствующие ответам «да» и «нет».

Такое определение алгоритма было дано Л. Блюм, С. Смейлом и М. Шубом в конце 80-х годов. Странно, что раньше никто не додумался до этого совершенно естественного определения. Подробно ознакомиться с теорией таких алгоритмов можно по книге L. Blum, F. Cucker, M. Shub., S. Smale. Complexity and Real Computation. Springer Verlag, 1997.

С описанным выше алгоритмом естественным образом связана функция «входа—выхода». Она определена на некотором множестве входных

данных (например, машина не может производить деление на нуль, поэтому при некоторых входных данных она останавливается).

Размером входных данных назовем число n, а временем работы при данном входе — длину пути от входа к выходу (на разных входах эти пути могут быть различными). Алгоритмы с полиномиальным временем удовлетворяют неравенству (1) для некоторой константы C при всех входах. Класс таких алгоритмов обозначается $\mathrm{P}/_{\mathbb{C}}$.

После этого определения вопрос о том, существует ли полиномиальный алгоритм для задачи $HN/_{\mathbb{C}}$, приобретает строгий математический смысл. Заметим, что утверждение о том, что такого алгоритма не существует, в точности эквивалентно утверждению $\mathrm{P} \neq \mathrm{NP}/_{\mathbb{C}}$ (определение $\mathrm{NP}/_{\mathbb{C}}$ пока не давалось и на этой лекции не будет дано).

Вместо поля $\mathbb{C}$ можно взять произвольное поле K и определить вычислительную машину над произвольным полем. Например, поле $K = \mathbb{Z}_2$ соответствует определению алгоритма, принятому в логике и computer science.

Можно также поставить вопрос об общих нулях многочленов над $\mathbb{Z}_2$. Гипотеза о том, что не существует полиномиального алгоритма для решения этой задачи, эквивалентна гипотезе $\mathrm{P} \neq \mathrm{NP}$ в ее классическом варианте.

Для числа $m \in \mathbb{Z}$ можно определить инвариант $\tau(m)$ по аналогии с инвариантом τ для многочленов. А именно, рассмотрим аналогичную последовательность $(1, m_1, \ldots, m_k = m)$ и определим $\tau(m)$ как минимальное возможное k. С помощью формулы Стирлинга можно доказать, что $\tau(m!) \leqslant (\ln m)^C$. Есть предположение, что верна и противоположная оценка такого же вида: $(\ln m)^{C'} \leqslant \tau(m!)$; эта проблема связана с разложением на простые множители.

На первый взгляд эти две проблемы (об инварианте τ для полиномов и для чисел) друг с другом не связаны.

Вернемся к проблеме $\mathrm{P} \neq \mathrm{NP}/_K$. Мы не будем определять, что такое $\mathrm{NP}/_K$, вместо этого будем говорить об эквивалентной проблеме Гильберта о нулях над произвольным полем K: $HN/_K \notin \mathrm{P}/_K$. (Если поле не алгебраически замкнуто, то теорема Гильберта о нулях неверна, но задача об общих нулях системы многочленов имеет смысл над произвольным полем; здесь имеется в виду именно эта задача.)

В случае не алгебраически замкнутого поля доказано следующее утверждение.

Теорема 2. *Если поле K не алгебраически замкнуто и* $\operatorname{char} K = 0$, *то* $\mathrm{P} \neq \mathrm{NP}/_K$.

Для поля $\mathbb{Z}_2$, которое тоже не алгебраически замкнуто, вопрос остается открытым (характеристика этого поля отлична от нуля).

Вернемся к алгебраически замкнутым полям. Для алгебраически замкнутого поля K ($\operatorname{char} K = 0$) проблема $\mathrm{P} \neq \mathrm{NP}/_K$ эквивалентна проблеме $\mathrm{P} \neq \mathrm{NP}/_{\mathbb{C}}$. Поэтому все сводится к одному полю, например, полю $\mathbb{C}$ или полю $\overline{\mathbb{Q}}$ (так обозначается алгебраическое замыкание поля $\mathbb{Q}$). Это — один из основных результатов упомянутой выше книги. Доказательство использует понятие высоты алгебраического числа.

Представляется весьма правдоподобным, что $\mathrm{P}/_K = \mathrm{P}/_{\mathbb{F}_2}$ для любого конечного поля K. Но для полей конечной характеристики вопросов больше, чем ответов.

Рассмотрим вопрос об эквивалентности проблем над $\mathbb{C}$ и над $\overline{\mathbb{Q}}$. Одна из основных проблем при переходе от комплексных чисел к алгебраическим связана с тем, что нужно избавиться от комплексных констант, которые могут использоваться при вычислениях. Вообще говоря, они могли бы сильно упростить вычисления. Однако доказано, что такого упрощения не происходит.

В упомянутой выше книге не рассматривался вопрос о связи проблем $\mathrm{P} \neq \mathrm{NP}/_{\mathbb{Z}_2}$ и $\mathrm{P} \neq \mathrm{NP}/_{\mathbb{C}}$. Именно первая из них относится к классической computer science. В предисловии к этой книге Дик Карп высказал предположение, что эти проблемы никак друг с другом не связаны. Но уже после того как книга была написана, Смейл заметил следующее.

Напомним, что интерес к алгоритмам с полиномиальным временем связан с тем, что именно их можно эффективно реализовывать на компьютерах. Но сейчас используется еще один важный класс алгоритмов — так называемые BPP-алгоритмы. В этих алгоритмах разрешается «подбрасывать монетку» и в зависимости от полученного результата производить те или иные вычисления. Требуется, чтобы правильный ответ получался в «квалифицированном большинстве» случаев. Тогда, повторив вычисления много раз, можно получить результат, который будет правильным с очень большой вероятностью. Например, если правильный результат получается с вероятностью 3/4, то после 50 повторений вычислений вероятность ошибки будет равна единице, деленной на число атомов во Вселенной.

С математической точки зрения условие BPP накладывает меньше ограничений, чем условие P, но с практической точки зрения BPP-алгоритмы столь же хороши, как и P-алгоритмы.

Теорема 3 (Смейл). *Если* $\mathrm{BPP} \not\supseteq \mathrm{NP}$, *то* $\mathrm{P} \neq \mathrm{NP}/_{\mathbb{C}}$.

С точки зрения современной computer science $\mathrm{BPP} \not\supseteq \mathrm{NP}$ — это нечто очень похожее на $\mathrm{P} \neq \mathrm{NP}$.

Пьер Картье

Значения ζ-функции

Лекция 21 мая 1999 года

1. Значения ζ-функции

Лекция будет состоять из следующих частей:

1. Значения ζ-функции.

2. Полилогарифмические функции.

3. Обобщения полилогарифмических функций и множественные значения ζ-функции (MZV-числа).

4. Гипотезы о природе некоторых чисел.

Начало будет очень классическим. Я начну с вычисления суммы ряда $\sum\limits_{n=1}^{\infty}\frac{1}{n^2}$. В 1739 г. Эйлер доказал, что сумма этого ряда равна $\frac{\pi^2}{6}$. Сначала он провел вычисления частных сумм, а затем угадал ответ. Следует отметить, что ряд сходится очень медленно. Поэтому Эйлеру для вычисления суммы этого ряда пришлось разработать специальные численные методы.

Эйлер придумал так называемое «доказательство», которое я повторю в нескольких словах. Пусть $P(x)=c_Nx^N+c_{N-1}x^{N-1}+\ldots+c_0$ — многочлен степени N, $\xi_1,\ldots,\xi_N$ — его корни. Тогда $\xi_1\ldots\xi_N=(-1)^N\frac{c_0}{c_N}$ и $\sum\limits_{i=1}^{N}\xi_1\ldots\hat{\xi}_i\ldots\xi_N=(-1)^{N-1}\frac{c_1}{c_N}$. Из этих двух равенств следует, что $\sum\limits_{i=1}^{N}\frac{1}{\xi_i}=$ $=-\frac{c_1}{c_0}$. По-другому это можно доказать, рассмотрев многочлен с корнями $\frac{1}{\xi_i}$. Можно также доказать формулу $\sum\limits_{i<j}\frac{1}{\xi_i\xi_j}=\frac{c_2}{c_0}$. Поэтому, если $c_1=0$, то $\sum\limits_{i=1}^{N}\frac{1}{\xi_i^2}=-2\frac{c_2}{c_0}$. Эйлер, разумеется, знал все эти формулы.

Предположим теперь, что есть «полином» с корнями 1, 2, 3, … Тогда с его помощью можно вычислить $\sum\limits_{n=1}^{\infty}\frac{1}{n^2}$. Эйлера при этом не смущало,

Пьер Картье (Pierre Cartier), профессор IHES (Франция).

что этот «полином» имеет бесконечно много корней. Рассмотрим функцию $\frac{\sin \pi x}{\pi x} = s(x)$. Корни «полинома» $s(x)$ — это корни уравнения $\sin \pi x = 0$, $x \neq 0$. Положительные корни — это как раз то, что нам нужно. Но с учетом отрицательных корней мы должны получить удвоенную сумму ряда $\sum\limits_{n=1}^{\infty} \frac{1}{n^2}$.

Ряд для синуса показывает, что

$$s(x) = 1 - \frac{\pi^2}{6} x^2 + \dots,$$

Таким образом, $c_0 = 1$, $c_1 = 0$ и $c_2 = \frac{\pi^2}{6}$. Следовательно, $2 \sum\limits_{n=1}^{\infty} \frac{1}{n^2} = -2\frac{c_2}{c_0}$, а значит, $\sum\limits_{n=1}^{\infty} \frac{1}{n^2} = -\frac{c_2}{c_0} = \frac{\pi^2}{6}$.

Разумеется, это доказательство неудовлетворительно. Например, если применить такие же рассуждения к функции e^{x^2}, то получится абсурдный результат. Только Вейерштрасс (1860) и Адамар (1895) в этом в конце концов разобрались, рассматривая факторизацию целых функций комплексного переменного. Если наложить дополнительные условия на рост целой функции на бесконечности, то для таких функций рассуждения Эйлера приводят к правильному результату.

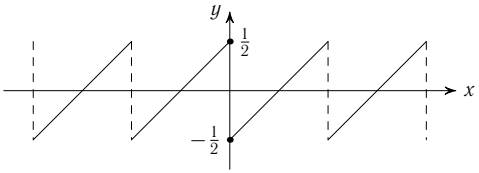


Рис. 1. Пилообразная функция

Есть также другое доказательство, которое легче выразить в терминах рядов Фурье. Ряды Фурье получили строгое обоснование лишь в XIX в., но и в XVIII в. ими пользовались достаточно успешно. Рассмотрим пилообразную функцию $\varphi(x)$ (рис. 1). Она периодична и $\varphi(x) = \frac{1}{2} - x$ при $x \in (0, 1)$. Ясно, что $\int\limits_0^1 \varphi(x)\, dx = 0$. Если ввести коэффициенты Фурье $c_n = \int\limits_0^1 \varphi(x) e^{-2\pi i n x}\, dx = \frac{1}{2\pi i n}$, то получим

$$\varphi(x) = \sum_n c_n e^{2\pi i n x} = \sum_{n \neq 0} \frac{e^{2\pi i n x}}{2\pi n}.$$

Теорема Парсеваля показывает, что

$$\int_0^1 |\varphi(x)|^2\,dx = \sum_{n\neq 0} |c_n|^2 = \frac{1}{2\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}.$$

С другой стороны,

$$\int_0^1 |\varphi(x)|^2\,dx = \int_0^1 \left|\frac{1}{2}-x\right|^2 dx = \frac{1}{12}.$$

Это уже настоящее доказательство.

В учебниках есть и другие доказательства. Например, можно воспользоваться формулой вычетов Коши. А именно, рассмотрим мероморфную функцию $\dfrac{1}{x^2(e^{2\pi ix}-1)}$. В точке $x=0$ эта функция имеет полюс порядка 3, а в точках $x=n\neq 0$ эта точка имеет полюсы порядка 1. Применив к этой функции формулу вычетов Коши, получим тот же самый результат.

Эйлер сделал большее: он вычислил суммы $1+2+3+\ldots$ и $1^2+2^2+3^2+\ldots$ Я не буду повторять этих вычислений. Они относятся к тому, что я называю «математией». Эти вычисления выполняются без особого обоснования и получаются некие результаты. После этого требуется много лет, чтобы их обосновать. В наше время такая же ситуация с фейнмановскими интегралами по траекториям. Их вычисляют, но никакого обоснования этих вычислений нет. Вычисления Фейнмана похожи на вычисления Эйлера. Они не имеют пока никакого обоснования, но когда-нибудь получат его.

Теперь я введу классическую дзета-функцию Римана $\zeta(s)=\sum\limits_{n=1}^{\infty}\dfrac{1}{n^s}$. Здесь предполагается, что $s\in\mathbb{C}$ и $\operatorname{Re} s>1$; в таком случае ряд абсолютно сходится. Мы уже вычислили, что $\zeta(2)=\pi^2/6$. Оба метода вычислений (с помощью многочленов и с помощью рядов Фурье) позволяют вычислить $\zeta(4)$ и дают один и тот же результат: $\zeta(4)=\pi^4/90$. Здесь обоснование вычислений с рядами Фурье даже проще, потому что в этом случае ряд Фурье сходится абсолютно. Аналогично можно доказать, что если k — натуральное число, то $\zeta(2k)=\pi^{2k}r$, где r — рациональное число. Но метод вычислений с рядами Фурье ничего не дает в случае $\zeta(3)$, $\zeta(5)$, … Единственное, что сейчас известно об этих числах, так это то, что $\zeta(3)\notin\mathbb{Q}$ (даже не известно, трансцендентно ли число $\zeta(3)$). Иррациональность числа $\zeta(3)$ была доказана Апери (R. Apéry) в 1978 г. Затем Don Zagier и H. Cohen упростили и прояснили доказательство Апери, и в августе 1978 г. Cohen представил доказательство Апери на Международном математическом конгрессе в Хельсинки.

1.1. Интегральное представление для функции $\zeta(s)$. Рассмотрим гамма-функцию $\Gamma(s) = \int_0^\infty e^{-x}x^{s-1}\,dx$; здесь $\operatorname{Re} s > 0$. Сделав замену $x = n\xi$, получим

$$n^{-s} = \frac{1}{\Gamma(s)}\int_0^\infty e^{-n\xi}\xi^{s-1}\,d\xi, \quad n = 1,\ 2,\ \ldots$$

Если мы рассмотрим сумму $\sum_{n=1}^{\infty} n^{-s}$, то в правой части получим выражение, содержащее геометрическую прогрессию, которую можно просуммировать. В итоге получаем формулу

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s} = \frac{1}{\Gamma(s)}\int_0^\infty \frac{x^{s-1}}{e^x - 1}\,dx. \tag{1}$$

Известен метод Адамара, хорошо разработанный в книге И. М. Гельфанда и Г. Е. Шилова «Обобщенные функции». Этот метод заключается в следующем. Рассмотрим интеграл вида $\Phi(s) = \int_0^\infty F(x)x^{s-1}\,dx$, где функция F класса C^∞, причем она разлагается в ряд Тэйлора даже в нуле и быстро убывает при $x \to \infty$ вместе со всеми производными.

Пример 1. $F(x) = e^{-x}$.

Функция $\Phi = \Phi_F$ продолжается на $\mathbb{C}$ как мероморфная функция с простыми полюсами в точках $0, -1, -2, \ldots$ Это легко доказать, интегрируя по частям. Действительно, при интегрировании по частям мы получаем функциональное уравнение, связывающее $\Phi_F(s)$, $\Phi_{F'}(s)$ и $\Phi_F(s-1)$; это функциональное уравнение является обобщением функционального уравнения для $\Gamma(s)$. Таким образом, функция $\Phi(s)/\Gamma(s)$ не имеет полюсов. Функция именно такого вида встречается в формуле (1). К сожалению, в интересующем нас случае функция $F(x) = \dfrac{1}{e^x - 1}$ имеет особенность в нуле, поэтому к ней нельзя применить эти рассуждения. Нужно изменить функцию, например, положить $F(x) = \dfrac{x}{e^x - 1}$ и воспользоваться разложением $x^{s-1} = x \cdot x^{s-2}$. В результате получим интегральное представление

$$\Gamma(s)\zeta(s) = \int_0^\infty \frac{x}{e^x - 1}\,x^{s-2}\,dx.$$

Мы сдвинули полюса на единицу: полюса функции $\Gamma(s)\,\zeta(s)$ находятся в точках $1, 0, -1, -2, \ldots$ Полюса функции $\Gamma(s)$ находятся в точках $0, -1, -2, \ldots$ Поэтому функция $\zeta(s)$ имеет единственный полюс в точке 1.

Общий результат такой: $\left(\dfrac{\Phi}{\Gamma}\right)(-k) = F^{(k)}(0)$. По-другому это можно выразить формулой $\dfrac{x^{s-1}}{\Gamma(s)}\Big|_{s=-k} = \delta^{(k)}(x)$, где $\delta^{(k)}$ — производная порядка k

функции Дирака. Эта точка зрения хорошо объяснена в упомянутой книге Гельфанда и Шилова. Применим это равенство к функции $\zeta(s)$. В результате получим

$$-k\zeta(1-k) = \frac{d^k}{dx^k}\left(\frac{x}{e^x-1}\right)\Big|_{x=0}; \qquad (2)$$

здесь мы учитываем сделанный ранее сдвиг на единицу.

Введем числа Бернулли B_k, которые определяются тождеством

$$\frac{x}{e^x-1} = \sum_{k=0}^{\infty} B_k \frac{x^k}{k!}.$$

Из формулы (2) получаем

$$\zeta(1-k) = -B_k/k, \quad k = 1, 2, 3, \ldots$$

Например, $\zeta(0) = -B_1 = 1/2$, $\zeta(-1) = -B_2/2 = -1/12$ и $\zeta(-2) = 0$. С формальной точки зрения $\zeta(-1) = 1+2+3+\ldots$ и $\zeta(-2) = 1^2+2^2+3^2+\ldots$ Это как раз те суммы, которые мы хотели вычислить.

Теперь еще немного математики. Формально положим $B_k = B^k = B \times \ldots \ldots \times B$ и рассмотрим e^{Bx}. Из равенства $\frac{x}{e^x-1} = e^{Bx}$ следует, что $x = (e^x-1)e^{Bx} = e^{(B+1)x} - e^{Bx}$. Поэтому при $n \neq 1$ выполняется равенство $(B+1)^n = B^n$. Например,

$$0 = (B+1)^2 - B^2 = 2B^1 + B^0 = 2B_1 + B_0.$$

Мы знаем, что $B_0 = 1$, поэтому $B_1 = -1/2$. Аналогично получаем равенство $3B_2 + 3B_1 + B_0 = 0$ и вычисляем B_2 и т. д.

Формально эту конструкцию можно описать так. Рассмотрим кольцо многочленов над полем $\mathbb{C}$ от одной переменной B. Определим линейное отображение $ev\colon \mathbb{C}[B] \to \mathbb{C}$ следующим образом: $ev(B^k) = B_k$; отображение ev задано на базисных элементах и продолжается по линейности. Отображение ev переводит формальный ряд от двух переменных B и x в формальный ряд от одной переменной x. После этого можно повторить приведенные выше вычисления, применяя в нужных местах отображение ev. Это объясняется в учебнике Бурбаки по элементарному анализу; в нем есть глава о числах Бернулли.

Можно также доказать, что верна формула

$$\zeta(2k) = (-1)^{k+1}\frac{(2\pi)^{2k}}{2\cdot(2k)!}B_{2k}.$$

Отметим, что $B_3 = 0$, $B_5 = 0$, ... Попытка вычислить значения дзета-функции в нечетных точках тем же самым методом приводит к равенству $0 = 0$, которое ничего не дает. Объяснение, почему так происходит, будет дано в конце второй части.

2. Полилогарифмические функции

Полилогарифмические функции задаются следующими равенствами:

$$\mathrm{Li}_k(z) = \sum_{n=1}^{\infty} \frac{z^n}{n^k}.$$

Это определение связано с тем, что $\mathrm{Li}_k(1) = \zeta(k)$. Мы надеемся получить информацию о дзета-функции, пользуясь полилогарифмическими функциями.

В комплексной области ряд для $\mathrm{Li}_k(z)$ сходится при $|z| < 1$. Первая задача — построить аналитическое продолжение. Сначала все очень просто:

$$\mathrm{Li}_0(z) = \sum_{n=1}^{\infty} z^n = \frac{z}{z-1}.$$

Это — рациональная функция с полюсом $z = 1$.

Мой учитель Анри Картан запрещал мне делать разрезы на комплексной плоскости, потому что такой подход неинвариантен. Он требовал всегда рассматривать только римановы поверхности. Если вы заглянете в его книгу по комплексному анализу (очень хорошую), то найдете там определение аналитического продолжения с помощью пучков или еще чего угодно, но только не с помощью разрезов. Тем не менее, я не вижу в разрезах плоскости никакого противоречия и поэтому сделаю разрез.

Рассмотрим открытое односвязное множество $U = \mathbb{C} \setminus [1, +\infty[$. Если функция Φ голоморфна на односвязном множестве, то на этом множестве у нее есть первообразная Ψ. Эта функция удовлетворяет нормировочному условию $\Psi(0) = 0$ и $\dfrac{d\Psi}{dz} = \Phi$. Функция Ψ тоже голоморфна в области U.

Легко проверить, что $z\dfrac{d}{dz}\mathrm{Li}_k(z) = \mathrm{Li}_{k-1}(z)$. Функция Li_0 голоморфна в области U и $\mathrm{Li}_0(0) = 0$, поэтому функция Li_1 тоже голоморфна в области U. Рассуждая дальше, получаем то же самое для Li_2, Li_3, ... Все эти функции аналитически продолжаются на U.

Чтобы идти дальше, нужно исследовать предельное поведение этих функций при подходе к разрезу с разных сторон: сверху и снизу. Я рассмотрю только случай

$$\mathrm{Li}_1(z) = z + \frac{z^2}{2} + \frac{z^3}{3} + \ldots = \ln\frac{1}{1-z}.$$

Значения логарифма сверху и снизу отличаются на одно и то же постоянное значение. Если γ_1 — монодромия при обходе вокруг 1, то $\gamma_1(\mathrm{Li}_1(z)) = 2\pi i$. Монодромия — это разность между двумя ветвями. Разность между двумя ветвями снова аналитически продолжается на всю плоскость. Эта монодромия подробно изучалась многими алгебраическими геометрами: Блохом, Делинем, Дринфельдом и другими.

Снова обращаясь к математии, зададим вопрос Эйлеру: как вычислить сумму ряда $\sum\limits_{n=-\infty}^{\infty} z^n$? Эйлер дает следующий ответ: эта сумма равна 0. Он рассуждает так. Рассмотрим сумму $z + z^2 + \ldots + z^n + \ldots = \dfrac{z}{1-z} = \mathrm{Li}_0(z)$. Она сходится при $|z| < 1$. Рассмотрим теперь сумму обратных величин: $z^{-1} + z^{-2} + \ldots + z^{-n} + \ldots = \dfrac{-1}{1-z} = \mathrm{Li}_0\left(\dfrac{1}{z}\right)$. Она сходится при $|z| > 1$. Но вспомните, что мы сделали аналитическое продолжение. Обе функции $\dfrac{z}{1-z}$ и $\dfrac{-1}{1-z}$ рациональные; они определены всюду, кроме полюса в точке 1. Чтобы вычислить требуемую сумму, к этим двум функциям нужно прибавить 1:

$$1 + \mathrm{Li}_0(z) + \mathrm{Li}_0\left(\frac{1}{z}\right) = 1 + \frac{z-1}{1-z} = 0.$$

Эти рассуждения можно пошагово обобщить и получить следующий результат. Функция $\mathrm{Li}_k(z)$ голоморфна вне $[1, +\infty[$, а функция $\mathrm{Li}_k\left(\dfrac{1}{z}\right)$ голоморфна вне $[0, 1]$. Поэтому функция $\mathrm{Li}_k(z) + (-1)^k \mathrm{Li}_k\left(\dfrac{1}{z}\right)$ голоморфна вне $[1, +\infty[\cup [0,\ 1] = [0, \infty[$.

Обычно логарифм определяют для разреза $]-\infty, 0]$, но его можно определить и для разреза $[0, \infty[$. Мы будем предполагать, что функция $\ln z$ определена именно с помощью разреза $[0, \infty[$, причем выберем ветвь этого логарифма так, что если подходить к этому разрезу сверху, то получим обычный вещественный логарифм.

При таком определении логарифма получаем следующую формулу:

$$\mathrm{Li}_k(z) + (-1)^k \mathrm{Li}_k\left(\frac{1}{z}\right) = \frac{(2\pi i)^k}{k!} B_k\left(\frac{\ln z}{2\pi i}\right). \qquad (1)$$

Здесь $B_k(t)$ — многочлен Бернулли. Неформально он определяется тождеством $B_k(t) = (B + t)^k$. Чтобы получить формальное определение, нужно воспользоваться отображением *ev*: сначала рассмотрим полином от переменных B и t, а затем каждый моном B^k заменим на B_k.

Многочлены Бернулли можно также задать следующими свойствами, которые их полностью характеризуют:

- $\dfrac{d}{dt} B_k(t) = kB_{k-1}(t)$;
- $B_k(t+1) - B_k(t) = kt^{k-1}$;
- $B_0(t) = 1$;
- $B_k(0) = B_k$.

Простое упражнение по алгебре — доказать, что этими свойствами однозначно задается некоторая последовательность многочленов.

Можно написать соответствующую производящую функцию:

$$\sum_{k=0}^{\infty} B_k(t)\,\frac{x^k}{k!} = \frac{xe^{xt}}{e^x - 1}.$$

В частности, при $t = 0$ получаем $B_k(0) = B_k$.

Если в формуле (1) положить $z = 1 + i\varepsilon$, $\varepsilon > 0$, и устремить ε к нулю, то получим

$$\zeta(k) + (-1)^k \zeta(k) = \frac{(2\pi i)^k}{k!}\, B_k.$$

Для нечетных k эта формула не дает никакой информации: получаем тождество $0 = 0$. А для четных k получаем ту самую формулу, о которой шла речь ранее.

3. Обобщения полилогарифмических функций

Мы видели, что $\mathrm{Li}_1(z) = \ln \dfrac{1}{1-z}$. В связи с этим Li_2 называют *дилогарифмом*, Li_3 называют *трилогарифмом* и т. д.

Мы хотим обобщить этот класс функций. При этом мы хотим включить в новый класс функций и обычный логарифм, а он определяется на комплексной плоскости с разрезом от $-\infty$ до 0. Именно таким стандартным определением логарифма мы теперь будем пользоваться, а не тем, которое мы использовали на с. 60.

Все функции, которые мы определим, будут голоморфны в комплексной плоскости с двумя разрезами — от 1 до $+\infty$ и от $-\infty$ до 0. Проще всего эти функции можно определить с помощью дифференциального уравнения. Напомню, что для полилогарифма мы доказали формулу

$$\partial_z \mathrm{Li}_k(z) = \frac{1}{z}\,\mathrm{Li}_{k-1}(z),$$

где ∂_z обозначает дифференцирование по z. Функция, которая будет определена, параметризуется индексами $k_1, \ldots, k_p$; она обозначается $\mathrm{Li}_{k_1,\ldots,k_p}(z)$.

Прежде чем перейти к разложению функции $\mathrm{Li}_{k_1,\ldots,k_p}(z)$ в ряд, я напишу для нее дифференциальное уравнение. Введем две некоммутирующие переменные X_0 и X_1. Рассмотрим последовательность $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_t)$, где $\varepsilon_i \in \{0, 1\}$. Каждой такой последовательности можно сопоставить произведение $X_\varepsilon = X_{\varepsilon_1} \ldots X_{\varepsilon_t}$. Например, $X_{010} = X_0 X_1 X_0$.

Рассмотрим дифференциальное уравнение

$$\partial_z \Lambda(z) = \left(\frac{X_0}{z} + \frac{X_1}{1-z}\right) \Lambda(z).$$

Например, если X_0 и X_1 — квадратные матрицы порядка p, то это уравнение представляет собой систему обыкновенных дифференциальных уравнений. Эта система голоморфна, но имеет особенности в точках 0 и 1. Решения этой системы нужно рассматривать в односвязной области, не содержащей точек 0 и 1. Например, можно взять комплексную плоскость с двумя разрезами, упомянутыми выше. В соответствии с общей теорией голоморфных дифференциальных уравнений в такой области система имеет голоморфное решение. Отметим, что уравнение Гаусса для гипергеометрической функции можно записать в таком виде при $p = 2$; матрицы X_0 и X_1 при этом являются некоторыми конкретными матрицами порядка 2. Решение уравнения есть вектор-функция — набор из двух функций, одна из которых и будет гипергеометрической функцией.

Вернемся к общим некоммутирующим переменным X_0 и X_1. Будем искать решения вида

$$\begin{aligned}\Lambda(z) &= \sum_{\varepsilon} \Lambda_\varepsilon(z) X_\varepsilon = \\ &= \Lambda_\varnothing(z) + \Lambda_0(z)X_0 + \Lambda_1(z)X_1 + \\ &\quad + \Lambda_{00}(z)X_0^2 + \Lambda_{01}(z)X_0X_1 + \Lambda_{10}(z)X_1X_0 + \Lambda_{11}(z)X_1^2 + \dots\end{aligned}$$

Если X_0 и X_1 — матрицы, то можно доказать, что этот бесконечный ряд сходится. Но сейчас меня не интересует вопрос сходимости; я рассматриваю формальный ряд. Решая дифференциальное уравнение для формальных рядов, получаем рекуррентные соотношения

$$\partial_z \Lambda_{0\varepsilon}(z) = \frac{1}{z}\Lambda_\varepsilon(z), \quad \partial_z \Lambda_{1\varepsilon}(z) = \frac{1}{1-z}\Lambda_\varepsilon(z).$$

Для простоты положим $\Lambda_\varnothing(z) = 1$. Тогда

$$\partial_z \Lambda_0(z) = \frac{1}{z}, \quad \partial_z \Lambda_{00}(z) = \frac{1}{z}\Lambda_0(z)$$

$$\partial_z \Lambda_1(z) = \frac{1}{1-z}, \quad \partial_z \Lambda_{10}(z) = \frac{1}{1-z}\Lambda_0(z), \dots$$

Следовательно, $\Lambda_0(z) = \ln z + C_0$; это выражение определено на комплексной плоскости с двумя разрезами. Далее,

$$\Lambda_{00}(z) = \frac{1}{2}\ln^2 z + C_0 \ln z + C_1.$$

Нормализацию решений зададим следующим образом: $\Lambda_{0\dots0}(z) = \dfrac{1}{p!}\ln^p z$.

Я воспользуюсь хорошо известным фактом, что интеграл $\displaystyle\int_0 x^\lambda \ln^p x\, dx$ сходится при $\lambda \geqslant 0$ (такая запись означает, что интеграл сходится в нуле). Единственная сингулярность в нуле происходит от логарифма.

Из другого уравнения получаем соотношение

$$\Lambda_1(z) = \ln \frac{1}{1-z}.$$

Здесь могла бы появиться константа, но я нормализую это следующим образом: при $z = 0$ функция обращается в нуль.

Если мы хотим полностью записать асимптотическое условие, нормализующее функцию Λ, то это можно сделать так. Запишем

$$\Lambda(z) = \tilde{\Lambda}(z) \exp(X_0 \ln z), \tag{1}$$

где

$$\exp(X_0 \ln z) = \sum_{p=0}^{\infty} \frac{1}{p!} \ln^p z \underbrace{X_0 \dots X_0}_{p}.$$

Здесь $\tilde{\Lambda}$ голоморфно в окрестности точки 0 и $\tilde{\Lambda}(0) = 1$. Этим определяется асимптотическое условие на функцию Λ.

Запишем исходное дифференциальное уравнение в виде

$$d\Lambda = \left(X_0 \frac{dz}{z} + X_1 \frac{dz}{1-z} \right) \Lambda$$

В окрестности нуля член $X_1 \dfrac{dz}{1-z}$ регулярен. Если же пренебречь этим регулярным членом, то решением уравнения будет как раз написанная выше экспонента. Формула (1) — это частный случай фуксовой формы решения дифференциального уравнения с особыми точками.

Разумеется, если мы заменим X_0 и X_1 матрицами, то нужно будет придать смысл всему тому, что было написано; в частности, нужно будет придать смысл экспонентам. Например, если $X_0 = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, то соответствующая функция равна $\operatorname{diag}(z^{\lambda_1}, \dots, z^{\lambda_n})$. Именно это возникает в обычной фуксовой теории. Одна из трудностей классической теории Фукса состоит в том, что она не работает в случае, когда одна из разностей $\lambda_i - \lambda_j$ является целым числом. Но у нас этой трудности не возникает, потому что мы рассматриваем формальные ряды.

Оказывается, что наши условия однозначно определяют функцию $\tilde{\Lambda}$: решение дифференциального уравнения с такими начальными условиями единственно.

Ситуация для $z = 0$ и для $z = 1$ симметрична. Действительно, отображение $z \mapsto 1 - z$ меняет местами два разреза и ситуация при этом получается та же самая. (Однако асимптотическое разложение функции $\Lambda(z)$ в точке $z = 1$ будет, конечно, другим.)

Запишем снова решение в виде бесконечного ряда

$$\Lambda(z) = \sum_{\varepsilon} \Lambda_\varepsilon(z) X_\varepsilon,$$

где ε — конечная последовательность нулей и единиц. Мне придется слегка изменить обозначения. Прежде всего я буду предполагать, что $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)$, где $\varepsilon_p = 1$. В таком случае легко убедиться, что Λ_ε не имеет особенностей в нуле: для тех членов, в правой части которых отсутствует X_0, после перемножения отрицательные степени z уничтожаются. Это и есть полилогарифмические функции, которые мы хотели определить.

Введем следующие обозначения: $Y_1 = X_1$, $Y_2 = X_0X_1$, $Y_3 = X_0X_0X_1$, ... Тогда произведение $X_{\varepsilon_1} \ldots X_{\varepsilon_{p-1}}X_1$ можно представить в виде $Y_{k_1} \ldots Y_{k_t}$, где $k_1 \geqslant 1, \ldots, k_t \geqslant 1$. Например,

$$X_1X_0X_0X_1X_0X_1 = Y_1Y_3Y_2.$$

Полилогарифмическая функция определяется следующим образом:

$$\mathrm{Li}_{k_1 \ldots k_t}(z) = \Lambda_\varepsilon(z).$$

Для остальных ε (с $\varepsilon_p = 0$) функции Λ_ε представляются в виде конечных сумм

$$\sum \mathrm{Li}_{\underline{k}}(z) \ln^? z,$$

где $\underline{k} = (k_1, \ldots, k_t)$, а $\ln^? z$ обозначает некоторую степень логарифма.

Для функции $\mathrm{Li}_{\underline{k}}$ разложение в степенной ряд имеет следующий вид:

$$\mathrm{Li}_{\underline{k}}(z) = \sum_{n_1 > \ldots > n_t} \frac{z^{n_1}}{n_1^{k_1} \ldots n_t^{k_t}}.$$

Радиус сходимости этого ряда равен 1.

Положим $\mathrm{Li}_{\underline{k}}(1) = \zeta(\underline{k})$. Например, $\zeta(3, 2) = \sum_{m>n} \frac{1}{m^3n^2}$. Так определенные числа $\zeta(\underline{k})$ будем называть *множественными значениями дзета-функции* (сокращенно MZV — multiple zeta values). Легко доказать, что рассматриваемый ряд сходится при $k_1 \geqslant 2$. Но, например, при $\underline{k} = (1, 1)$ получаем ряд $\sum_{m>n} \frac{1}{mn}$, который расходится. Действительно, сумма $\sum_{n=1}^{m-1} \frac{1}{n}$ приблизительно равна $\ln m$, а ряд $\sum \frac{\ln m}{m}$ расходится. Те же самые рассуждения показывают, что ряд $\sum_{m>n} \frac{1}{m^2n}$ сходится, поскольку ряд $\sum \frac{\ln m}{m^2}$ сходится.

Это означает следующее. Полилогарифмическая функция регулярна в нуле. Но при подходе к точке $z = 1$ для регулярности нужны какие-то дополнительные предположения. Например, при $k_1 \geqslant 2$ ряд абсолютно сходится на единичной окружности. Таким образом, в точке $z = 1$ функция имеет предел, но она не продолжается до голоморфной функции в окрестности этой точки.

Можно также рассматривать полилогарифмические функции от многих переменных:

$$\mathrm{Li}_{\underline{k}}(z_1, \ldots, z_s) = \sum_{n_1 > \ldots > n_t} \frac{z_1^{n_1} \ldots z_s^{n_s}}{n_1^{k_1} \ldots n_t^{k_t}}, \qquad s \leqslant t.$$

Множественные значения дзета-функции мы определили, положив $z = 1$. Гончаров и некоторые физики обнаружили, что очень интересные числа получаются также, если в качестве z взять корень из единицы, т. е. рассмотреть ряд $\sum\limits_{n=1}^{\infty} \frac{\alpha^n}{n^k}$, где $\alpha^p = 1$. Сумму такого ряда можно также записать в виде $\sum\limits_{j=0}^{p-1} \alpha^j \sum\limits_{n \equiv j \pmod p} \frac{1}{n^k}$.

4. Гипотезы о природе некоторых чисел

Мы будем рассматривать числа $\zeta(k_1, \ldots, k_t)$, где $k_1 \geqslant 2$, $t \geqslant 1$, $k_2 \geqslant 1, \ldots, k_t \geqslant 1$. Все эти числа вещественны. Среди них есть известные нам числа $\zeta(2) = \frac{\pi^2}{6}$ и $\zeta(4) = \zeta(2)^2 \times \frac{2}{5}$. Числа $\zeta(2k)$ нас интересовать не будут. А вот числа $\zeta(3)$, $\zeta(5)$, ... для нас весьма интересны.

Назовем число $p = p(\underline{k}) = k_1 + \ldots + k_t$ *весом* числа $\zeta(k_1, \ldots, k_t)$, а число t назовем *глубиной*. Обозначим через $\mathscr{Z}_p$ (в честь Zagier'a) множество всех линейных комбинаций с рациональными коэффициентами всех MZV веса p. Условно полагаем $\mathscr{Z}_0 = \mathbb{Q}$ и $\mathscr{Z}_1 = (0)$, поскольку в $\mathscr{Z}_1$ могло бы войти только число $\zeta(1) = \sum \frac{1}{n} = +\infty$. (Цагир предложил заменить $\zeta(1)$ суммой ряда $\sum \frac{1}{n}$ «по Эйлеру», т. е. постоянной Эйлера

$$C = \lim_{n \to \infty} \left(1 + \frac{1}{2} + \ldots + \frac{1}{n} - \ln n\right) = Res_{s=1} \zeta(s),$$

но не очевидно, что это правильное решение.) Мы уже выясняли, что $\mathscr{Z}_2 = \mathbb{Q}\pi^2$. Можно также доказать, что $\mathscr{Z}_3 = \mathbb{Q}\zeta(3)$, поскольку $\zeta(3) = \zeta(2, 1)$ (см. ниже).

О множествах $\mathscr{Z}_p$ есть несколько гипотез, которые не вполне независимы друг от друга.

1. Сумма множеств $\mathscr{Z}_0, \mathscr{Z}_1, \ldots, \mathscr{Z}_p$ прямая, т. е. если $z_0 + z_1 + \ldots + z_p = 0$, где $z_j \in \mathscr{Z}_j$ при $j = 0, 1, \ldots, p$, то $z_j = 0$ при всех j.

В связи с этой гипотезой проверено, что если $p(\underline{k}_j) = j$, $0 \leqslant j \leqslant 17$ и $|m_j| \leqslant 10^{10}$, $m_j \in \mathbb{Z}$, то $|\sum_j m_j \zeta(\underline{k}_j)| \geqslant 10^{-50}$.

2) Пусть $d_p = [\mathscr{Z}_p : \mathbb{Q}]$ (размерность $\mathscr{Z}_p$ над $\mathbb{Q}$). Тогда, как предположил Цагир,

$$\sum_{p=0}^{\infty} d_p t^p = \frac{1}{1 - t^2 - t^3}.$$

Более детальные гипотезы основаны на следующем наблюдении. Рассмотрим, например, произведение $\zeta(2)\,\zeta(3) = \sum \frac{1}{n^2} \sum \frac{1}{m^3}$. Здесь m и n пробегают часть целочисленной решетки, расположенную в положительном квадранте. Разобьем это множество на три подмножества: диагональ, наддиагональные элементы и поддиагональные элементы (рис. 2).

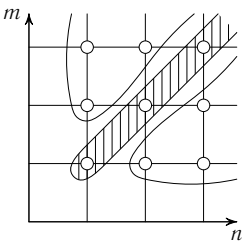


Рис. 2

Легко проверить, что суммирование по диагональным элементам дает $\zeta(5)$, суммирование по поддиагональным элементам дает $\zeta(2, 3)$, а суммирование по наддиагональным элементам дает $\zeta(3, 2)$. Поэтому

$$\zeta(2)\,\zeta(3) = \zeta(5) + \zeta(2, 3) + \zeta(3, 2).$$

Обобщая эти рассуждение, можно получить соотношение вида

$$\zeta(a_1, \ldots, a_s)\,\zeta(b_1, \ldots, b_t) = \sum \zeta(c_1, \ldots).$$

Здесь число $\zeta(a_1, \ldots, a_s)$ получается суммированием по наборам $n_1 > \ldots > n_s$, а число $\zeta(b_1, \ldots, b_t)$ получается суммированием по наборам $m_1 > \ldots > m_t$. Эти наборы нужно перетасовать, т. е. объединить их и после этого расположить числа в порядке убывания.

Это — первое множество мультипликативных соотношений. Есть еще одно множество мультипликативных соотношений. Оно получается следующим образом. Ранее было доказано, что

$$\zeta(3) = \frac{1}{2!} \int_0^{\infty} \frac{x^2}{e^x - 1}\, dx.$$

Сделав замену $x = \ln t$, получим

$$\zeta(3) = \int \frac{(\ln t)^2}{t - 1} \frac{dt}{t}.$$

После некоторых преобразований можно получить следующее выражение:

$$\zeta(3) = \iiint\limits_{1 > x_1 > x_2 > x_3 > 0} \frac{dx_1}{x_1} \frac{dx_2}{x_2} \frac{dx_3}{1 - x_3}.$$

Напомню, что в комбинаторных обозначениях $\zeta(3)$ соответствует произведению $Y_3 = X_0 X_0 X_1$; при этом в интеграле знаменатели трех дробей

имеют вид $x_1 - 0$, $x_2 - 0$ и $1 - x_3$. В общем случае формула выглядит следующим образом. Положим $\omega_0(x) = \dfrac{dx}{x}$ и $\omega_1(x) = \dfrac{dx}{1-x}$. Тогда

$$\zeta_{\varepsilon_1,\ldots,\varepsilon_p} = \int\limits_{1>x_1>\ldots>x_p>0} \ldots \int \omega_{\varepsilon_1}(x_1)\ldots\omega_{\varepsilon_p}(x_p).$$

Чтобы вычислить $\zeta(k_1, \ldots, k_t)$, нужно рассмотреть произведение $Y_{k_1}\ldots$ $\ldots Y_{k_t} = X_{\varepsilon_1}\ldots X_{\varepsilon_p}$ (здесь $p = k_1 + \ldots + k_t$). По определению $\zeta(k_1, \ldots, k_t) = \zeta_{\varepsilon_1,\ldots,\varepsilon_p}$.

Число $\zeta(2)$ аналогично выражается в виде двойного интеграла:

$$\zeta(2) = \iint\limits_{1>x_1>x_2>0} \frac{dx_1}{x_1}\,\frac{dx_2}{1-x_2}$$

Поэтому $\zeta(2)\,\zeta(3)$ представляется в виде пятикратного интеграла. Переменные нужно перетасовать (расположив их при этом по порядку). Те же самые рассуждения, что и раньше, дают некоторую сумму выражений. Но в этом случае есть некоторое упрощение: диагональ имеет нулевую меру, и поэтому ее можно не рассматривать.

Такие рассуждения дают второе множество мультипликативных соотношений.

Для $\zeta(2)\zeta(3)$ получаем два выражения. Одно из них имеет вид $\zeta(2)\zeta(3) = \zeta(5) + \zeta(2, 3) + \zeta(3, 2)$. Другое соотношение тоже можно записать явным образом, но мы не будем этого делать. В результате получаем одно линейное соотношение между MZV данного веса $p \geqslant 4$. Из интегрального представления получаем также равенство $\zeta(3) = \zeta(2, 1)$

Основная гипотеза такова: все независимые линейные соотношения между MZV данного веса получаются таким способом. Возможно, из этой гипотезы чисто алгебраически можно вывести те две гипотезы, которые были сформулированы ранее. Мои ученики работают сейчас над этой редукцией, но пока это не закончено. Это не легко.

Эта гипотеза весьма сильная. Из нее, в частности, следует, что числа $\zeta(3)$, $\zeta(5)$, $\ldots$ трансцендентны и алгебраически независимы над полем рациональных чисел. Например, сейчас не известно, является ли число $\zeta(3)$ трансцендентным и является ли число $\zeta(5)$ иррациональным.

Пьер Картье

Комбинаторика деревьев

Лекция 24 мая 1999 года

Я начну с определения чисел Каталана, которые часто встречаются в комбинаторике. Числа Каталана — это числа 1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, ... Эта последовательность чисел задается первым членом $c_1 = 1$ и рекуррентным соотношением $c_n = \sum_{p+q=n} c_p c_q = \sum_{p=1}^{n-1} c_p c_{n-p}$. (Здесь удобно положить $c_0 = 0$.)

Чтобы получить явную формулу для чисел Каталана, рассмотрим производящую функцию $c(t) = c_1 t + c_2 t^2 + \ldots$ Из рекуррентного соотношения следует, что $c(t) = t + c(t)^2$. Решая это квадратное уравнение и учитывая, что $c_0 = 0$, получаем $c(t) = \frac{1}{2}(1 - \sqrt{1-4t})$. Следовательно, $c_n = \frac{1}{n}\binom{2n-2}{n-1}$.

Числа Каталана имеют много различных комбинаторных интерпретаций. Для меня наиболее важной будет их интерпретация, предложенная Кэли (1860). А именно: c_n — это число различных триангуляций выпуклого правильного $(n+1)$-угольника. Например, у квадрата есть ровно две триангуляции (рис. 1); это соответствует тому, что $c_3 = 2$. У шестиугольника

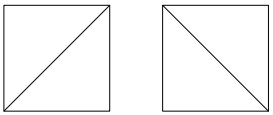


Рис. 1. Триангуляции квадрата

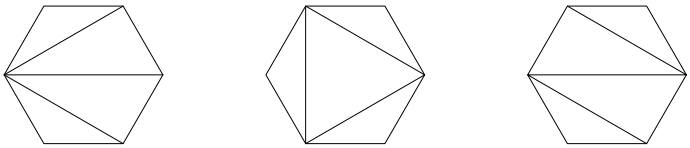


Рис. 2. Триангуляции шестиугольника

есть 3 разных типа триангуляций (рис. 2); под каждой из этих триангуляций написано количество различных триангуляций такого типа. Поэтому количество различных триангуляций шестиугольника равно $6 + 2 + 6 = 14 = c_5$.

Пусть c'_n — количество различных триангуляций $(n+1)$-угольника. Чтобы доказать равенство $c'_n = c_n$, достаточно проверить, что $c'_n = \sum_{p+q=n} c'_p c'_q$. (Очевидно, что $c'_1 = 1$.) Доказательство проведем на примере шестиугольника. Пометим его стороны числами 0, 1, 2, 3, 4, 5. Выделим треугольник триангуляции, содержащий сторону с пометкой 0 (рис. 3). На диагоналях, являющихся сторонами этого треугольника, тоже поставим пометки 0. В результате (в ситуации, изображенной на рис. 3) получим (3+1)-угольник и (2+1)-угольник, которые нужно триангулировать каким-то способом. В общем случае получаем $(p+1)$-угольник и $(q+1)$-угольник, где $p+q=n$.

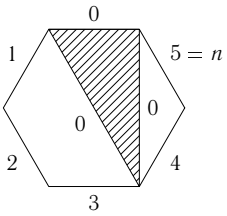


Рис. 3. Пометки сторон и диагоналей

Упражнение. Доказать, что для триангуляции выпуклого $(n+1)$-угольника требуется ровно $n-2$ диагонали и при этом получается ровно $n-1$ треугольник.

Другая интерпретация чисел Каталана связана с планарными бинарными корневыми деревьями с n листьями. Триангуляции $(n+1)$-угольника можно сопоставить связное дерево с $n+1$ свободными (т. е. принадлежащими только одному ребру) вершинами, как это показано на рис. 4. При этом одна свободная вершина, соответствующая стороне с пометкой 0, будет выделенной. Помеченную вершину дерева будем называть *корнем*, а остальные свободные вершины будем называть *листьями*.

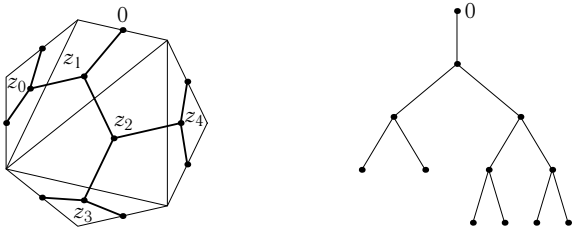


Рис. 4. Построение дерева по триангуляции

Объясним теперь, что такое дерево и что такое планарное бинарное дерево. Приведенный граф — это подмножество $E \subset \binom{V}{2}$, где $\binom{V}{2}$ — двухэлеметные подмножества в V. Элементы множества V — вершины графа, а элементы множества E — ребра графа. Приведенность графа означает, что у него нет петель (ребер с началом и концом в одной и той же вершине) и двойных ребер. Граф называют несвязным, если множество вершин V

можно разбить на два (непустых) непересекающихся множества V' и V'' так, что оба конца любого ребра лежат одновременно либо в V', либо в V''. Дерево — это связный граф, не содержащий циклов. (Цикл — это последовательность ребер, образующая многоугольник.) Несвязный граф, не содержащий циклов, называют лесом.

Планарное дерево — это дерево с выбранным вложением в плоскость. С комбинаторной точки зрения это эквивалентно тому, что для каждой вершины задан порядок обхода ребер, одним из концов которых является эта вершина.

Определим теперь, что такое *бинарное корневое дерево*. Прежде всего — это дерево, каждая вершина которого является концом либо одного ребра, либо трех ребер. (Концом одного ребра являются листья и корень.) Кроме того, ребра дерева должны быть ориентированы таким образом, чтобы из каждой внутренней вершины дерева выходило ровно два ребра и входило в нее ровно одно ребро.

Обозначим множество планарных бинарных корневых деревьев с n листьями через $\mathscr{T}_n$. Можно доказать, что планарные бинарные корневые деревья с n листьями находятся во взаимно однозначном соответствии с триангуляциями $(n+1)$-угольника; при этом треугольники триангуляции соответствуют внутренним вершинам дерева, а диагонали триангуляции соответствуют внутренним ребрам дерева. Таким образом, $c_n = |\mathscr{T}_n|$.

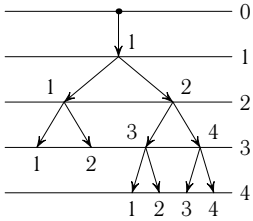


Рис. 5. Расстояние от корня

Для любых двух вершин дерева существует ровно один путь, соединяющий их. Поэтому для любой вершины можно определить расстояние от нее до корня. Расположим вершины, находящиеся на равном расстоянии от корня, на одном уровне (рис. 5). Пусть G_k — множество вершин уровня k. Сопоставив каждой вершине начало входящего в него ребра, получим последовательность отображений

$$G_0 \leftarrow G_1 \leftarrow G_2 \leftarrow G_3 \leftarrow G_4 \leftarrow \dots$$

Чтобы закодировать эту последовательность отображений, занумеруем вершины одного уровня слева направо (для планарного дерева этот порядок определен однозначно). Введем следующее обозначение: $[n] = \{1, 2, \dots, n\}$. Тогда $G_j \simeq [\gamma_j]$, где γ_j — количество элементов множества G_j. Последовательность отображений принимает вид

$$[\gamma_0] \leftarrow [\gamma_1] \leftarrow [\gamma_2] \leftarrow [\gamma_3] \leftarrow [\gamma_4] \leftarrow \dots$$

Эти отображения неубывающие: если числа p и q отображаются в p' и q', то из неравенства $p > q$ следует неравенство $p' \geqslant q'$.

1. Универсальная алгебра

Назовем произвольное отображение $\mu\colon X\times X\to X$ *магмой*; это название соответствует тому, что на бинарный закон композиции $(ab)=\mu(a,b)$ не накладываются никакие условия (магма не обладает никакой структурой).

Планарное бинарное корневое дерево с n листьями позволяет расставить скобки в последовательности из n элементов, т. е. оно позволяет задать порядок, в котором нужно выполнять бинарную операцию μ на множестве из n элементов. Например, дереву, изображенному на рис. 6, соответствует следующая расстановка скобок: $(((ab)c)(d(ef)))$. Эта расстановка скобок (т. е. последовательность выполнения операций) получается следующим образом. Возьмем пару свободных ребер с общей вершиной, например пару ребер с концами a и b. Отрежем эту пару ребер и сопоставим образовавшейся при этом новой свободной вершине элемент $\mu(a,b)$. В результате получим дерево с меньшим числом листьев. Для некоторой пары его свободных ребер с общей вершиной повторим ту же самую операцию и т. д.

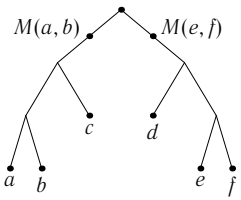


Рис. 6. Расстановка скобок с помощью дерева

Таким образом, каждое дерево $T\in\mathscr{T}_n$ задает отображение $\mu_T\colon X^n\to X$. Это отображение можно записать следующим образом. Пусть $l(T)$ — число листов корневого дерева T. Введем обозначение: $\bigsqcup\limits_T\{T\}\times X^{l(T)}=M(X)$. Тогда магма μ индуцирует отображение $\tilde{\mu}\colon M(X)\to X$, которое переводит $(T, x_1, \ldots, x_{l(t)})$ в $\mu_T(x_1, \ldots, x_{l(t)})$.

Универсальное свойство. 1) Для любого множества X с заданным умножением на множестве $M(X)$ есть умножение, индуцированное произведением деревьев $(T', T'')\mapsto T'*T''$ (рис. 7).

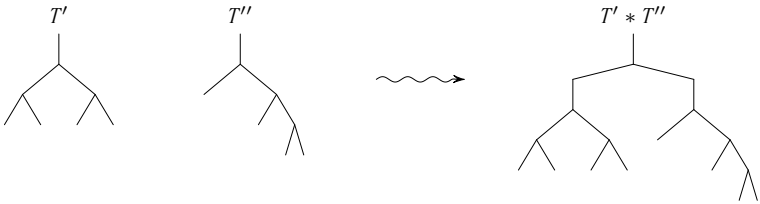
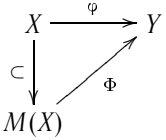


Рис. 7. Произведение деревьев

2) Если есть какое-то другое множество Y с умножением $\nu\colon Y\times Y\to Y$, то любое отображение $\varphi\colon X\to Y$ единственным образом можно продолжить

до отображения $\Phi\colon M(X) \to Y$, согласованного с умножением:



Вложение $X \subset M(X)$ устроено следующим образом. Множество $\mathscr{T}_1$ состоит из единственного дерева τ; элементу $x \in X$ мы сопоставляем элемент $(\tau, x_{l(\tau)})$.

Универсальное свойство означает, что $M(X)$ — свободная магма над X. Таким образом, мы получаем явную конструкцию свободной магмы.

Дадим теперь комбинаторную интерпретацию тождества $c(t) = t + c(t)^2$. Пусть $|X| = t$. Обозначим через $M(X)_p$ множество тех деревьев, для которых $l(T) = p$. Тогда $M(X)_p = \mathscr{T}_p \times X^p$ и множество $M(X)$ разлагается следующим образом: $M(X) = \bigsqcup\limits_{p=1}^{\infty} M(X)_p$. Рассмотрим для $M(X)$ ряд Пуанкаре $\sum c_p t^p$. Тождество $c(t) = t + c(t)^2$ вытекает из того, что $M(X) = X \sqcup M(X)^2$. Действительно, любой элемент множества $M(X)$ либо лежит в $M(X)_1$, либо единственным образом представляется в виде произведения двух элементов $M(X)$ (рис. 8).

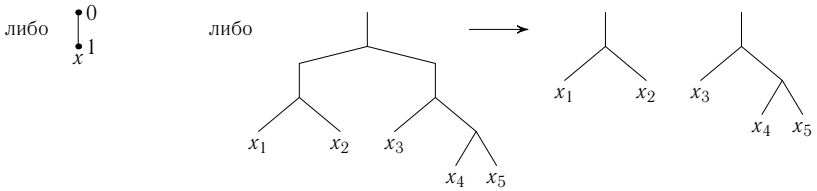


Рис. 8. Разложение дерева

2. Многогранники Сташефа

Многогранник $\mathscr{P}_2$ — это одна точка; $\dim \mathscr{P}_2 = 0$. Число вершин этого многогранника равно $1 = c_2$.

Многогранник $\mathscr{P}_3$ — это отрезок; $\dim \mathscr{P}_3 = 1$. Число вершин этого многогранника равно $2 = c_3$.

Многогранник $\mathscr{P}_4$ — это пятиугольник; $\dim \mathscr{P}_4 = 2$. Число вершин этого многогранника равно $5 = c_4$.

Что такое многогранник $\mathscr{P}_5$, объяснить уже сложнее. Число его вершин должно быть равно $c_5 = 14$. Его размерность должна быть равна 3. Кроме того, все многогранники Сташефа простые (в размерности 3 это означает,

что сечение вблизи каждой вершины является симплексом). Многогранник Сташефа $\mathscr{P}_5$ получается следующим образом. Склеим два тетраэдра. У полученного многогранника есть 3 вершины, вблизи которых сечения — четырехугольники. Но если мы отсечем каждую из этих трех вершин плоскостью, то в результате получим простой многогранник (рис. 9). Это и есть многогранник $\mathscr{P}_5$.

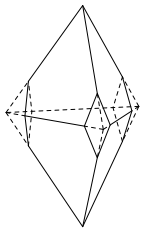


Рис. 9. Многогранник Сташефа $\mathscr{P}_5$

Многогранники Сташефа связаны с триангуляциями многоугольников, а именно, вершины многогранника $\mathscr{P}_n$ находятся во взаимно однозначном соответствии с триангуляциями $(n+1)$-угольника. Наиболее простая ситуация для $\mathscr{P}_3$ (рис. 10); вершины отрезка $\mathscr{P}_3$ соответствуют двум триангуляциям квадрата, а внутренность отрезка соответствует самому квадрату.

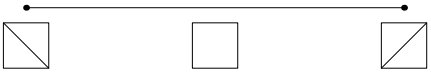


Рис. 10. Многогранник Сташефа $\mathscr{P}_3$

Следующий по сложности случай — многогранник $\mathscr{P}_4$ (5-угольник); его вершины находятся во взаимно однозначном соответствии с триангуляциями 5-угольника. (рис. 11). Внутренности $\mathscr{P}_4$ соответствует 5-угольник без диагоналей; каждой стороне $\mathscr{P}_4$ соответствует 5-угольник с одной диагональю; каждой вершине $\mathscr{P}_4$ соответствует 5-угольник с двумя диагоналями. При этом две диагонали, относящиеся к вершине $\mathscr{P}_4$, — это в точности те две диагонали, которые относятся к тем двум сторонам $\mathscr{P}_4$, которым принадлежит эта вершина.

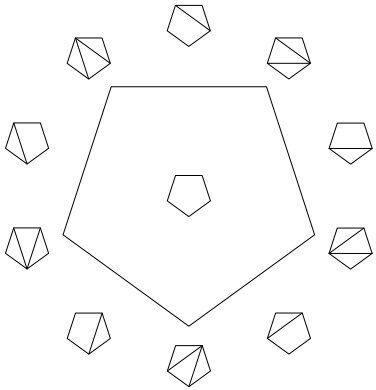


Рис. 11. Многогранник Сташефа $\mathscr{P}_4$

Аналогично для клеток $\mathscr{P}_5$ можно установить следующее соответствие: вершинам (0-мерным клеткам) $\mathscr{P}_5$ соответствуют 6-угольники с 3 непересекающимися диагоналями; 1-мерным клеткам $\mathscr{P}_5$ соответствуют 6-угольники с 2 непересекающимися диагоналями; 2-мерным клеткам $\mathscr{P}_5$ соответствуют 6-угольники с 1 диагональю; 3-мерной клетке $\mathscr{P}_5$ соответствует сам 6-угольник.

Граница $\partial\mathscr{P}_5$ состоит из шести 5-угольников (т. е. многогранников $\mathscr{P}_4$) и трех 4-угольников (т. е. многогранников $\mathscr{P}_3 \times \mathscr{P}_3$). Многогранник $\mathscr{P}_2$ —

это точка, поэтому $\mathscr{P}_4 = \mathscr{P}_2 \times \mathscr{P}_4 = \mathscr{P}_4 \times \mathscr{P}_2$. Таким образом, $\partial\mathscr{P}_5$ состоит из $2\mathscr{P}_2 \times \mathscr{P}_4$, $4\mathscr{P}_4 \times \mathscr{P}_2$ и $3\mathscr{P}_3 \times \mathscr{P}_3$. В общем случае естественно ожидать, что $\partial\mathscr{P}_n = \bigcup\limits_{p+q=n+1} p\mathscr{P}_p \times \mathscr{P}_q$. В таком случае можно определить отображения $\partial_k\colon \mathscr{P}_p \times \times \mathscr{P}_q \to \partial\mathscr{P}_n$, $1 \leqslant k \leqslant p$.

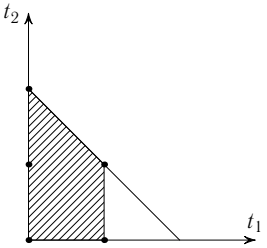


Рис. 12. Модель многогранника $\mathscr{P}_4$

Многогранник $\mathscr{P}_n$ представляется в пространстве $\mathbb{R}^{n-2}$ с координатами $t_1, \ldots, t_{n-2}$ следующей моделью. Рассмотрим множество, заданное системой неравенств $t_1 \geqslant 0$, $t_2 \geqslant 0$, $\ldots$ $\ldots$, $t_{n-2} \geqslant 0$ и $t_1 \leqslant 1$, $t_1 + t_2 \leqslant 2$, $t_1 + t_2 + t_3 \leqslant 3$, $\ldots$ $\ldots$, $t_1 + \ldots + t_{n-2} \leqslant n-2$ (на рис. 12 изображено такое множество для $n = 2$). Многогранник $\mathscr{P}_n$ можно получить, сделав дополнительное разбиение некоторых граней этого многогранника. Я оставляю это в качестве упражнения.

Приведем теперь комбинаторную конструкцию $\mathscr{P}_n$ как абстрактного клеточного комплекса. Пусть $n \geqslant 2$. Фиксируем выпуклый $(n+1)$-угольник и фиксируем число $0 \leqslant k \leqslant n-2$. Определим Γ_k (множество k-мерных клеток) следующим образом: элемент $\gamma \in \Gamma_k$ соответствует набору из $n-2-k$ непересекающихся диагоналей в $(n+1)$-угольнике. Например, для $k = n-2$ имеется ровно одна клетка Γ_{n-2}, а 0-мерные клетки Γ_0 соответствуют триангуляциям $(n+1)$-угольника.

Осталось определить отношение инцидентности. Пусть $\gamma \in \Gamma_k$ и $\delta \in \Gamma_l$, причем $k > l$. Тогда клетка δ смежна с клеткой γ (т. е. $\delta \subset \partial\gamma$ геометрически), если $\gamma \subset \delta$ (как множества диагоналей).

Многогранник $\mathscr{P}_n$ определяется как геометрическая реализация этого комплекса. При таком подходе нетривиально доказательство того, что геометрическая реализация этого комплекса гомеоморфна $(n-2)$-мерной клетке.

3. Пространство (вещественных) конфигураций

Рассмотрим вещественную проективную прямую. Топологи обозначают ее $\mathbb{RP}^1$, а алгебраические геометры используют обозначение $\mathbb{P}^1(\mathbb{R})$ или даже просто $\mathbb{P}^1$. Стандартное определение таково: точка $\mathbb{P}^1$ — это прямая в пространстве V, где $\dim V = 2$. Пусть $u_0, \ldots, u_n$ — различные точки $\mathbb{P}^1$, рассматриваемые с точностью до проективного преобразования. По-другому это можно сказать так: наборы $u_0, \ldots, u_n$ и $v_0, \ldots, v_n$ эквивалентны, если существует такой элемент $g \in G = \mathrm{GL}_2(\mathbb{R})$, что $gu_i = v_i$.

Помимо группы G есть другая важная группа $G_+ \subset G$, состоящая из матриц с положительным определителем. Индекс $[G : G_+]$ равен 2. Преобразования из группы G_+ сохраняют ориентацию $\mathbb{P}^1$.

Нас будут интересовать наборы точек $u_0, \ldots, u_n$, рассматриваемые с точностью до преобразования из группы G_+. Можно ввести следующую нормализацию: $u_0 = \infty$. Тогда $u_1, \ldots, u_n \in \mathbb{R}$ — различные точки, рассматриваемые с точностью до преобразования $u_i \mapsto au_i + b$, где $a > 0$ и $b \in \mathbb{R}$.

На этом множестве действует симметрическая группа S_n: элемент $\sigma \in S_n$ переводит набор $u_1, \ldots, u_n$ в набор $u_{\sigma(1)}, \ldots, u_{\sigma(n)}$.

Пространство наборов точек состоит из $n!$ связных компонент. Каждый элемент группы S_n переставляет эти связные компоненты. Выберем ту из компонент, для которой $u_1 < \ldots < u_n$. Такие наборы точек называют *вещественными конфигурациями*. Пространство вещественных конфигураций обозначают conf^+_{n+1}.

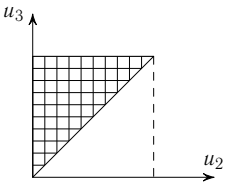


Рис. 13. Пространство конфигураций

Рассматриваемая группа преобразований позволяет нормализовать точки так, что $u_1 = 0$ и $u_n = 1$. Например, для $n = 4$ пространство конфигураций изображено на рис. 13. Действительно, в этом случае $0 = u_1 < u_2 < u_3 < u_4 = 1$.

Из пространства конфигураций conf^+_5 легко получить пятиугольник $\mathscr{P}_5$ (рис. 14). Точки пересечения трех прямых соответствуют нестабильным конфигурациям, в которых сталкиваются 3 точки. В этих точках нужно сделать раздутие.

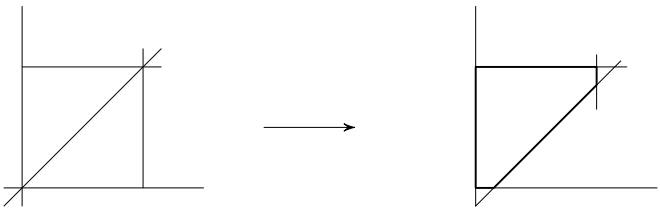


Рис. 14. Компактификация пространства конфигураций

В общем случае Капранов доказал, что многогранник $\mathscr{P}_n$ является естественной компактификацией пространства conf^+_n. При этом внутренность многогранника $\mathscr{P}_n$ отождествляется с пространством conf^+_n; размерности этих пространств равны $n - 2$.

Внутренность многогранника $\mathscr{P}_n$ параметризует конфигурации. Рассмотрим теперь внутренние точки граней многогранника $\mathscr{P}_n$ размерности $n - 3$. Такие точки принадлежат одному из множеств $\mathrm{int}\,\mathscr{P}_p \times \mathrm{int}\,\mathscr{P}_q$, где

$p+q=n+1$. Сопоставим точке такого множества композицию двух конфигураций (рис. 15). Для этого склеим две проективные прямые в одной точке. Конфигурации на одной проективной прямой параметризуются $\operatorname{int}\mathscr{P}_p$, а конфигурации на другой параметризуются $\operatorname{int}\mathscr{P}_q$. Для первой прямой точку касания пометим числом k, а для второй прямой эту точку пометим числом 0. При этом $1 \leqslant k \leqslant p$, как и должно быть (с. 75).

В общем случае, склеивая проективные прямые, получаем бинарное дерево (рис. 16).

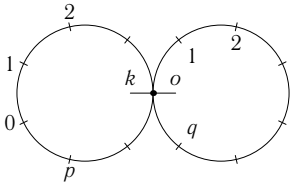


Рис. 15.
Композиция конфигураций

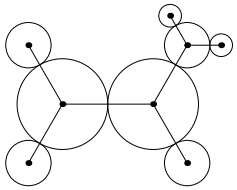


Рис. 16.
Дерево конфигураций

Пьер Картье

Что такое операда?

Лекция 27 мая 1999 года

Лекция будет состоять из следующих частей:

1. Что такое операда?
2. (Ко)гомологические операции.
3. Возвращение к конфигурационному пространству.
4. Заключение: физика.

Теория операд — это теория комбинирования операций. Предположим, что задана магма $X \times X \to X$, $a, b \mapsto (ab)$. Из этой бинарной операции можно строить более сложные операции.

Рассмотрим функцию $f(x_1, \ldots, x_n) \in X$, $x_i \in X$. На языке теории вычислений можно считать, что на вход подаются $x_1, \ldots, x_n$, а на выходе получается $f(x_1, \ldots, x_n)$ (рис. 1).

Операции можно сопоставить $(n+1)$-угольник; сторона с номером 0 соответствует выходу, а стороны с номерами $1, \ldots, n$ соответствуют входу (стороны нумеруются по часовой стрелке: рис. 2).

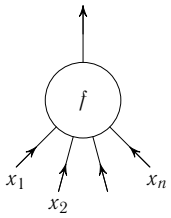


Рис. 1. Операция

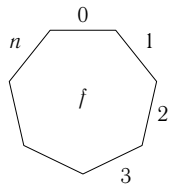


Рис. 2. Многоугольник, соответствующий операции

Определим теперь операцию $f \circ_i g$. Пусть заданы функция $f(x_1, \ldots, x_n)$, число $1 \leqslant i \leqslant n$ и функция $g(y_1, \ldots, y_p)$. Тогда можно сделать подстановку $x_i = g(y_1, \ldots, y_p)$. Функцию

$$f(x_1, \ldots, x_{i-1}, g(y_1, \ldots, y_p), x_{i+1}, \ldots, x_n)$$

мы будем рассматривать как функцию от $n+p-1$ переменных $x_1, \ldots$ $\ldots, x_{n+p-1}$; для этого ее нужно записать в виде

$$f(x_1, \ldots, x_{i-1}, g(x_i, \ldots, x_{i+p-1}), x_{i+p}, \ldots, x_{n+p-1}) = h(x_1, \ldots, x_{n+p-1}).$$

Операцию $f \circ_i g$ можно проинтерпретировать как на языке графов (рис. 3), так и на языке многоугольников (рис. 4). Графы удобны тем, что их вершины не нужно нумеровать: достаточно задать ориентацию в горизонтальном направлении (например, слева направо, как на рис. 3). Для многогранников сторона многоугольника g с номером 0 приклеивается к стороне многоугольника f с номером i.

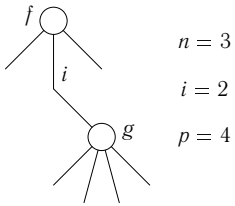


Рис. 3. Склейка графов

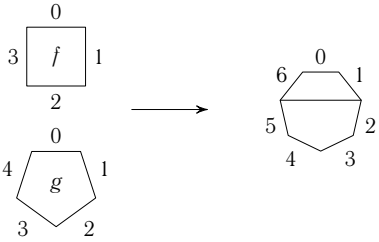


Рис. 4. Склейка многоугольников

Эти интерпретации показывают, что комбинирование операций связано как с деревьями, так и с многоугольниками, в которых выделены непересекающиеся диагонали.

Бинарные деревья связаны с бинарными операциями (рис. 5).

Пусть X — множество с бинарной операцией $\mu\colon X \times X \to X$. Положим $\mathscr{P}_n = Map(X^{\times n}, X)$. В частности, $\mathscr{P}_0 = X$ и $\mathscr{P}_1 = Map(X, X)$. Если $f \in \mathscr{P}_n$ и $g \in \mathscr{P}_p$, то $f \circ_i g \in \mathscr{P}_{n+p-1}$. Поэтому возникает отображение

$$\circ_i\colon \mathscr{P}_n \times \mathscr{P}_p \to \mathscr{P}_{n+p-1}.$$

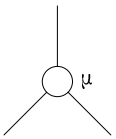


Рис. 5. Бинарная операция

Будем предполагать, что бинарная операция на X коммутативна: $\mu(a, b) = \mu(b, a)$.

На $\mathscr{P}_n$ действует симметрическая группа S_n посредством перестановок переменных:

$$\sigma f(x_1, \ldots, x_n) = f(x_{\sigma(1)}, \ldots, x_{\sigma(n)}).$$

Для $x = (x_1, \ldots, x_n)$ введем обозначение $x \cdot f = f(x_1, \ldots, x_n)$. Будем предполагать, что $x\sigma \cdot f = x \cdot \sigma f$.

Операда функций на X — это набор $\mathscr{P}_n$, $n = 0, 1, \ldots$, с заданным действием S_n на $\mathscr{P}_n$ и с заданными композициями $\circ_i$.

Линейная операда получается, если $\mathscr{P}_n$ — векторное пространство (над полем K), действие группы S_n линейное, а композиции билинейные.

Например, пусть

$$\mathscr{P}_n = \mathrm{Hom}_K(V^{\otimes n}, V), \quad \text{где} \quad V^{\otimes n} = \underbrace{V \otimes \ldots \otimes V}_{n}.$$

Здесь K — поле нулевой характеристики и V — векторное пространство размерности $d < +\infty$. На пространстве $V^{\otimes n}$ действует группа $G = \mathrm{GL}(V)$:

$$g(v_1 \otimes \ldots \otimes v_n) = gv_1 \otimes \ldots \otimes gv_n.$$

На пространстве $V^{\otimes n}$ действует также группа S_n:

$$\sigma(v_1 \otimes \ldots \otimes v_n) = v_{\sigma^{-1}(1)} \otimes \ldots \otimes v_{\sigma^{-1}(n)}.$$

Эти действия коммутируют. Кроме того, $V^{\otimes n}$ является полупростым S_n-модулем, т. е.

$$V^{\otimes n} = P_1 \oplus \ldots \oplus P_l,$$

где P_i — неприводимые (простые) S_n-модули. В этом разложении можно объединить одинаковые слагаемые, воспользовавшись тем, что $P \oplus P = P \otimes K^2$, $P \oplus P \oplus P = P \otimes K^3$, ... В результате получим разложение

$$V^{\otimes n} = \bigoplus_D P_D \otimes F_D.$$

При этом $\sigma(p_D \otimes f_D) = \sigma p_D \otimes f_D$ и $g(p_D \otimes f_D) = p_D \otimes gf_D$. В стабильной области, где $d \geqslant n$, представления группы S_n и представления группы $\mathrm{GL}_d(K)$ параметризуются одними и теми же диаграммами Юнга D. Это — двойственность Шура—Вейля.

Рассмотрим еще один пример (Макдональд и Милнор). Пусть Vect_K^f — категория конечномерных векторных пространств над полем K, Vect_K — категория векторных пространств (не обязательно конечномерных) над полем K. Пусть T — функтор из категории Vect_K^f в ту же самую категорию Vect_K^f. Функтор T сопоставляет векторному пространству V векторное пространство $T(V)$, а линейному отображению $\varphi\colon V \to W$ он сопоставляет линейное отображение $T(\varphi)\colon T(V) \to T(W)$. Линейные отображения φ и $T(\varphi)$ представляются матрицами (разных размеров). Потребуем, чтобы выполнялось следующее условие: элементы матрицы $T(\varphi)$ выражаются через элементы матрицы φ как однородные многочлены степени t. В таком случае будем говорить, что T — однородный функтор степени t. Пусть J_t — векторное пространство над K с заданным действием группы S_t. Положим

$$T_t(V) = \left(\left(\underbrace{V \otimes \ldots \otimes V}_{t} \right) \otimes J_t \right)^{S_t}$$

и $T = \bigoplus_{t=0}^{\infty} T_t$. В результате получим функтор из Vect_K^f в Vect_K.

Теперь пример из универсальной алгебры. Пусть V — векторное пространство (над полем K) с базисом $e_1, \ldots, e_d$ и $\mathrm{Sym}(V) = K[e_1, \ldots, e_d]$. Тогда

$$\mathrm{Sym}(V) = \bigoplus_{t=0}^{\infty} \mathrm{Sym}^t(V),$$

где $\mathrm{Sym}^t(V) = \left(\underbrace{V \otimes \ldots \otimes V}_{t}\right)^{S_t}$ — симметрическая часть. Это соответствует предыдущему примеру для $J_t = K$ с тривиальным действием группы S_t.

Пусть Com_K — категория коммутативных ассоциативных алгебр с единицей над полем K. Учитывая, что $\mathrm{Sym}^1(V) = V$, получаем вложение $V \subset \mathrm{Sym}(V)$; при этом $\mathrm{Sym}(V)$ — объект категории Com_K. Пусть A — объект категории Com_K. Тогда любое линейное отображение $\lambda\colon V \to A$ однозначно продолжается до гомоморфизма $\Lambda\colon \mathrm{Sym}(V) \to A$. Будем считать, что $J_t = K$ с тривиальным действием группы S_t. Операцию умножения в A обозначим $\mu(a, b) = a \cdot b$. Пусть

$$\mathscr{P}(A)_n = \mathrm{Hom}_K(A^{\otimes n}, A).$$

Ясно, что $\mu \in \mathscr{P}(A)_2$. Рассмотрим наименьшую подопераду в $\mathscr{P}(A)$, содержащую μ. Из ассоциативности умножения следует, что все деревья дают одну и ту же операцию. Операцию степени n можно обозначить μ_n.

Если помимо ассоциативности есть коммутативность, то есть инвариантность относительно действия группы S_n. Если же есть только ассоциативность, то получаем универсальную алгебру

$$T(V) = \bigoplus_{t=0}^{\infty} \underbrace{V \otimes \ldots \otimes V}_{t}.$$

В случае, когда $J_t = KS_t$ — регулярное представление ($\dim J_t = t$), получаем операду Ass_K.

Пусть V — векторное пространство. Свободная алгебра Ли $\mathrm{Lie}(V)$ над V определяется следующим образом. Рассмотрим линейное отображение $\Delta\colon T(V) \to T(V) \otimes T(V)$, которое для $v \in V$ имеет вид $\Delta v = v \otimes 1 + 1 \otimes v$. Тогда

$$\mathrm{Lie}(V) = \{u \in T(V)\colon \Delta u = u \otimes 1 + 1 \otimes u\}.$$

В частности, $V \subset \mathrm{Lie}(V)$. Определим в $\mathrm{Lie}(V)$ коммутатор $[u, u'] = uu' - u'u$.

Положим

$$\mathrm{Lie}_t(V) = \left(\underbrace{V \otimes \ldots \otimes V}_{t} \otimes \mathscr{L}_t\right)^{S_t},$$

где $\mathscr{L}_0 = 0$, $\mathscr{L}_1 = 0$; пространство $\mathscr{L}_2$ порождено $[x, y] = -[y, x]$; пространство $\mathscr{L}_3$ порождено элементами $[x, [y, z]]$, $[y, [z, x]]$ и $[z, [x, y]]$, связанными соотношением $[x, [y, z]] + [y, [z, x]] + [z, [x, y]] = 0$ (тождество Якоби).

Определим теперь когомологии Хохшильда. Пусть A — ассоциативная алгебра с единицей над полем K. Положим

$$C^p = \mathrm{Hom}_K(A^{\otimes p}, A).$$

Элемент $c(a_1, \ldots, a_p) \in C^p$ называют коцепью степени p. Ясно, что $\mu \in C^2$, где $\mu(a_1, a_2) = a_1 a_2$. Определим произведение коцепей следующим образом:

$$c \circ c' = c \circ_1 c' - c \circ_2 c' + c \circ_3 c' - \ldots$$

Положим

$$[c, c'] = c \circ c' + (-1)^{|c|+|c'|+|c||c'|} c' \circ c.$$

Тогда, как заметил Герстенхабер (60-е годы), с точностью до знака выполняется тождество Якоби:

$$[[c, c'], c''] = [c, [c', c'']] + (-1)^{|c|+|c'|+|c||c'|}[c', [c, c'']].$$

Кроме того,

$$[c', c] = (-1)^{|c|+|c'|+|c||c'|}[c, c'].$$

При этом $\deg[c, c'] = \deg c + \deg c' - 1$. В частности, $\mu_2 = [\mu, \mu]$ имеет степень 2. Из-за ассоциативности $\mu_2(a, b, c) = (ab)c - a(bc) = 0$.

Положим $\mathfrak{b}c = [\mu, c]$. Тогда $\mathfrak{b}\colon C^p \to C^{p+1}$ и из тождества Якоби следует, что $\mathfrak{b}\mathfrak{b} = 0$. Гомологии комплекса C с дифференциалом $\mathfrak{b}$ называют когомологиями Хохшильда и обозначают $HH^*(A, A)$.

В когомологиях Хохшильда есть две операции: коммутатор Герстенхабера $[\ ,\ ]_G$ (его степень равна -1) и $\cup$-произведение, которое определяется следующим образом:

$$c \cup c'(a_1, \ldots, a_p, a_{p+1}, \ldots, a_{p+q}) = c(a_1, \ldots, a_p)c'(a_{p+1}, \ldots, a_{p+q});$$

степень $\cup$-произведения равна 0.

Свойства этих операций таковы: $\cup$-произведение ассоциативно, а на уровне когомологий оно даже коммутативно с точностью до знака; операция $[\ ,\ ]_G$ коммутативна с точностью до знака и для нее с точностью до знака выполняется тождество Якоби. На уровне когомологий выполняется правило Лейбница:

$$[c, c' \cup c'']_G = [c, c']_G \cup c'' \pm c' \cup [c, c'']_G.$$

1. (Ко)гомологические операции

Дадим теперь описание конструкции топологической операды, которую предложил Сташеф в 60-е годы. Начнем с определения умножения Понтрягина в гомологиях. Пусть X — топологическое пространство и $\mu\colon X \times X \to X$ — непрерывное отображение. Рассмотрим гомологии с рациональными коэффициентами $\mathbb{Q}$. По формуле Кюннета $H_*(X \times X) = H_*(X) \otimes H_*(X)$,

поэтому отображение μ индуцирует отображение $H_*(\mu)\colon H_*(X)\otimes H_*(X) = H_*(X\times X)\to H_*(X)$ (умножение Понтрягина). Таким образом произведение в пространстве дает произведение в гомологиях. Если произведение μ ассоциативно, то произведение в гомологиях тоже ассоциативно; то же самое верно для коммутативности.

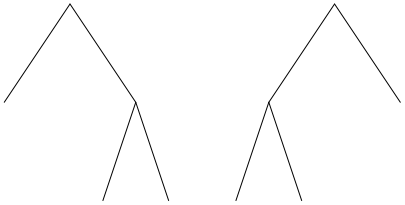


Рис. 6. Два дерева

Два дерева, изображенные на рис. 6, соответствуют двум разным расстановкам скобок в произведении трех элементов. Для топологического пространства с произведением $\mu\colon X\times X\to X$ эти два способа расстановки скобок определяют два отображения $X\times X\times X\to X$. Если эти отображения гомотопны, то умножение в H_* ассоциативно.

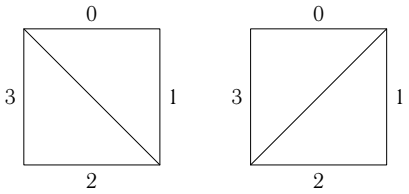


Рис. 7. Две триангуляции

Двум деревьям, рассмотренным выше, соответствуют две триангуляции квадрата (рис. 7), а этим двум триангуляциям соответствует многогранник Сташефа $\mathscr{P}_3$ (рис. 8).

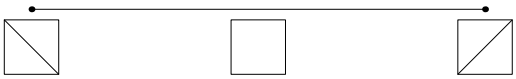


Рис. 8. Многогранник $\mathscr{P}_3$

Аналогично по отображению $\mathscr{P}_3\times X^3\to X$ можно определить отображение $\mathscr{P}_n\times X^n\to X$. Напомню, что $\partial\mathscr{P}_n = \bigcup\limits_{p+q=n+1} p\mathscr{P}_p\times\mathscr{P}_q$ и имеются отображения $\mathscr{P}_p\times\mathscr{P}_q\to\mathscr{P}_{p+q-1}$. Поэтому получаем отображение

$$H_*(\mathscr{P}_n)\otimes H_*(X)\otimes\ldots\otimes H_*(X)\to H_*(X).$$

Здесь $\mathscr{P}_n$ — топологическая клетка размерности $n-2$, поэтому $H_0(\mathscr{P}_n)=\mathbb{Q}$ и $H_i(\mathscr{P}_n)=0$ при $i\geqslant 1$. В результате для каждого $n=2,3,\ldots$ получаем операцию порядка n в $H_*(X)$. Например, при $n=2$ получаем умножение Понтрягина, при $n=3$ получаем препятствие к ассоциативности умножения Понтрягина.

2. Возвращение к конфигурационному пространству

Ввиду того, что пространство $\mathscr{M}_{0,n+1}(\mathbb{R})$ гомеоморфно $S_n \times \operatorname{int} \mathscr{P}_n$, $\overline{\mathscr{M}_{0,n+1}(\mathbb{R})} \approx S_n \times \mathscr{P}_n$. В положительных размерностях гомологии $H_*(\overline{\mathscr{M}_{0,n+1}(\mathbb{R})})$ тривиальны.

Положим $\Pi_n = H_0(\overline{\mathscr{M}_{0,n+1}(\mathbb{R})}, \mathbb{Q})$. Тогда $\Pi_n = \mathbb{Q}S_n$ — регулярное представление группы S_n.

Рассмотрим набор множеств $\mathscr{X}_n = \overline{\mathscr{M}_{0,n+1}(\mathbb{R})}$, $n = 0, 1, \ldots$ Группа S_n действует на $\mathscr{X}_n$. Для $1 \leqslant i \leqslant p$ можно образовать композицию $\circ_i$, как показано на рис. 9. В результате получаем вырожденную конфигурацию; все вырожденные конфигурации такого вида образуют пространство, гомеоморфное $\operatorname{int}(\mathscr{P}_p \times \mathscr{P}_q)$. При компактификации получаем точку пространства $\mathscr{P}_{p+q-1}$. Таким образом, многогранники Сташефа (или конфигурационные пространства) могут быть организованы в топологическую операду $\mathscr{X}$.

Если есть топологическая операда, то из нее с помощью гомологий можно получить алгебраическую операду. Из топологической операды Сташефа $\mathscr{X}$ получается тривиальная алгебраическая операда $Ass_{\mathbb{Q}}$. Но в комплексном случае теория нетривиальна.

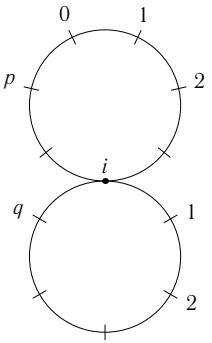


Рис. 9. Композиция

Комплексная теория получается следующим образом. Рассмотрим пространство $\mathscr{M}_{0,n+1}(\mathbb{C})$, которое получается из

$$\mathbb{C}^n_* = \{z = (z_0, \ldots, z_n) \in \mathbb{C}^n : z_j \neq z_k \text{ при } j \neq k\}$$

факторизацией по действию аффинной группы $z \mapsto az + b$. Компактификация $\overline{\mathscr{M}_{0,n+1}(\mathbb{C})}$ определяется довольно сложно; это было сделано Делинем и Мамфордом. Из пространств $\overline{\mathscr{M}_{0,n+1}(\mathbb{C})}$ можно организовать топологическую операду. В вещественном случае было объяснено, как это делается. В вещественном случае композиция представляется вещественными многочленами с рациональными коэффициентами. Их можно комплексифицировать, заменив вещественные переменные комплексными.

В вещественном случае гомологии получаются тривиальными, а в комплексном случае гомологии $H_*(\overline{\mathscr{M}_{0,n+1}(\mathbb{C})}, \mathbb{Q})$ очень интересны. Они образуют алгебраическую операду $\Pi_{\mathbb{C}}$. В этой алгебраической операде есть два специальных элемента, которые в некотором смысле ее порождают. А именно, cup-произведение $\cup$ и произведение Герстенхабера $[\ ,\ ]_G$. Это означает, что операда $\Pi_{\mathbb{C}}$ действует на $HH^*(A)$. Это получается вычисле-

ниями. Вопрос заключается в том, чтобы объяснить, почему это происходит.

В 1970 г. Арнольд вычислил гомологии $H_*(\mathbb{C}^n_*)$. Это позволяет вычислить гомологии некомпактифицированного пространства. Вопрос заключается в том, чтобы понять, почему гомологии $H_*(\overline{\mathscr{M}_{0,n+1}(\mathbb{C})}, \mathbb{Q})$ — это как раз то, что управляет операциями в когомологиях Хохшильда. Желательно построить действие на уровне коцепей.

Теперь коротко о связи с физикой. Одна из важнейших задач сейчас — точное решение уравнений Янга—Бакстера. Их решения можно получить как решения некоторых дифференциальных уравнений — так называемых уравнений Книжника—Замолодчикова, которые строятся на $\mathbb{C}^n_*$. При построении решений этих уравнений (Дринфельд) возникают степенные ряды, связанные с MZV-числами, о которых шла речь на первой лекции.

Есть также связь с диаграммами Фейманна и с матрицей рассеяния.

А. А. Кириллов

Метод орбит за пределами групп Ли. Бесконечномерные группы

Лекция 2 сентября 1999 года

Лекции, которые будут прочитаны сегодня и завтра, в каком-то смысле являются продолжением моих предыдущих лекций (см. «Студенческие чтения МК НМУ», выпуск 1). Прошлые две лекции назывались «Метод орбит и конечные группы», а эти две лекции будут называться «Метод орбит за пределами групп Ли». Я не буду сейчас останавливаться на том, что такое метод орбит. Скажу лишь, что он применяется к группам Ли. Так что основной рассматриваемый объект — это группы Ли. Но метод орбит можно применять и к другим группам, которые не являются группами Ли. Я приготовил три серии таких примеров:

1) бесконечномерные группы;
2) конечные группы;
3) квантовые группы.

Мои прошлогодние лекции были посвящены второй части — конечным группам. Поэтому об этом я говорить больше не буду, хотя и в этом направлении есть интересные продвижения. Сегодня я буду рассказывать про бесконечномерные группы, а завтра про квантовые группы. Квантовые группы — это одно из очень модных направлений современной математики. Их успех во многом обязан звучному названию. Тонкость заключается в том, что квантовые группы группами не являются — это объект другой природы. Но кое-что от групп в них все-таки есть, и метод орбит к ним можно попытаться применить. Об этих попытках я буду завтра рассказывать. А сегодня я буду говорить о бесконечномерных группах, которые тоже группами Ли не являются.

Обычные группы Ли — это (конечномерные) многообразия, которые одновременно снабжены структурой группы, причем структура многообразия и структура группы определенным образом согласованы. Бесконечномерные группы — это почти то же самое, но только теперь многообразие бесконечномерное, т. е. локальные системы координат бесконечномерные. Но если все конечномерные пространства одной и той же размерности над полем вещественных чисел изоморфны, то в бесконечномерном случае это

уже не так. Разных бесконечномерных пространств много, поэтому не всегда бывает ясно, какое именно из них нужно рассматривать. Типичный пример бесконечномерного пространства — это пространство функций. Например, можно рассмотреть функции на прямой. Но функции на прямой бывают разные. Во-первых, можно рассмотреть вообще все функции, но это бессмысленное понятие; обычно рассматривают функции, которые имеют какие-то свойства. Можно рассматривать непрерывные функции; это уже более осмысленное понятие. Еще лучше рассматривать гладкие функции, которые имеют одну, две, три... или бесконечное число производных. Можно рассматривать аналитические функции. Можно также вводить какие-то ограничения на бесконечности, например, рассматривать быстро убывающие функции или финитные функции, которые отличны от нуля только в конечной области. Как видите, бесконечномерных пространств очень много. Поэтому бесконечномерные группы в каком-то смысле, как и алгебраические многообразия, тоже не являются множествами. Обычно они состоят из функций и становятся множествами только после того, как мы уточним, какие именно функции рассматриваем. А до этого просто имеется некоторое правило, задающее групповой закон.

Задачи, которые здесь возникают при попытках применения метода орбит, очень интересны. Иногда они совпадают с уже известными классическими задачами, как решенными, так и не решенными. Иногда появляются новые задачи. Часть из этих новых задач я хочу сегодня обсудить.

Теперь я сделаю небольшое отступление. Люди часто интересуются, что происходит нового в математике, поэтому я обычно по мере сил стараюсь об этом рассказывать, чтобы удовлетворить любопытство аудитории. Сейчас математика помимо общих теорий, которые охватывают многочисленные частные примеры, проявляет особый интерес к исключениям. Любители исключений были всегда, но сейчас исключения играют все бóльшую и бóльшую роль. В каждой науке бывают исключительные объекты. Например, среди комплексных простых групп Ли есть 4 бесконечные серии и 5 особых примеров групп, которые ни в какие серии не укладываются. Подобные вещи есть и в других науках. Один такой пример я расскажу более подробно. Есть такой объект, связанный с группами Ли, с геометрией и со многими физическими приложениями, — это понятие *решетки*. Решетка — это дискретная подгруппа в n-мерном евклидовом пространстве, факторгруппа по которой компактна (иногда такие группы называют *кокомпактными*). Для любой решетки на прямой можно выбрать масштаб так, что решетка будет состоять из всех целых чисел. В этом случае задача классификации решеток тривиальна.

С точки зрения теории групп все решетки в n-мерном пространстве одинаковы: они изоморфны $\mathbb{Z}^n$. Но с геометрической точки зрения эти решетки могут быть разными. Например, на плоскости есть стандартная

целочисленная решетка. А еще можно построить решетку следующим образом. Рассмотрим на плоскости три координатные оси, образующие друг с другом угол 120° (рис. 1). Тогда каждой точке плоскости будут соответствовать не две координаты, а три: (x, y, z); здесь берутся проекции с учетом знака. Эти координаты связаны соотношением $x+y+z=0$. Если мы потребуем, чтобы все три координаты x, y, z были целыми, то получим другую решетку (рис. 2).

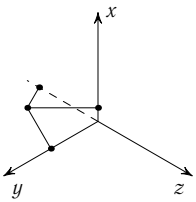


Рис. 1. Координаты на плоскости

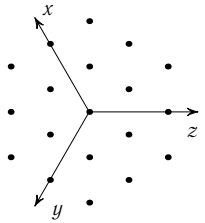


Рис. 2. Решетка на плоскости

Можно сравнить плотность этих двух решеток. Плотность решетки определяется следующим образом. Поместим в каждую точку решетки одинаковые шарики максимального размера так, чтобы эти шарики не пересекались. Отношение площади, покрытой кругами, ко всей площади — это и есть плотность. Видно, что вторая решетка плотнее прямоугольной решетки. В действительности, прямоугольная решетка — одна из наименее плотных, а вторая решетка — наиболее плотная.

Теорию решеток начали разрабатывать еще в XIX в. Сначала это было весьма популярное направление, потом оно было слегка подзабыто, а сейчас оно снова привлекло внимание математиков. В теории решеток помимо серий известны также особые, экзотические, решетки, которые не укладываются в серии. Одна из таких решеток связана с тем, что в 4-мерном пространстве есть гораздо больше правильных многогранников, чем в других пространствах. В 3-мерном пространстве есть 5 правильных выпуклых многогранников: правильный тетраэдр, куб, октаэдр, икосаэдр и додекаэдр. В пространстве размерности 4 имеется 6 правильных многогранников (там добавляется многогранник, ограниченный 24 октаэдрами, который не имеет аналогов в пространствах других размерностей). Во всех остальных пространствах есть только 3 правильных многогранника (правильный симплекс, куб и многогранник, двойственный кубу). Таким образом, 4-мерное пространство является исключительным, и в нем есть исключительная решетка, связанная с правильным 24-гранником. Эта решетка состоит из точек с целыми координатами x_1, x_2, x_3, x_4, которые удовлетворяют следующему условию: они одновременно либо все четные, либо все нечетные. Можно проверить, что у каждой точки будет ровно 24 ближайших к ней

точки; расстояние между соседними точками равно 2. Например, у начала координат есть 8 соседних точек вида $(\pm 2, 0, 0, 0)$ и 16 соседних точек вида $(\pm 1, \pm 1, \pm 1, \pm 1)$. Все точки решетки равноправны: любую точку можно перевести в любую другую симметрией решетки.

Это все были достаточно простые примеры решеток. Более интересный пример — *решетка Лича* в пространстве $\mathbb{R}^{24}$. Эта решетка была открыта геометрами, а в последнее время она часто использовалась физиками в связи со струнными теориями. Я дам не то определение решетки Лича, которое было дано первоначально, а то, которое связано с физикой. Многие физики считают, что мы живем в 26-мерном пространстве Минковского $\mathbb{R}^{1,25}$, в котором есть одна координата временная и 25 координат пространственных: $(t, x_0, \ldots, x_{24})$. Непосредственно наши чувства говорят нам, что мы живем в 4-мерном мире, в котором есть одна временная координата и три пространственные. Но наши чувства грубые, они не могут отличить очень маленькое 10-мерное многообразие от точки. Поэтому размеры пространства в 22 направлениях могут быть очень маленькими, порядка 10^{-33} см; этого вполне достаточно для того, чтобы никакой прибор эти направления не обнаружил. Остальные 3 пространственных направления будут видны явно. Но это отступление. Вернемся к математике. В пространстве Минковского $\mathbb{R}^{1,25}$ есть квадратичная форма $t^2 - x_0^2 - \ldots - x_{24}^2$, и на световом конусе есть замечательный целочисленный вектор $(70, 0, 1, 2, \ldots, 24) = \rho$. Рассмотрим в $\mathbb{R}^{1,25}$ целочисленную решетку M. Выберем в $\mathbb{R}^{1,25}$ ортогональное дополнение $\rho^\perp$; при этом $\rho^\perp \supset \rho$. Рассмотрим факторпространство $\rho^\perp/\mathbb{R}\rho$, которое имеет размерность 24. Если взять целочисленные точки в $\rho^\perp$ и профакторизовать их по $\mathbb{R}\rho$, то получится целочисленная решетка в 24-мерном пространстве, которая замечательна во всех отношениях. Эта решетка не имеет аналогов в пространствах других размерностей.

Теперь я возвращаюсь к основной теме — метод орбит для бесконечномерных групп. Я хочу рассмотреть две бесконечномерные группы, первая из которых хорошо изучена, а вторая почти совсем не изучена. Обычно абстрактные группы возникают из групп преобразований. А группы преобразований — это обычно преобразования чего-нибудь, которые что-нибудь сохраняют, т. е. обычно берется множество, на котором есть дополнительная структура, и рассматриваются все взаимно однозначные преобразования, которые сохраняют эту структуру. Например, можно взять плоскость, а на ней дополнительную структуру — расстояние между точками. Тогда получим группу движений и отражений, т. е. группу изометрий. Можно взять более обширную группу преобразований, сохраняющих проективную структуру на плоскости или сохраняющих топологию на плоскости, т. е. непрерывных преобразований. Группа непрерывных преобразований плоскости бесконечномерная.

Одна из наиболее изученных бесконечномерных групп — группа взаимно однозначных гладких преобразований окружности; ее обозначают $\mathrm{Diff}(S^1)$. Окружность можно представлять как факторпространство: $S^1 = \mathbb{R}/\mathbb{Z}$, т. е. мы вводим координату $t \pmod 1$ (две точки прямой отождествляются, если их разность — целое число). Диффеоморфизм окружности задается функцией $s = f(t)$; здесь s — новая координата. При этом разность между $f(t+1)$ и $f(t)$ должна быть целым числом. Функция f может быть либо монотонно возрастающей, либо монотонно убывающей. Ясно, что это — две связные компоненты группы $\mathrm{Diff}(S^1)$. Мы будем рассматривать связную компоненту, содержащую единичный элемент группы; эту группу обозначают $\mathrm{Diff}_+(S^1)$. Для нее условие такое: $f(t+1) = f(t) + 1$. Функция f гладкая; для нее выполняется условие $f'(t) > 0$. Условие $f(t+1) = f(t) + 1$ в терминах производной запишется так:

$$\int_0^1 f'(t)\,dt = 1.$$

Кроме того, $f'(t+1) = f'(t)$.

Группа получает достаточно наглядное описание. Если в качестве параметра брать не f, а f', то группа состоит из положительных периодических функций (с периодом 1), для которых интеграл по периоду равен 1. Топологически это множество тривиально: оно стягиваемо, потому что представляет собой выпуклое множество в линейном пространстве. Оба условия (положительность и равенство интеграла 1) выпуклые. Но чтобы получить саму группу, надо от производной перейти к функции. Для этого достаточно задать начальное условие, которое принадлежит окружности. В результате получим, что интересующая нас группа $\mathrm{Diff}_+(S^1)$ гомотопически эквивалентна окружности. Иногда бывает удобно рассматривать односвязные группы. Чтобы получить односвязную группу, можно не факторизовать по функциям, принимающим целые значения, а рассматривать все монотонные функции. Полученная группа $\widetilde{\mathrm{Diff}}_+(S^1)$ будет универсальным накрытием группы $\mathrm{Diff}_+(S^1)$.

Обсудим теперь, как к группе $G = \widetilde{\mathrm{Diff}}_+(S^1)$ применить метод орбит. Для этого прежде всего нужно перейти от группы Ли к алгебре Ли $\mathfrak{g} = \mathrm{Lie}(G)$. Алгебра Ли является касательным пространством в единице, т. е. она состоит из инфинитезимальных преобразований окружности. Как геометрический объект, алгебра Ли состоит из векторных полей на окружности. Действительно, мы рассматриваем преобразования $t \mapsto t + \varepsilon v(t)$. В рассматриваемой группе групповой закон соответствует суперпозиции функций: $f_1 \circ f_2(t) = f_1(f_2(t))$. На уровне алгебры Ли групповой закон превращается в коммутатор $[v, w] = vw' - v'w$; здесь $v' = dv(t)/dt$. Известно, что каждая группа Ли действует на своей алгебре Ли линейными преобразованиями (присоединенное представление группы Ли). В нашем случае

группа состоит из замен переменных (систем координат) на окружности: $t = t(s)$. Для векторных полей тоже можно делать замены систем координат, а именно, векторное поле можно записать в одной системе координат, а можно и в другой. Оказывается, что присоединенное представление группы это и есть та же самая замена переменных, но для векторных полей.

Векторное поле можно задавать функцией $v(t)$. Но геометрический смысл этой функции состоит в том, что она задает векторное поле. Поэтому правильнее писать не $v(t)$, а $v(t)\dfrac{d}{dt}$. Если мы делаем замену переменных $t = t(s)$, то для векторных полей получаем

$$v(t)\frac{d}{dt} \mapsto v(t(s))\frac{d}{dt(s)} = v(t(s))\frac{d}{t'_s\,ds}.$$

Если мы следим только за коэффициентом, то, введя обозначение $t = f(s)$, получаем

$$v \mapsto \frac{v \circ t}{t'} = \frac{v \circ f}{f'}.$$

Для метода орбит нужна не сама алгебра Ли $\mathfrak{g}$, а двойственное пространство $\mathfrak{g}^*$, которое состоит из линейных функционалов на $\mathfrak{g}$. Какой геометрический смысл имеют объекты, двойственные векторным полям? Двойственный объект в паре с векторным полем позволяет получить число. Оказывается, что двойственный объект — это квадратичный дифференциал, т. е. выражение вида $p(t)(dt)^2$. При замене переменных $t = f(s)$ квадратичный дифференциал меняется так: $p(t)(dt)^2 \mapsto p \circ f\,(f')^2$. Можно проверить, что если есть векторное поле $v(t)\dfrac{d}{dt}$ и квадратичный дифференциал $p(t)(dt)^2$, то составленное из них выражение $p(t)v(t)\,dt$ является дифференциальной формой, т. е. при замене переменных оно умножается на первую степень производной. Как и всякую дифференциальную форму, ее можно интегрировать. Интеграл $\displaystyle\int_{S^1} p(t)v(t)\,dt$ не зависит от выбора параметра. Тем самым квадратичный дифференциал является линейным функционалом на пространстве векторных полей. Обозначения специально выбирались так, что естественное действие группы на пространстве $\mathfrak{g}^*$, дуальное действию на самом пространстве $\mathfrak{g}$, тоже будет естественным: оно будет обычной заменой переменных для квадратичных дифференциалов по формуле, написанной выше.

Функция p, вообще говоря, является обобщенной функцией. Но нам интересен тот случай, когда p — обычная гладкая функция, потому что наиболее интересные примеры коприсоединенных орбит связаны именно с такими функциями p.

Я должен сделать еще одно замечание. Метод орбит нужен для того, чтобы строить и изучать бесконечномерные унитарные представления.

Но в основных приложениях бесконечномерных унитарных представлений (квантовой теории поля и квантовой механике вообще) обычные представления не так нужны, как проективные представления. Множеством симметрий квантовой системы является не множество унитарных операторов, а соответствующее проективное пространство. Два пропорциональных унитарных оператора не различаются, потому что две волновые функции, отличающиеся множителем, по модулю равным 1, не различаются. Поэтому понятие обычного представления, как функции на группе, удовлетворяющей соотношению

$$\pi(g_1 g_2) = \pi(g_1)\,\pi(g_2),$$

заменяется на понятие проективного представления, как функции на группе, удовлетворяющей соотношению

$$\pi(g_1 g_2) = c(g_1, g_2)\,\pi(g_1)\,\pi(g_2),$$

где $c(g_1, g_2) \in \mathbb{C}$, $|c(g_1, g_2)| = 1$.

Проективные представления группы $\mathbb{G}$ сводятся к обычным представлениям чуть большей группы, а именно, центрального расширения группы G. Если у группы по каким-то причинам нет нетривиальных центральных расширений, то каждое проективное представление на самом деле является обычным представлением. Но у многих интересных групп есть нетривиальные центральные расширения, а значит, есть нетривиальные проективные представления, не сводящиеся к обычным представлениям. Например, группа диффеоморфизмов окружности $\mathrm{Vect}(S^1)$ имеет нетривиальное центральное расширение.

Метод орбит должен на это отреагировать. Мы заменяем группу на ее центральное расширение; алгебра Ли при этом тоже заменяется на центральное расширение и коприсоединенное действие тоже заменяется на чуть более сложное действие. Коприсоединенное действие центрального расширения отличается от исходного действия тем, что оно заменяет линейное действие на аффинное действие.

Группа G и ее центральное расширение $\widetilde{G}$ включаются в точную последовательность

$$0 \to \mathbb{R} \to \widetilde{G} \to G \to 1.$$

Я начал последовательность с нуля, потому что $\mathbb{R}$ — аддитивная группа, а заканчиваю последовательность единицей, потому что обычно группа G записывается мультипликативно; обе группы 0 и 1 состоят ровно из одного элемента.

Для алгебр Ли точная последовательность выглядит так:

$$0 \to \mathbb{R} \to \tilde{\mathfrak{g}} \to \mathfrak{g} \to 0.$$

Имеется также двойственная последовательность

$$0 \leftarrow \mathbb{R} \leftarrow \tilde{\mathfrak{g}}^* \leftarrow \mathfrak{g}^* \leftarrow 0.$$

На уровне линейных пространств эта точная последовательность расщепляется, поэтому можно считать, что $\tilde{\mathfrak{g}}^* = \mathfrak{g}^* \oplus \mathbb{R}$ как линейные пространства. Как мы знаем, пространство $\mathfrak{g}^*$ состоит из квадратичных дифференциалов; элементы этого пространства называют *моментами*. Расширенный момент состоит из квадратичного дифференциала и числа. Нужно задать действие группы на множестве пар, состоящих из квадратичного дифференциала и числа. Это действие должно выглядеть так:

$$K(f)(p, c) = (p \circ f\,(f')^2 + cS(f),\ c). \tag{1}$$

То, что c остается без изменений — это общий факт для всех центральных расширений; дополнительный параметр не изменяется при коприсоединенном действии. Из-за этого его иногда называют *зарядом*. Аффинное преобразование для p состоит из двух частей. Линейную часть $p \circ f\,(f')^2$ мы уже знаем. Добавочный член от p не зависит, а от c зависит линейно, поэтому он имеет вид $cS(f)$. Чтобы выражение (1) было действием, квадратичный дифференциал $S(f)$ должен обладать следующим свойством:

$$S(f_1 \circ f_2) = S(f_1) \circ f_2\,(f_2')^2 + S(f_2).$$

Любители когомологий могут сказать, что это выражение является уравнением коцикла.

Существует ровно одно S, обладающее требуемым свойством, а именно,

$$S(f) = \frac{f'''}{f'} - \frac{3}{2}\left(\frac{f''}{f'}\right)^2. \tag{2}$$

Теперь можно забыть обо всем ранее сказанном и относиться к этому лишь как к мотивировке. Математическая постановка задачи такова. Нужно получить классификацию коприсоединенных орбит, т. е. получить классификацию периодических функций от аргумента t относительно преобразований вида (1).

Выражение (2) было хорошо известно математикам еще в XIX веке. Его открыл немецкий математик Шварц; в связи с этим $S(f)$ называют *производной Шварца*. Вполне естествен вопрос, как устроены функции, для которых $S(f) = 0$? Функции, для которых производная Шварца равна нулю, обладают замечательным свойством: они образуют группу относительно композиции. Эти функции являются решениями дифференциального уравнения порядка 3, поэтому соответствующая группа имеет размерность 3. Трехмерных групп Ли не так уж много. На самом деле эта группа изоморфна группе PSL_2 и реализуется как группа дробно-линейных преобразований. Поэтому общее решение уравнения $S(f) = 0$ выглядит так: $f(x) = \dfrac{ax+b}{cx+d}$.

Еще более замечательное свойство производной Шварца заключается в том, что если вы вычислите производную Шварца от какой-нибудь

хорошей функции, то обычно в результате получается число простого вида. Например, если $f(x) = e^{\lambda x}$, то

$$S(f) = \frac{\lambda^3 e^{\lambda x}}{\lambda e^{\lambda x}} - \frac{3}{2}\left(\frac{\lambda^2 e^{\lambda x}}{\lambda e^{\lambda x}}\right)^2 = -\frac{\lambda^2}{2}.$$

Для функции $\operatorname{tg} x$ тоже получается число, но я не буду говорить, какое.

Нам предстоит решить следующую трудную задачу: привести к наиболее простому виду функцию, используя преобразования, действующие по описанному выше сложному правилу. Казалось бы, эта задача очень искусственная. Но я ее сейчас сформулирую по-другому, и тогда она будет выглядеть совершенно естественно. Оказывается, что эта задача равносильна вот какой геометрической задаче. Рассмотрим окружность S^1 и определим для нее понятие структуры проективного многообразия. Обычное гладкое многообразие получается, если мы покрываем топологическое пространство локальными системами координат и требуем, чтобы переход от одной системы координат к другой задавался гладкими функциями. Наложим теперь дополнительное условие. Пусть две локальные системы координат с координатами t и s пересекаются. Тогда на пересечении возникают две координаты, и одна из них будет функцией от другой, например $t = f(s)$. В обычном определении многообразия требуется только, чтобы функция f была гладкой. Мы сужаем класс функций и требуем, чтобы функция была дробно-линейной, т. е. $t = \dfrac{as+b}{cs+d}$. То, что в результате получится, называют проективной структурой на окружности. Вопрос такой: можно ли вообще ввести проективную структуру на окружности, а если можно, то сколькими способами?

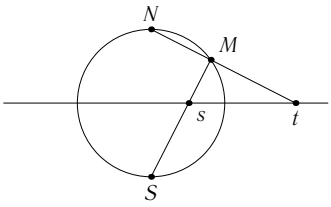


Рис. 3. Проективная структура на окружности

Как известно, окружность одной картой покрыть нельзя. Но с помощью стереографической проекции можно ввести две координаты на окружности (рис. 3). Пусть N — северный полюс, S — южный полюс, M — точка окружности. Точку M можно спроектировать на ось x из точки N и из точки S. В результате получим две координаты t и s. Координата t определена всюду, кроме северного полюса N; координата s определена всюду, кроме южного полюса S. Там, где обе координаты определены, они связаны соотношением $st = 1$. Это соотношение является дробно-линейным: $t = 1/s$. Таким образом, на окружности определена проективная структура. Оказывается, что на окружности есть и другие проективные структуры.

Теорема. *Классификация моментов для центрального расширения группы диффеоморфизмов окружности эквивалентна описанию проективных структур на окружности.*

Окружность — это не самое простое одномерное многообразие. Самое простое одномерное многообразие — это прямая. Попробуем решить задачу описания проективных структур на прямой. На прямой есть одна тривиальная проективная структура: на прямой можно ввести одну локальную координату x (она же будет глобальной координатой). Другой пример проективной структуры можно получить так. Отобразим прямую на окружность: $x \mapsto e^{ix}$. На окружности есть проективная структура; ее можно перенести на прямую. Получится не та проективная структура, которая задается координатой x. Эти проективные структуры не эквивалентны, потому что не существует функции, которая выражает x через координату на окружности и является дробно-линейной.

Теперь я приведу еще один пример проективной структуры на прямой, который даст возможность описать все проективные структуры на прямой. Прежде всего заметим, что интервал (0, 1) с координатой y, изменяющейся от 0 до 1, проективно не эквивалентен прямой. Как гладкие многообразия прямая и интервал — одно и то же: существует гладкая функция, отображающая прямую на интервал, для которой обратное отображение тоже гладкое. Но эта функция не может быть дробно-линейной. Ясно также, что луч с координатой z, изменяющейся от 0 до $+\infty$, проективно эквивалентен интервалу. Действительно, в качестве функции перехода можно взять $y = \dfrac{z}{z+1}$.

Обозначим прямую с тривиальной проективной структурой $\mathbb{R}$, а прямую с проективной структурой луча $\frac{1}{2}\mathbb{R}$. Изготовим теперь проективное многообразие, которое назовем $\frac{3}{2}\mathbb{R}$. Это делается следующим образом (рис. 4). Возьмем две прямые с координатами x и y и склеим положительный луч $(0, +\infty)$ первой прямой с отрицательным лучом $(-\infty, 0)$ второй прямой по следующему правилу: $x = -\dfrac{1}{y}$. Проективное многообразие $\frac{3}{2}\mathbb{R}$ содержит проективные подмногообразия $\frac{1}{2}\mathbb{R}$ и $\mathbb{R}$, но отлично от них.

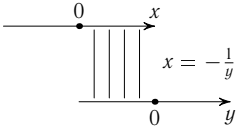


Рис. 4. Проективная структура $\frac{3}{2}\mathbb{R}$

Конструкцию можно продолжить дальше и получить проективное многообразие $2\mathbb{R}$. Возьмем для этого три прямые с координатами x, y, z и склеим еще дополнительно положительный луч второй прямой с отрицательным лучом третьей прямой по правилу $y = -\dfrac{1}{z}$

(рис. 5). Проективное многообразие $\frac{3}{2}\mathbb{R}$ не содержит проективного подмногообразия $2\mathbb{R}$.

Аналогично можно определить на прямой проективную структуру $\frac{m}{2}\mathbb{R}$, где m — любое натуральное число или ∞. На самом деле есть три варианта бесконечной прямой: бесконечная вправо, бесконечная влево и бесконечная в обе стороны. (Последнему случаю соответствует соответствует как раз та проективная структура, которая переносится с окружности на прямую). Можно доказать, что этим исчерпываются все проективные структуры на прямой. Инвариант, который различает эти проективной структуры, — минимальное количество карт, которыми можно покрыть всю прямую.

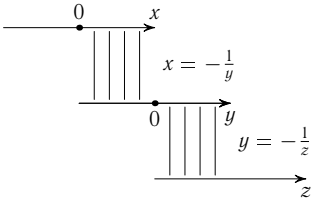


Рис. 5. Проективная структура $2\mathbb{R}$

Таков ответ для прямой. Для окружности все гораздо интереснее. Ответ для окружности равносилен описанию орбит коприсоединенного представления для группы диффеоморфизмов окружности.

Я много времени потратил на формулировку решенных задач, а теперь сформулирую одну нерешенную задачу. Рассмотрим группу $G = \mathrm{Diff}(D, \partial D, \sigma)$, где

$$D = \{(x, y) \in \mathbb{R}^2 \colon x^2 + y^2 \leqslant 1\}, \quad \partial D = \{(x, y) \in \mathbb{R}^2 \colon x^2 + y^2 = 1\},$$

а $\sigma = dx\,dy$ — форма площади. Здесь имеется в виду, что G — группа диффеоморфизмов диска D, гладко продолжающихся на границу ∂D и сохраняющих форму σ.

Пример преобразования из группы G — вращение диска. Не так уж просто написать явную формулу какого-нибудь другого преобразования из группы G, кроме

$$r \mapsto r, \quad \varphi \mapsto \varphi + f(r).$$

Для этой группы я хочу поставить задачу классификации коприсоединенных орбит и обсудить ее. Прежде всего нужно выяснить, как устроена алгебра Ли этой группы. Если бы мы брали все диффеоморфизмы, то получили бы алгебру векторных полей, которые на границе касаются границы. Мы же хотим рассматривать векторные поля, малый сдвиг вдоль которых сохраняет площадь. Как известно, такие векторные поля называют бездивергентными. Векторное поле

$$v = a(x, y)\frac{\partial}{\partial x} + b(x, y)\frac{\partial}{\partial y}.$$

бездивергентно, если

$$\operatorname{div} v = \frac{\partial a}{\partial x} + \frac{\partial b}{\partial y} = 0.$$

При таком условии малый сдвиг вдоль векторного поля сохраняет площадь.

Обсудим теперь, какова размерность рассматриваемой группы. Для бесконечномерных групп тоже есть понятие размерности, только не обычной, а функциональной. Для конечномерных пространств в топологии доказывается топологическая инвариантность размерности: пространства разной размерности не гомеоморфны. Для бесконечномерных многообразий ситуация не столь простая, но тем не менее все, кто занимается так называемым глобальным анализом, знают, что существует понятие функциональной размерности. Например, функции одной переменной образуют пространство функциональной размерности 1, а функции двух переменных образуют пространство функциональной размерности 2. Поэтому функцию двух переменных нельзя записать с помощью функций одной переменной. Мне могут возразить, что есть 13-я проблема Гильберта, которая была решена Колмогоровым и Арнольдом: им удалось выразить любую функцию n переменных через суперпозицию функций одной переменной. Но тут есть некоторый обман, потому что рассматриваются непрерывные функции, а непрерывные функции — вещь необозримая, не поддающаяся никакому описанию. Если же рассматривать функции гладкие или аналитические, то функция двух переменных содержит гораздо больше информации, чем функция одной переменной; любого конечного числа функций одной переменной не хватит, чтобы заменить функцию двух переменных. Так вот, группа диффеоморфизмов плоскости имеет функциональную размерность, соответствующую двум функциям от двух переменных, т. е. $2\infty^2$. Действительно, диффеоморфизм плоскости задается двумя функциями от двух переменных. Алгебра Ли в этом случае имеет такую же размерность: векторное поле на плоскости тоже задается парой функций. Но у нас есть дополнительное условие $\operatorname{div} v = 0$. В таком случае существует функция h, для которой $a = -\dfrac{\partial h}{\partial y}$ и $b = \dfrac{\partial h}{\partial x}$. (Векторное поле v называют косым градиентом функции h и обозначают $v = \text{s-grad}\, h$; косой градиент — это обычный градиент, повернутый на 90°.) Таким образом, рассматриваемая группа имеет размерность ∞^2.

Условие, что векторное поле касается границы, выражается очень просто: $h|_{\partial D} = \text{const}$. Действительно, у функции h косой градиент касается границы, поэтому обычный градиент перпендикулярен границе. Следовательно, граница является линией уровня, т. е. на границе функция постоянна. Косой градиент, так же как и обычный, не изменяется при добавлении к функции константы. Поэтому можно считать, что $h|_{\partial D} = 0$. Тогда алгебра

Ли $\mathfrak{g} = C^\infty(D, \partial D)$ — пространство гладких функций на диске, обращающихся в нуль на границе. Коммутатор в этой алгебре Ли — обычная скобка Пуассона $[f, g] = f'_x g'_y - f'_y g'_x$. Двойственное пространство $\mathfrak{g}^*$ состоит из обобщенных функций F, для которых $\langle F, f\rangle = \iint_D Ff\sigma$. Мы ограничимся гладкой частью пространства $\mathfrak{g}^*$, т. е. будем считать, что F — гладкая функция. Тогда коприсоединенное действие — это обычное действие диффеоморфизмов на функции.

Мы получаем такую задачу. Имеется гладкая функция на диске. Две функции считаются эквивалентными, если одну можно перевести в другую заменой переменных, сохраняющей площадь. Какие есть инварианты?

Эта задача совсем не тривиальна. Она, по-видимому, имеет обозримое решение. Я сейчас его намечу. С каждой функцией $f \in C^\infty(D, \partial D)$ связан интересный топологический инвариант Y_f — дерево компонент линий уровня функции f (рис. 6). Оно определяется так. Рассмотрим линию уровня. Она может состоять из нескольких компонент связности. Каждую компоненту связности будем считать отдельной точкой. На этом множестве есть естественная топология. Соответствующее топологическое пространство и есть дерево компонент уровня.

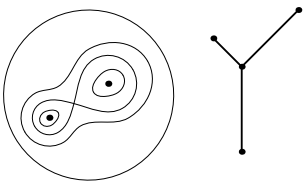


Рис. 6. Дерево компонент линий уровня

Ясно, что дерево компонент линий уровня является инвариантом относительно замен переменных, в том числе и относительно замен переменных, сохраняющих площадь. Для функций, у которых дерево компонент линий уровня имеет простейший вид (две вершины, соединенные ребром), есть ровно один инвариант относительно замен переменных, сохраняющих площадь. Этот инвариант не обычный, а функциональный. А именно, рассмотрим функцию

$$S(c) = \operatorname{area}\{x\colon f(x) \leqslant c\}.$$

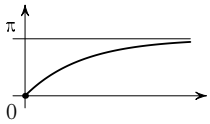


Рис. 7. График функции S

Если функция f неотрицательна, то $S(0) = 0$, $S(\infty) = \pi$ и график функции имеет такой вид, как на рис. 7. Не очень простая, но и не очень сложная теорема утверждает, что для функций с простейшим деревом компонент этот инвариант единствен. Это означает, что если известна функция S, то известна и функция f с точностью до замены переменных, сохраняющих площадь. А именно, замену переменных можно сделать так, чтобы функция f зависела только от радиуса (как именно она зависит от радиуса, легко вычислить, зная функцию S).

Соответствующая теорема предполагается верной и для любого дерева компонент, но эта теорема не доказана. Функцию S при этом нужно модифицировать, указывая отдельно ее значения для каждого ребра дерева. Такая функция по-прежнему является инвариантом, но не известно, есть ли другие инварианты.

Бесконечная группа Ли $\mathrm{Diff}(D, \partial D, \sigma)$ и ее алгебра Ли, по-видимому, являются богатым источником для новых исследований. Дело в том, что эта алгебра сама себе двойственна, потому что она имеет инвариантную относительно коприсоединенного представления форму $\langle F, f\rangle = \iint_D Ff\sigma$. Таким образом, присоединенное представление эквивалентно коприсоединенному. А это позволяет обращаться с такой алгеброй Ли, как с обычной компактной алгеброй Ли. И перенесение основных конечномерных конструкций на этот бесконечномерный случай — это очень перспективное занятие. В частности, описание многообразия флагов для этой группы — это очень вызывающая задача, на которую я хочу обратить внимание молодых математиков. Правда, здесь будет не одно многообразие флагов, а несколько, отвечающих разным типам деревьев.

Я еще не успел рассказать о симплектической структуре орбит и о комплексной структуре, которая тоже здесь иногда встречается.

А. А. Кириллов

Метод орбит за пределами групп Ли. Квантовые группы

Лекция 3 сентября 1999 года

Сегодня я буду говорить о квантовых группах. Сначала я расскажу о том, как я понимаю, что такое квантовая группа. Обычная группа Ли — это одновременно гладкое многообразие и группа. Структуру группы мы пока обсуждать не будем, а о структуре гладкого многообразия немного поговорим. Есть разные определения гладкого многообразия. Одно из них — алгебраическое, которое с вычислительной точки зрения часто бывает самым полезным. Общий принцип вычислений в математике состоит в том, чтобы сводить все вопросы к алгебраическим вопросам, а они уже решаются алгоритмическим путем. Как же заменить такую геометрическую конструкцию, как гладкое многообразие, чисто алгебраическим понятием? Для этого вместо гладкого многообразия M мы рассмотрим $A(M)$ — алгебру гладких (вещественных) финитных функций на M. Финитность означает, что функция равна нулю вне некоторого компактного множества. Если многообразие компактно, то говорить о финитности функций не нужно. Для компактных многообразий весь этот подход выглядит проще; теоремы короче формулируются и проще доказываются. Но чтобы результат был общим, я сформулирую его для всех многообразий.

Алгебра $A(M)$ топологическая; в ней определено понятие предела. На компактных многообразиях сходимость означает сходимость самих функций и всех их производных. Алгебра $A(M)$ полностью описывает многообразие M. Тем самым вся геометрия изгоняется, и остается одна алгебра.

Как можно восстановить многообразие M? Если есть другое многообразие N и задано гладкое отображение $\varphi\colon M \to N$, то можно построить встречное отображение на алгебрах функций $\varphi^*\colon A(N) \to A(M)$. А именно, функции $f \in A(N)$ сопоставляется $\varphi^*(f) = f \circ \varphi$. Здесь возникает вопрос, останутся ли финитные функции финитными? Для компактного многообразия ответ ясен, а для некомпактного финитная функция может перейти не в финитную. Поэтому для некомпактных многообразий нужно ограничить класс отображений и рассматривать только так называемые *собственные* отображения, для которых прообраз любого компактного множества

компактен. Тогда финитные функции будут переходить в финитные. Это не очень удобно, например потому, что отобразить некомпактное многообразие в компактное тогда вообще невозможно, поскольку прообраз самого компактного многообразия будет некомпактным многообразием.

Отображение φ^* является гомоморфизмом алгебр, поэтому геометрическое понятие гладкого отображения многообразий заменяется алгебраическим понятием. Самое замечательное здесь то, что любой гомоморфизм алгебр функций порождается гладким отображением многообразий. Тут тоже нужны некоторые уточнения, о которых я сейчас не буду говорить.

Рассмотрим сначала самый простой случай, когда многообразие M_0 состоит из одной точки. Ясно, что $A(M_0) = \mathbb{R}$; поэтому гладкое отображение $\varphi\colon M_0 \to M$ индуцирует гомоморфизм $\varphi^*\colon A(M) \to \mathbb{R}$. Частный случай только что сформулированной теоремы выглядит следующим образом.

Теорема 1. *Любой ненулевой непрерывный гомоморфизм*

$$\chi\colon A(M) \to \mathbb{R}$$

имеет вид χ_m, $m \in M$, *где* $\chi_m(f) = f(m)$.

Доказательство. Для гомоморфизма χ его ядро $\operatorname{Ker}\chi = \{f \in A(M)\colon \chi(f) = 0\}$ является идеалом; обозначим этот идеал I_χ.

Лемма. *Для любого нетривиального идеала в алгебре функций на многообразии существует точка многообразия, в которой все функции идеала обращаются в нуль.*

Доказательство. Предположим, что такой точки нет. Тогда для любой точки $m \in M$ существует такая функция $f_m \in I_\chi$, что $f_m(m) \neq 0$. Выберем окрестность U_m точки m так, чтобы в ней функция f_m была отлична от нуля.

Возьмем произвольную функцию $g \in A(M)$. Мы хотим доказать, что эта функция тоже принадлежит идеалу I_χ. Тогда $I_\chi = A(M)$, а это противоречит нетривиальности идеала.

По определению функция g финитна, т. е. существует компактное множество K, вне которого она обращается в нуль. Множество K можно покрыть конечным числом определенных выше окрестностей $U_1, \ldots, U_N$. Им соответствуют функции $f_1, \ldots, f_N$ из идеала. Рассмотрим функцию $f = \sum_{i=1}^{N} f_i^2$. Ясно, что $f > 0$ на K. Поэтому $g = fh$, где h — некоторая гладкая функция. Но $f \in I_\chi$, поэтому $g \in I_\chi$. □

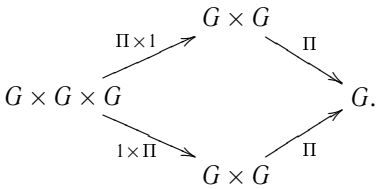
В рассматриваемом случае идеал имеет коразмерность 1, потому что вся алгебра отображается в одномерное пространство. Для таких идеалов точка, о которой идет речь в условии леммы, может быть только одна.

Тем самым, наш гомоморфизм является вычислением значения функции в данной точке. □

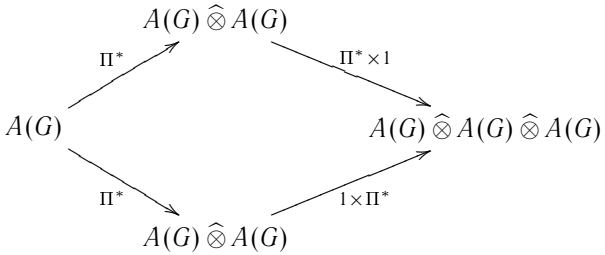
Итак, всю теорию гладких многообразий можно изложить в чисто алгебраических терминах, если вместо точек многообразия говорить об алгебре функций. Теперь мы сделаем самый важный шаг. Зададимся таким вопросом: нельзя ли вместо алгебры гладких функций рассматривать какие-нибудь более общие алгебры, например некоммутативные? Оказывается, что можно. То, что в результате получается, сейчас принято называть *некоммутативным многообразием*. Оно не является многообразием в классическом смысле; говорить о точках этого многообразия нельзя. Можно говорить только об алгебре функций на этом многообразии, а в качестве алгебры функций рассматривать произвольную алгебру, в том числе и некоммутативную. Смысл этого в том, что интуиция работы с гладкими многообразиями переносится на работу с некоммутативными алгебрами. Можно пробовать перенести конструкции, известные в теории гладких многообразий, на некоммутативные алгебры. При этом получается много интересного. В частности, можно определить, что такое некоммутативная группа Ли; здесь имеется в виду не то, что она некоммутативна как группа, а то, что она является некоммутативным многообразием.

Мы подошли к следующему определению: квантовая группа — это некоммутативное многообразие, которое одновременно является группой. Нужно только объяснить, что означает выражение «некоммутативное многообразие, которое одновременно является группой», потому что некоммутативное многообразие не является множеством. Мы должны сформулировать то свойство, что некоторое множество является группой, на языке не множеств, а алгебр функций на множествах. Для этого нужно не говорить о точках, а говорить только о функциях.

Обычное определение группы таково. Группа — это множество G, для которого задано отображение $G \times G \xrightarrow{\Pi} G$, обладающее определенными свойствами. Для группы Ли это отображение должно быть гладким. Нужно еще, чтобы переход к обратному элементу был гладким, но формулировку соответствующего утверждение я оставлю в качестве упражнения. А сейчас я переведу на язык функций свойство ассоциативности умножения. При обычном определении ассоциативность — это коммутативность следующей диаграммы:



Для алгебр функций получаем отображение $A(G) \xrightarrow{\Pi^*} A(G \times G)$. Если группа конечная, то $A(G \times G)$ алгебраически совпадают с обычным тензорным произведением $A(G) \otimes A(G)$. Для бесконечных групп это уже неверно. В этом случае нужно взять пополненное (в топологическом смысле) тензорное произведение $A(G) \mathbin{\widehat{\otimes}} A(G)$. Ассоциативность группы соответствует коммутативности диаграммы



Это определение имеет смысл и для некоммутативных многообразий.

Теперь можно определить *квантовую группу* как алгебру A, для которой задано такое отображение $A \xrightarrow{\Pi^*} A \mathbin{\widehat{\otimes}} A$, что соответствующая диаграмма коммутативна. Еще должно выполняться свойство, соответствующее существованию у каждого элемента обратного, но его формулировка остается в качестве упражнения.

Частным случаем квантовой группы является обычная группа Ли.

В действительности квантовыми группами называют очень специальный класс некоммутативных многообразий, для которых алгебра функций хотя и некоммутативна, но является небольшой деформацией коммутативной. Сформулированное выше определение годится для произвольных некоммутативных алгебр, но в такой общности мало что можно сказать. Наиболее интересные примеры требуют введения дополнительных структур. Поэтому давайте ограничим класс некоммутативных алгебр.

С чем связано название «квантовая»? Дело в том, что один из способов придать математический смысл слову «квантование» состоит в том, чтобы в нужном месте заменить коммутативную операцию некоммутативной. Есть и другие подходы. Например, некоторые считают, что квантование состоит в том, чтобы в некотором месте заменить непрерывный параметр на дискретный. Или заменить какие-нибудь простые конструкции на сложные. Слово «квантование» адекватного перевода на математический язык не имеет. Оно имеет несколько переводов, но эти переводы не эквивалентны. Что вполне естественно, поскольку физика и математика — это два разных языка, и объем понятий не обязан совпадать. Просто в математике нет такого слова, которое имело бы тот же объем понятий, что и слово «квантование» в физике.

Мы будем рассматривать специальный класс некоммутативных алгебр. А именно, рассмотрим пространство формальных степенных рядов от параметра h над некоторой коммутативной (и ассоциативной) алгеброй A. В этом пространстве можно ввести некоммутативное умножение следующим образом. Пусть $a = a_0 + ha_1 + h^2a_2 + \dots$ и $b = b_0 + hb_1 + h^2b_2 + \dots$ Эти элементы нужно перемножить. Их можно перемножить как формальные степенные ряды, но такое умножение коммутативно. Попробуем построить некоммутативное умножение, которое в нулевом приближении было бы устроено так же, а для членов с более высокими степенями h могло бы отличаться. Если мы будем считать, что h коммутирует с остальными элементами, то достаточно научиться перемножать элементы самой алгебры A. Степенные ряды тогда мы сможем перемножать. Определим теперь умножение так:

$$a \textcircled{h} b = a_0b_0 + h\{a, b\}_1 + h^2\{a, b\}_2 + \dots$$

Знак $\textcircled{h}$ символизирует, что умножение зависит от h. Каждая скобка $\{a, b\}_i$ — билинейная операция в алгебре. Свойство ассоциативности умножения накладывает на скобки много соотношений. В каком-то смысле эти соотношения можно разрешить. Я не буду описывать всю эту науку, которая называется теорией деформаций ассоциативных алгебр, а сформулирую только результаты, к которым она пришла.

Теория деформаций ассоциативных алгебр, как и все остальные науки, сильно зависит от понятия гомологий. Ее результаты записываются некоторыми классами когомологий, т. е. какими-то коциклами по модулю кограниц. На простом языке это означает, что мы могли бы сделать некоторую замену переменных в самой алгебре A, а именно, заменить степенной ряд новым степенным рядом, все члены которого однозначно определяются нулевым членом: $a = a_0 + hA_1(a_0) + h^2A_2(a_0) + \dots$, $b = b_0 + hA_1(b_0) + h^2A_2(b_0) + \dots$ Тогда получим новый закон умножения:

$$\{a, b\}_1^{\text{новое}} = \{a, b\}_1^{\text{старое}} + aA_1(b) - A_1(ab) + A_1(a)b.$$

Это изменение следует считать тривиальным: мы записали то же самое умножение в других координатах. Оказывается, что то уравнение, которому удовлетворяет скобка, говорит, что общая скобка является суммой такой тривиальной добавленной скобки и кососимметричного выражения. Поэтому с точностью до замены переменных можно считать, что умножение в первом члене антикоммутативно; симметричную часть можно убрать. В нулевом члене умножение всегда коммутативно, а в первом члене его можно сделать антикоммутативным. После этого оказывается, что для новой скобки ассоциативность умножения эквивалентна тождеству Якоби. Это можно было предвидеть, потому что есть общая теорема о том, что тождество Якоби — это нечетный аналог ассоциативности.

Для алгебры гладких функций на многообразии это выглядит так. Пусть $f_1, f_2 \in C^\infty(M)$. Тогда

$$f_1 \textcircled{h} f_2(m) = f_1(m) f_2(m) + h\{f_1, f_2\}(m).$$

Этот ряд можно продолжать и дальше, но достаточно интересно ограничиться первыми членами. Если мы наложим еще такое физически понятное требование, что носитель произведения должен содержаться в пересечении носителей сомножителей, то любая билинейная операция, не увеличивающая носитель, должна задаваться дифференциальным оператором. Если еще выполняется и тождество Якоби, то можно доказать, что в дифференциальном операторе не встречаются производные выше первого порядка. Поэтому скобка в локальных координатах имеет вид

$$c^{ij}(m)\, \partial_i f_1\, \partial_j f_2.$$

Константы c^{ij} антисимметричны по i и j и удовлетворяют квадратичному соотношению, вытекающему из тождества Якоби. Такая геометрическая структура на многообразии называется *структурой Пуассона*, а само многообразие с такой структурой — *пуассоновым многообразием*.

Геометрически пуассонова структура задается бивектором c^{ij}. Краткая запись такова: $c^{ij}\, \partial_i \wedge \partial_j$. Можно определить поливекторные поля на многообразии как выражения, которые в локальных системах координат имеют вид

$$c^{i_1 \ldots i_k}\, \partial_{i_1} \wedge \ldots \wedge \partial_{i_k}.$$

При переходе от одной системы координат к другой эти выражения преобразуются стандартным образом, как частные производные; а внешнее произведение тоже преобразуется стандартным образом, как антисимметричное ассоциативное умножение. Оказывается, что существует естественная билинейная операция

$$\mathrm{Vect}^k(M) \times \mathrm{Vect}^l(M) \to \mathrm{Vect}^{k+l-1}(M).$$

где $\mathrm{Vect}^k(M)$ — множество всех k-векторных полей на многообразии M. В частности, паре векторных полей сопоставляется снова векторное поле, а именно, обычный коммутатор векторных полей. Паре бивекторных полей сопоставляется тривекторное поле. Операция естественна в том смысле, что она не зависит от выбора локальной системы координат. На векторных полях операция антикоммутативна, а на бивекторных полях коммутативна. И всегда она либо коммутативна, либо антикоммутативна.

На бивекторах операция устроена так:

$$[c, c]^{ijk} = c^{s[i} \partial_s c^{jk]};$$

квадратные скобки в правой части означают антисимметризацию по верхним индексам. Тождество Якоби эквивалентно тому, что $[c, c] = 0$. Это легко

восстановить, если вспомнить, что

$$\{f_1, f_2\} = c^{ij}(m)\,\partial_i f_1\,\partial_j f_2,$$

и потом записать для левой части тождество Якоби. Получится кососимметричное выражение от трех функций, которое задает тривектор. Этот тривектор как раз и есть $[c, c]$.

Вернемся к квантовым группам. Я напомню, что квантовая группа связана с некоммутативным многообразием. У нас теперь есть первое приближение к некоммутативному многообразию: мы рассматривали только два члена ряда, а именно, нулевой (коммутативный) и первый (антикоммутативный). Такое квазиклассическое приближение к квантовой группе задается вполне классическим объектом — группой Ли с пуассоновой структурой. Такой объект называют *группой Пуассона—Ли*.

Группа Пуассона—Ли состоит из группы Ли G и бивекторного поля c на G, удовлетворяющего двум условиям:

- выполняется тождество Якоби $[c, c] = 0$;
- если L_x — левый перенос на x, R_y — правый перенос на y, то

$$c(xy) = L_x c(y) + R_y c(x). \tag{1}$$

Первое условие можно сформулировать для любого многообразия; не только для группы Ли. Второе условие соответствует ассоциативности умножения в переводе на язык гомоморфизмов алгебр функций.

Второе условие неудобно с точки зрения вычислений, потому что значение c является не функцией, а сечением некоторого расслоения. Значение бивектора c в данной точке x лежит в пространстве бивекторов, касательных к G в точке x. Поэтому значения в разных точках лежат в разных пространствах. Но группа Ли параллелизуема, поэтому можно ввести новую функцию, а именно, пусть

$$\gamma(x) = L_x^{-1} c(x).$$

Тогда γ — обычная функция на G со значениями в $\mathfrak{g} \wedge \mathfrak{g}$, потому что $c(x)$ принимает значения во внешнем квадрате касательного пространства в точке x, а $L^{-1}(x)$ переводит касательное пространство в точке x в касательное пространство в единице, т. е. в $\mathfrak{g}$. Для функции γ равенство (1) перепишется в виде

$$\gamma(xy) = \gamma(y) + \mathrm{Ad}\, y^{-1} \gamma(x). \tag{2}$$

Это — функциональное уравнение. Из него легко получить дифференциальное уравнение. Для этого положим $y = \exp(tY)$ и продифференцируем по t:

$$Y\gamma(x) = Y\gamma(1) - e^{\mathrm{ad}\, Y} Y(x). \tag{3}$$

Мы получили систему обыкновенных дифференциальных уравнений относительно функции $\gamma(x)$; в правую часть входит первая производная в начальной точке. Кстати сказать, если мы в (2) положим $x = y = 1$, то получим, что $\gamma(1) = 0$. В правую часть (3) входят только первые производные γ в точке 1, поэтому бивектор γ полностью определяется этим дифференциальным уравнением, если известны его первые производные в точке 1.

Первая производная бивектора в точке 1 — это тензор с тремя индексами: бивектор действует на функции, но еще сам он линейно зависит от касательного вектора. Все это вместе дает элемент пространства $\mathfrak{g}^* \otimes \mathfrak{g} \wedge \mathfrak{g}$, которое изоморфно $\mathrm{Hom}(\mathfrak{g}, \mathfrak{g} \wedge \mathfrak{g})$. Группа Пуассона—Ли полностью определяется этим элементом. Этот элемент должен удовлетворять определенным условиям, но вся информация о группе Пуассона—Ли в нем уже содержится. Эта ситуация во многом похожа на само понятие группы Ли. Известно, что группа Ли во многом определяется своей алгеброй Ли, а алгебра Ли определяется набором структурных констант. Структурные константы — это тоже тензор третьего ранга. Набор структурных констант дает отображение $\mathfrak{g} \wedge \mathfrak{g} \to \mathfrak{g}$. А для алгебры Пуассона—Ли получается отображение в обратную сторону: $\mathfrak{g} \to \mathfrak{g} \wedge \mathfrak{g}$. Но от одного легко перейти к другому, рассмотрев двойственное отображение. Для алгебры Пуассона—Ли двойственное отображение будет вида $\mathfrak{g}^* \wedge \mathfrak{g}^* \to \mathfrak{g}^*$. Таким образом, получается полная аналогия. Сама группа Ли задается структурными константами c_{ij}^k, а в группе Пуассона—Ли добавляется еще бивектор c_k^{ij}, задающий произведение на двойственном пространстве. Оказывается, что условия, которые накладываются на этот бивектор, гарантируют, что второй набор c_k^{ij} тоже является набором структурных констант в двойственном пространстве, т. е. определяемое им произведение тоже удовлетворяет тождеству Якоби.

Получается очень красивая алгебраическая конструкция. Имеется пространство $\mathfrak{g}$, которое является алгеброй Ли с коммутатором $[\ ,\ ]$, причем двойственное пространство $\mathfrak{g}^*$ тоже является алгеброй Ли с некоторым коммутатором $[\ ,\ ]_*$. Конечно, чтобы в результате получилась группа Пуассона—Ли, обе эти структуры должны быть между собой связаны. Алгебраисты занимались этим вопросом. Окончательное решение таково. Если мы рассмотрим отображение $\mathfrak{g} \to \mathfrak{g} \wedge \mathfrak{g}$ как 1-цепь на алгебре Ли $\mathfrak{g}$ со значениями в $\mathfrak{g} \wedge \mathfrak{g}$, то эта 1-цепь должна быть 1-циклом. Двойственное условие для отображения $\mathfrak{g}^* \to \mathfrak{g}^* \wedge \mathfrak{g}^*$ эквивалентно первому условию.

Десять лет назад была открыта замечательная интерпретация обоих этих равносильных условий, которую легко может запомнить даже человек, далекий от когомологий. Как проще всего согласовать коммутаторы в самом пространстве и в сопряженном пространстве? Начнем с того, что закодируем понятие сопряженного пространства. Для этого рассмотрим прямую сумму $\mathfrak{g} \oplus \mathfrak{g}^* = \mathfrak{D}$. Пространство $\mathfrak{D}$ имеет дополнительную

структуру: в нем можно определить билинейную форму, которая на $\mathfrak{g}$ обращается в нуль, на $\mathfrak{g}^*$ тоже обращается в нуль, а если один аргумент лежит в $\mathfrak{g}$, а другой в $\mathfrak{g}^*$, то берется обычное значение функционала на векторе. Эта форма на $\mathfrak{D}$ невырожденная, а два подпространства $\mathfrak{g}$ и $\mathfrak{g}^*$ изотропны относительно ее. Кроме того, отдельно на $\mathfrak{g}$ и отдельно на $\mathfrak{g}^*$ задана структура алгебры Ли. Оказывается, что все дополнительные условия с коциклами равносильны следующему условию: на $\mathfrak{D}$ существует структура алгебры Ли, которая сохраняет эту билинейную форму (т. е. все операторы ad кососимметричны относительно этого скалярного произведения) и ограничения этой алгебры Ли на $\mathfrak{g}$ и на $\mathfrak{g}^*$ совпадают с теми алгебрами Ли, которые были заданы на этих пространствах.

Инвариантность билинейной формы соответствует тому, что

$$([x, y], z) = ([y, z], x).$$

Окончательный объект, который отвечает группе Пуассона—Ли, — это так называемая *тройка Манина*. Она состоит из алгебры Ли $\mathfrak{D}$ с невырожденной инвариантной билинейной формой и из двух подалгебр $\mathfrak{g}$ и $\mathfrak{g}^*$, изотропных относительно этой билинейной формы. Этот набор алгебраических данных полностью кодирует такой геометрический объект, как группа Пуассона—Ли.

Теперь я расскажу о том, какое это имеет отношение к методу орбит. Рассмотрим тривиальную группу Пуассона—Ли, т. е. группу Ли G с бивектором $c = 0$. Ей соответствует тройка Манина $\mathfrak{g} \oplus \mathfrak{g}^*$, для которой коммутатор в $\mathfrak{g}$ тот же, что и в обычной алгебре Ли, коммутатор в $\mathfrak{g}^*$ нулевой. Билинейная форма такая, как и положено в прямой сумме пространства с двойственным. Уже поэтому коммутатор элементов из $\mathfrak{g}$ и из $\mathfrak{g}^*$ определяется однозначно из условия инвариантности формы. А именно, коммутатор — это коприсоединенное действие элемента из $\mathfrak{g}$ на функционал на алгебре Ли. Таким образом, алгебре Ли $\mathfrak{D}$ отвечает группа Ли $G \ltimes \mathfrak{g}^*$ — полупрямое произведение группы Ли G на пространство $\mathfrak{g}^*$, двойственное алгебре Ли. Групповое действие такое: в G — как и положено, $\mathfrak{g}^*$ — абелева группа по сложению, а G действует на $\mathfrak{g}^*$ коприсоединенно. Мы получаем основной объект метода орбит — коприсоединенное действие группы Ли G на пространстве $\mathfrak{g}^*$. Этот объект возникает, если мы рассматриваем тривиальную группу Пуассона—Ли с нулевым бивектором c. В общем же случае получается группа Ли D с двумя подгруппами G и G^*. Следует заметить, что в этой конструкции G и G^* играют полностью симметричную роль. Квантовые группы, которые связаны с деформациями, встречаются парами. С каждой группой G связана группа G^*. Если G — обычная группа Ли, то G^* — пространство, двойственное алгебре Ли.

Алгебра Ли группы D является прямой суммой алгебр Ли групп G и G^*, поэтому в первом приближении группа D является произведением G и G^*,

но в целом это не так. Это примерно так, как разложить матричную группу в произведение верхних треугольных матриц и нижних треугольных: для матриц общего вида это можно сделать, а для исключительных нельзя. Так что не каждый элемент D является произведением элемента из G и элемента из G^*.

Я считаю, что полезно изучить действие элементов G не только на пространстве $\mathfrak{g}^*$, но и на группе G^*. Это уже будет нелинейное действие.

Если $G = \mathrm{SO}(3)$ или $\mathrm{SL}(2)$, то G^* — разрешимая группа. Этот факт имеет общую природу: группа, двойственная простой группе Ли, всегда разрешима. Группа G^* зависит от того, какие берутся скобки. Но для простой группы G можно описать все группы G^*. Это, как я уже говорил, связано с некоторыми коциклами, а у простой группы все коциклы тривиальны, т. е. все коциклы — кограницы. Поэтому нужный коцикл можно явно написать. Если этот коцикл (который является кограницей) достаточно невырожден, то получается довольно явное описание. Оно получено Белавиным и Дринфельдом.

Для $G = \mathrm{SO}(3)$ получается одна двойственная группа, а для $\mathrm{SL}(2)$ их три, потому что в $\mathrm{SL}(2)$ есть элементы трех типов: эллиптические, параболические и гиперболические.

Совершенно неизученная задача — попробовать исследовать тройку Манина, построенную на бесконечномерной алгебре Ли, а именно, на той алгебре, которая обсуждалась на прошлой лекции. В качестве алгебры Ли $\mathscr{A}$ я предлагаю взять алгебру гладких комплексных функций на диске D, обращающихся в нуль на границе, а в качестве инвариантного скалярного произведения взять

$$(f_1, f_2) = \mathrm{Im} \int_D f_1 f_2 \, d\sigma.$$

В конечномерной ситуации любая алгебра Ли соответствует некоторой группе Ли. А эта бесконечномерная алгебра Ли не соответствует никакой группе Ли; в бесконечномерной ситуации есть много других таких примеров. Зато $\mathscr{A}$ можно представить в виде прямой суммы подпространств $\mathfrak{g}$ и $\mathfrak{g}^*$, которые изотропны относительно указанного скалярного произведения и являются алгебрами Ли, причем этим алгебрам Ли соответствуют группы Ли. А именно, в качестве $\mathfrak{g}$ возьмем вещественные функции $C^\infty_{\mathbb{R}}(D, \partial D)$ (это пространство изотропно, поскольку мы берем мнимую часть интеграла, а для вещественных функций мнимая часть интеграла равна нулю). Пространство $\mathfrak{g}^*$ строится следующим образом. Запишем функцию на диске в виде

$$f = f(r, \varphi) = \sum_{k \in \mathbb{Z}} c_k(r) e^{ik\varphi}.$$

Удобно заменить r на другую координату. В механике есть канонические координаты, называемые действие—угол. В них форма записывается в виде произведения дифференциалов координат. В нашем случае

$$\sigma = r\,dr \wedge d\varphi = d(\pi r^2) \wedge d\left(\frac{\varphi}{2\pi}\right) = dS \wedge d\theta;$$

здесь S — площадь (в механике — действие), θ — нормированный угол. Положим

$$\mathfrak{g}^* = \left\{ f = \sum_{k \geqslant 0} c_k(S) e^{2\pi i k\theta} \right\}.$$

Операция в алгебре Ли — скобка Пуассона. Скобку Пуассона можно брать в любых координатах, лишь бы площадь в них записывалась, как обычно, $dx\,dy$. В частности, S и θ можно взять за канонические координаты. Тогда

$$\{f_1(S)e^{2\pi i k\theta},\ f_2(S)e^{2\pi i l\theta}\} = 2\pi i\big(l f_1'(S) f_2(S) - k f_1(S) f_2'(S)\big) e^{2\pi i (k+l)\theta}.$$

Числа k и l складываются, поэтому функции, для которых $k \geqslant 0$, образуют подалгебру (положительные гармоники образуют подалгебру).

Итак, из алгебры Ли $\mathscr{A}$ можно изготовить тройку Манина. Пространство здесь бесконечномерное, поэтому все привычные конечномерные конструкции приходится заново осмысливать. Все это пока совершенно не исследовано.

В науке часто бывает, что, когда известно, что задача очень трудная, никто не берется ее решать. А потом приходит молодой человек, который не знал, что эта задача трудная. Он берет и решает ее. Его спрашивают: «Как же ты ее решил? Ведь она трудная.» А он отвечает: «А я не знал, что она трудная.» Поэтому я призываю тех людей, которые пока не могут оценить всех трудностей, которые здесь возникают, заняться этой задачей.

И. М. Кричевер

Конформные отображения и уравнение Уизема

Лекция 23 декабря 1999 года

Тема, которой посвящена первая часть лекции, всем известна со студенческой скамьи. Моей конечной целью будет — показать, как связаны теория интегрируемых уравнений, которая активно развивалась последние 20 лет, и теория Уизема, которая имеет уже 10-летнюю историю, с классической задачей комплексного анализа. Теорема Римана гласит, что если на комплексной плоскости есть область с границей, содержащей более чем две точки, то существует конформное отображение этой области на единичный диск. Эта теорема носит характер теоремы существования. Построением конкретных таких конформных отображений занимаются многие из прикладных наук, и не только из прикладных наук, потому что с этими задачами связаны приложения в гидродинамике, в теории нефтяных месторождений, в аэродинамике. Во многих ситуациях возникает необходимость строить конформные отображения тех или иных областей.

Я хочу в какой-то степени рассказать о недавнем замечательном наблюдении Забродина и Вигмана, которые пару месяцев тому назад обнаружили связь между классической задачей о конформных отображениях областей и бездисперсионной линеаризованной цепочкой Тодда. Я расскажу о том, какое развитие это получило в нашей совместной работе, которая пока не опубликована. А именно, о том, как это обобщается для неодносвязных областей и какую роль в этом играют методы алгебраической геометрии.

Прежде чем перейти собственно к этой задаче, я хочу рассказать обо всем контексте, в котором она возникла, чтобы было яснее, что такое уравнение Уизема. Удивительным образом одинаковые структуры, связанные с уравнением Уизема, возникают в разных областях математики. Помимо конформных отображений они возникают в задаче об n-ортогональных криволинейных координатах, которая была центральной задачей дифференциальной геометрии в XIX веке. Пусть в $\mathbb{R}^n$ задана криволинейная система координат $x^i(u)$, где x^i — декартовы координаты, выраженные через криволинейные координаты u. Эту систему координат называют

n-ортогональной, если все гиперповерхности уровня $u_i = \text{const}$ пересекаются под прямым углом. Примером такой системы координат служат полярные координаты. В 2-мерном случае задача тривиальна, но начиная с размерности 3 эта задача становится очень богатой. Теоретически она была решена Дарбу, опять же на уровне теоремы существования. Он доказал, что локально задача построения n-ортогональной системы криволинейных координатах зависит от $n(n-1)/2$ функций двух переменных. Известно довольно много конкретных примеров n-ортогональных систем координат. Один из таких примеров — эллиптические координаты. По существу, решение конкретной системы дифференциальных уравнений — это и есть удачное построение системы координат, в которой система становится тривиальной. Поэтому так важны хорошие системы координат: у нас появляется больше шансов решить уравнение.

Задачу об n-ортогональных системах координат можно переформулировать во внутренних терминах как задачу отыскания диагональных метрик $ds^2 = \sum H_i^2(u)(du)^2$, которые являются плоскими. Егоров рассматривал такие метрики, для которых выполняется дополнительное условие $H_i^2 = \partial_i \Phi$. Это — условие симметрии метрики. Такие метрики называют *метриками Дарбу—Егорова*. Этот класс метрик обладает многими специальными свойствами.

И эта задача, и уравнение Уизема, и задача о конформных отображениях — это все единый комплекс идей и методов. Немного позже я расскажу о том, как задача об n-ортогональных криволинейных координатах связана с топологическими квантовыми моделями теории поля.

Другой сюжет, который в конечном итоге объединил все эти разноплановые задачи, — это теория интегрируемых уравнений, или, как теперь говорят, — солитонных уравнений. Эта теория возникла примерно 30 лет назад. Самое известное из солитонных уравнений, которое и возникло первым, — уравнение Кортевега—де Фриза (КдФ)

$$u_t - \frac{3}{2}uu_x + \frac{1}{4}u_{xxx} = 0.$$

Имеется также много других солитонных уравнений, по счастью имеющих важное прикладное значение, которые интегрируемы методами теории солитонов. Я не буду рассказывать об этих методах; они в основном развивались лет 10—20 тому назад и по-прежнему продолжают развиваться.

Простейшее решение уравнения КдФ было известно Кортевегу и де Фризу почти в момент написания этого уравнения. Это — стационарное решение $u(x, t)$, не зависящее от t. В этом случае $u_t = 0$ и уравнение можно проинтегрировать:

$$\frac{3}{4}u^2 = \frac{1}{4}u_{xx} + g_2.$$

Затем можно умножить на u_x и снова проинтегрировать:

$$\frac{1}{2}(u_x)^2 = u^3 + g_2 u + g_3.$$

Решения такого уравнения выражаются через функцию Вейерштрасса:

$$u = 2\wp(x + \text{const}; \omega_1, \omega_2),$$

Функция Вейерштрасса $\wp$ — это двоякопериодическая функция с периодами $2\omega_1$ и $2\omega_2$ и с полюсом второго порядка в нуле: $\wp = \frac{1}{x^2} + O(x)$. Это решение зависит от трех констант, т. е. мы получили полный набор решений уравнения 3-го порядка.

Стационарное решение строится по эллиптической кривой — кривой рода 1. В 70-е годы и в начале 80-х годов в цикле работ Дубровина, Новикова, Матвеева, моих и ряда других возникли так называемые алгебро-геометрические методы построения решений солитонных уравнений, которые по набору алгебро-геометрических данных выдают решение разных нелинейных уравнений, включая уравнение КдФ, sin-Гордон и другие уравнения в рамках этой науки. Решение возникает как результат обработки данных некой машиной, называемой конечнозонной теорией интегрирования. Решение тоже записывается в явном виде, но только обычно не через эллиптические функции, а через тэта-функции Римана. Алгебро-геометрический набор данных состоит из римановой поверхности Γ_g рода g с фиксированными точками $P_1, \ldots, P_N$ и с фиксированными локальными координатами $z_1, \ldots, z_N$ в окрестностях этих точек и еще из фиксированной точки комплексного многомерного тора $J(\Gamma_g)$ — якобиана этой поверхности. По такому набору данных строится решение. В зависимости от того, сколько выбрать точек, какие классы кривых выделить, получаются решения самых разных уравнений.

Для стационарного решения уравнения КдФ алгебро-геометрический набор данных включает эллиптическую кривую $y^2 = E^3 + g_2 E + g_3$ и фиксированную точку на бесконечности. Эта часть данных играет роль интегралов — они не изменяются с течением времени. А точка на якобиане движется со временем. Фазовое пространство уравнения выглядит таким образом: есть пространство интегралов, которые являются кривыми с отмеченными точками и фиксированными локальными координатами в этих точках, и над каждой точкой пространства интегралов висит тор. Движение на торе — прямолинейная обмотка, в полном соответствии с духом теории вполне интегрируемых конечномерных систем, т. е. с теорией Лиувилля.

Так выглядит ответ для солитонных уравнений. Как строятся решения — это отдельная история. Об этом я сейчас не буду говорить. Я хочу рассказать, о том, что происходило в этой науке потом — начиная с середины 80-х годов. Тогда основной упор переместился на теорию возмущений интегрируемых уравнений. Обычно нас интересует не только какое-то

вполне конкретное уравнение, но и то, что происходит в его окрестности. Основной элемент теории возмущений интегрируемых уравнений — это и есть теория Уизема.

Прежде чем перейти к теории Уизема, я хочу написать одну формулу, которую мы потом будем в разных видах встречать во всех перечисленных выше науках. Как я уже сказал, описание движения для солитонных уравнений посредством системы интегралов и прямолинейной обмотки тора вполне согласуется с теорией Лиувилля. Конечная цель теории Лиувилля — построение переменных «действие—угол». Гамильтонова система строится по многообразию M^{2n} (фазовое пространство), симплектической структуре ω на нем и гамильтониану H. Гамильтонова система называется вполне интегрируемой, если помимо гамильтониана есть набор n интегралов в инволюции: $\{F_i, F_j\} = 0$. Тогда компактные поверхности уровня этих интегралов должны быть n-мерными торами, а движение должно быть прямолинейной обмоткой тора. На торе есть естественные координаты — циклы. Если Φ_i — угловые координаты для базисных циклов, то переменными действия называют координаты A_i, которые канонически сопряжены угловым переменным, т. е. симплектическая структура в этих координатах приобретает стандартную форму Дарбу: $\omega = \sum dA_i \wedge d\Phi_i$. Отдельная нетривиальная задача — как выделить такие системы координат среди всех других систем координат. Теорема Лиувилля в изложении Арнольда дает такой ответ: нужно примитивную форму проинтегрировать по базисным циклам. Но непонятно, как этот n-мерный тор в $2n$-мерном многообразии описать явным образом. Так что эта теорема тоже носит характер скорее теоремы существования. Явным образом построить переменные «действие—угол» не удалось. Замечательным было наблюдение Новикова и Веселова в начале 80-х годов. Анализируя первые известные к тому времени интегрируемые гамильтоновы уравнения, они обнаружили, что во всех этих задачах переменные «действие—угол» всегда имеют один и тот же вид. А именно, интегрировать нужно не по циклу в n-мерном пространстве, а по циклу на соответствующей римановой поверхности:

$$A_i = \oint_{a_i} Q\, dE. \tag{1}$$

Здесь Q — некоторый мероморфный дифференциал; каждой гамильтоновой системе соответствует свой дифференциал Q. Эти дифференциалы могут быть и многозначными. Почему так происходит, было не известно. Новиков и Веселов называли это аналитическими скобками Пуассона. Объяснение природы этих формул было получено совсем недавно, 3 года назад, в нашей совместной работе с Fong'ом (в Journal of Differential Geometry) после того, как мы стали анализировать ответы для симплектических структур, которые возникают в теории Виттена—Зайберга по суперсим-

метричной модели Янга—Миллса. Оказалось, что те же самые симплектические скобки, которые описывают случай гиперэллиптических кривых (я должен сказать, что все то, что рассматривали Новиков и Веселов, относится к случаю гиперэллиптических кривых), были переоткрыты Виттеном и Зайбергом.

Формулу (1) нужно запомнить, потому что точно такой же интеграл многозначного дифференциала решает задачу о конформных отображениях областей.

Что же такое уравнение Уизема? Предположим, что мы немножко изменили уравнение — возмутили его. Тогда интегралы исходного уравнения перестанут быть интегралами. Они будут медленно изменяться; как говорят физики, они станут адиабатическими интегралами. Для невозмущенного уравнения точка фазового пространства бегает по тору. Как только мы возмутим уравнение, начинается медленный дрейф вдоль пространства интегралов. Возникает система уравнений на пространстве модулей кривых с отмеченными точками, которая описывает это движение. Эти уравнения и есть уравнения Уизема. Оказалось, что они сами интегрируемы.

Я больше не буду говорить об алгебро-геометрических данных и перейду на совсем элементарный уровень, когда рассматриваются только кривые рода 0. Дело в том, что самое простое решение уравнения КдФ — это не то решение, которое использует функцию Вейерштрасса; самое простое решение — это константа. В теории уравнения КдФ этим решением никто не интересовался; родом 0 пренебрегали ввиду его полной тривиальности. Но если рассматривать не постоянное решение, а его возмущения, то теория становится содержательной. В теории Уизема род 0 стал играть нетривиальную роль. Этот случай называется бездисперсионным пределом солитонных уравнений. К нему можно относиться как к частному случаю более общей задачи, а можно и заниматься им отдельно.

С чем связано название «бездисперсионный»? В уравнении КдФ коэффициенты не существенны, потому что масштабными преобразованиями его можно привести к виду $u_t = uu_x + u_{xxx}$. В дальнейшем я не буду следить за коэффициентами. Если решение почти постоянное, то про третью производную можно забыть. Хорошим приближением будет уравнение $u_t = uu_x$. Это уравнение и называют бездисперсионным пределом, потому что в уравнении КдФ член u_{xxx} отвечает за дисперсию. Уравнение $u_t = uu_x$ — простейший пример уравнения Уизема.

Если уравнение КдФ — это бесконечномерный аналог интегрируемых (по Лиувиллю) гамильтоновых систем, то уравнение $u_t = uu_x$ тоже интегрируемо, только совсем в другом смысле. Ничего не стоит написать решение уравнения $u_t = uu_x$ (его называют уравнением Римана—Хопфа). А именно, возьмем произвольную функцию $f(\xi)$ и запишем уравнение $u = f(x + ut)$. Это уравнение в неявном виде определяет функцию $u(x, t)$. Эта функция

$u(x, t)$ — решение уравнения $u_t = uu_x$. Более того, все решения так получаются.

На основе этого решения обычно объясняют роль нелинейности в гидродинамике. Если воспринимать u как высоту (амплитуду волны), то видно, что точка движется со скоростью, пропорциональной высоте. Поэтому если функция не монотонна, то «горб» начинает обгонять все остальное, волна становится все более крутой, а потом опрокидывается (рис. 1). В том месте, где происходит опрокидывание, уже нельзя пренебрегать третьими производными; они там становятся большими. Гидродинамика воспринимает это так, что дисперсия (вязкость) регуляризует поведение волны.

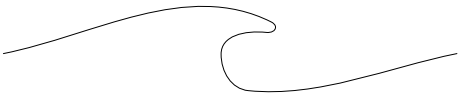


Рис. 1. Опрокидывающаяся волна

Все уравнения Уизема интегрируются сходным методом — пишется некое неявное выражение для решения.

Какова более общая постановка бездисперсионных уравнений Лакса? Я напомню, что в основе построения решений солитонных уравнений лежит представление Лакса $\dot{L} = [L, A]$. Для уравнения КдФ $L = \partial^2 + u(x, t)$ (оператор Штурма—Лиувилля) и $A = \partial^3 + \frac{3}{2}u\partial + \frac{3}{4}u_4$. Обобщения уравнения КдФ возникли, когда стали рассматривать операторы с матричными коэффициентами и операторы более высокого порядка:

$$L = \partial^n + u_{n-2}\partial^{n-2} + \ldots + u_0.$$

Представление Лакса — это условие совместности переопределенной системы линейных задач $L\psi = E\psi$, $L_2\psi = A\psi$. Вообще, идея метода обратной задачи заключается в том, чтобы идти не от уравнения, а в обратную сторону: исходя из функции ψ строить оператор и решение.

Коль скоро мы договорились для начала рассматривать самые простейшие решения лаксовых уравнений (когда u — константа), очень просто решить линейное дифференциальное уравнение. Решение — экспонента. Тогда собственные значения будут полиномами. Если взять собственную функцию $\psi = e^{px}$, то $E(p) = p^n + u_{n-2}p^{n-2} + \ldots + u_0$ (символ соответствующего дифференциального оператора). Уравнение Уизема записывается в виде $\partial_i E = \{E_+^{i/n}, E\}$; здесь $\{f, g\} = f_p g_x - f_x g_p$ — скобка Пуассона. Мы будем выражать $u_i(X, T)$ через медленные переменные $X = \varepsilon x$ и $T = \varepsilon t$.

Появление индекса i здесь не случайно. Каждое интегрируемое уравнение возникает не само по себе, а как часть большой иерархии — целой совокупности коммутирующих с ним интегралов. Это в духе интегрируе-

мости по Лиувиллю: если есть набор интегралов в инволюции, то каждый из них, рассматриваемый как гамильтониан, порождает свою гамильтонову динамику. То, что интегралы в инволюции, означает, что соответствующие динамики коммутируют.

Объясню теперь, что такое $E_+^{i/n}$. Запишем лорановское разложение $E^{1/n}(p) = p + \sum v_i p^{-i}$. Тогда $E^{i/n}(p) = p^i + \ldots + O(p^{-1})$; $E_+^{i/n}$ означает, что здесь берутся только неотрицательные степени p, т. е. нужно вычеркнуть $O(p^{-1})$. Получаем полином, коэффициенты которого — полиномы от u. Поэтому получается замкнутая система уравнений — бездисперсионный предел лаксовых уравнений. В простейшем случае, если $E = p^2 + u$ и $i = 3$, то получаем уравнения Римана—Хопфа, которые я записывал раньше.

Как выглядит общая процедура решения для бездисперсионного предела? Рассмотрим пространство пар (Q, E), где полиномы E и Q имеют вид соответственно: $p^n + u_{n-2}p^{n-2} + \ldots + u_0$ и $Q = b_0 p + \ldots + b_{m-1}p^m$. На этом пространстве можно ввести координаты Уизема $T_i = \frac{1}{i}\operatorname{res}_\infty(E^{-i/n}QdE)$. Так определенные T_i являются функциями от u и b (от b они зависят линейно, а от u полиномиально). Эти T_i обращаются в нуль при больших i; ненулевых T_i ровно столько, сколько нужно. Локально T_i как функции от u и b можно обратить и получить функции $u(T)$, $b(T)$. Подставив эти значения $u(T)$ в E, получим функцию $E(T)$. Утверждение таково: $E(T)$ — решение уравнения бездисперсионной иерархии. При этом Q, вроде бы, выступает как некий вспомогательный объект. Но $Q(T)$ тоже решение того же самого уравнения с тем же самым гамильтонианом, а именно, $\partial_i Q = \{E_+^{i/n}, Q\}$. Более того, оказывается, что всегда $\{Q, E\} = 1$. Уравнение $\{Q, E\} = 1$ называют *струнным уравнением*.

В отличие от обычной иерархии лаксовских уравнений, в бездисперсионном пределе выживает только одно решение: все решения параметризуются разными высшими временами; общее решение удовлетворяет подходящему струнному уравнению.

Бездисперсионная наука была известна уже несколько лет, когда появилась работа Дирграфа, Верлинде и Виттена, которые занимались совсем другой задачей — классификацией топологических моделей теории поля. Решая эту задачу, они написали те же самые формулы совершенно в другом контексте. Стало ясно, что за всей этой бездисперсионной наукой скрывается еще один очень важный элемент, который теперь называют тау-функцией. Вся структура, связанная с бездисперсионным пределом уравнения КдФ или общего лаксовского уравнения, закодирована в одной функции

$$F(t) = \frac{1}{2}\operatorname{res}_\infty(T_i k^i dS).$$

Здесь $dS = QdE = \sum_{i=1}^{\infty} T_i dk^i + O(k^{-1})$ и $k = E^{1/n}(p) = k(p) = p + O(p^{-1})$. Я напомню, что мы имеем дело со случаем кривой рода 0; отмеченную точку можно загнать на бесконечность. Единственный выживающий параметр — локальная координата p. Можно проверить, хотя это и далеко не очевидно, что производные функции F по временам T_i дают все остальные коэффициенты. Например, $\partial_i F = \operatorname{res}_\infty(k^i\, dS)$ и $\partial^2_{ij} F = \operatorname{res}_\infty(k^i\, d\Omega_j)$, где $\Omega_j = E_+^{j/n}$. Есть совершенно замечательная формула

$$\partial^3_{ijk} F = \sum_{q_s} \operatorname{res}_\infty \Big(\frac{d\Omega_i\, d\Omega_j\, d\Omega_k}{dQ\, dt} \Big),$$

где суммирование ведется по критическим точкам полинома E: $dE(q_s) = 0$.

Теперь я хочу вернуться к исходной задаче о конформных отображениях. Я буду рассматривать только случай областей, ограниченных аналитическими кривыми. Внутреннюю область обозначим D, а внешнюю $\overline{D}$. Меня будут интересовать однолистные конформные отображений внешности единичного круга на $\overline{D}$. Для чтения на эту тему я хочу рекомендовать книгу Варченко и Этингофа (А. Н. Варченко, П. И. Этингоф. Почему граница круглой капли превращается в инверсный образ эллипса. М.: Наука, 1995). Там построено много красивых конкретных примеров конформных отображений, которые связаны со следующей задачей, возникающей в нефтяной отрасли. Представьте себе, что область — это месторождение нефти. Имеется несколько скважин, через которые нефть откачивают. Тогда область как-то деформируется. Уравнение, описывающее динамику границы области таково. Пусть Φ — решение уравнения

$$\Delta \Phi = \sum q_i\, \delta(z - z_i)$$

с нулевым граничным условием: $\Phi|_{\partial D}$. Тогда $\operatorname{grad} \Phi$ — скорость движения границы.

Эта задача в каком-то смысле интегрируема. Оказывается, что конечная форма капли не зависит от того, в каком порядке производится откачивание нефти через скважины, как и полагается для коммутирующих потоков. Результат зависит только от того, сколько нефти откачано через каждую скважину; в каком порядке это делается — не имеет значения.

Основной вклад в эту науку внес Ричардсон, который обнаружил бесконечный набор интегралов. О них мы сейчас и поговорим.

Довольно просто доказать, что любая область (односвязная или неодносвязная) полностью задается своими гармоническими моментами. Гармонические моменты области D определяются следующим образом. Пусть $u(x, y)$ — гармоническая функция. Тогда гармонический момент равен

$$t_u = \iint_D u(x, y)\, dx\, dy.$$

Если область меняется, то меняется гармонический момент для некоторой функции. Это локальное утверждения. Гармонические моменты — локальные координаты.

Не обязательно рассматривать все гармонические моменты; можно ограничиться некоторыми функциями. Например, набор функций

$$t_n = \iint_D z^{-n}\, dz\, d\bar{z}, \quad n \geqslant 1,$$

вместе с функцией

$$t_0 = \iint_{\overline{D}} dz\, d\bar{z},$$

где $\overline{D}$— внешняя область, для односвязной области является локальным набором координат.

Основное наблюдение, которое сделали Вигман и Забродин, таково. Рассмотрим еще моменты дополнения

$$u_n = \iint_{\overline{D}} z^n\, dz\, d\bar{z}.$$

Ясно, что функции v_n выражаются через $t_0, t_1, \ldots$ Оказывается, что при этом

$$\frac{\partial v_n}{\partial t_m} = \frac{\partial v_m}{\partial t_n}.$$

Это означает, что существует функция $F(t)$, для которой $\partial_n F(t) = v_n$. При этом оказывается, что $\partial_0\, \partial_n F$ — коэффициенты разложения однолистной функции, осуществляющей конформное отображение. Предполагается, что эта функция нормирована следующим образом. В дополнении к единичному кругу есть координата w, а в $\overline{D}$— координата z. Мы рассматриваем отображение внешностей и предполагаем, что бесконечность переходит в бесконечность и, более того, $z = w + O(w^{-1})$. В таком случае $w(z) = z + \sum(\partial_0\, \partial_n F) z^{-n}$. Снова оказывается, что все конформные отображения закодированы одной функцией. Эта функция та же самая, о которой я говорил раньше.

Прежде всего я хочу дать новое доказательство того, что локально координаты t_n являются полной системой координат. Из этого доказательства будет видно, как это все связать с бездисперсионной наукой.

Мне потребуется понятие функции Шварца. Локально гладкую кривую можно задать в виде $y = f(x)$. В комплексном виде это можно записать так: $\bar{z} = S(z)$. Функцию S называют функцией Шварца. Например, для единичной окружности получаем уравнение $\bar{z} = z^{-1}$.

Для вещественно аналитической кривой (без углов) функцию S можно продолжить до комплексно аналитической функции в малой окрестности кривой.

Первое утверждение, которое я хочу доказать, таково. Предположим, что контур деформируется, т. е. имеется семейство функций Шварца $S(z, t)$, где t — параметр деформации. Тогда если при такой деформации не изменяются все гармонические моменты t_n, то кривая неподвижна, т. е. деформация тривиальна.

Утверждение: 1-дифференциал $S_t(z, t)\,dz$ чисто мнимый на контуре ∂D, т. е. все его значения на касательных к контуру векторах чисто мнимые. Это легко следует из определения функции Шварца.

Следующее **утверждение** уже использует специфику рассматриваемых координат t_n. Если $\partial_t t_n = 0$, то голоморфный дифференциал $\partial_t S\,dz$, определенный в малой окрестности кривой, продолжается до голоморфного дифференциала на всей внешности.

Прежде чем доказывать второе утверждение, я объясню, как из этих двух утверждений следует то, что нужно. Если есть любая область $D \subset \mathbb{C}$ с координатой z, то по ней можно построить замкнутую риманову поверхность. Для этого нужно взять другой экземпляр той же самой области с координатой $\bar{z}$ и склеить их по границе. Полученную риманову поверхность называют дублем Шоттки. Воспользуемся принципом симметрии Шварца: если есть функция, аналитическая в верхней полуплоскости и вещественная на вещественной оси, то ее можно аналитически продолжить в нижнюю полуплоскость. У нас есть голоморфный дифференциал в D. Он продолжается на комплексное сопряжение, потому что на границе он чисто мнимый. В результате получаем голоморфный дифференциал на сфере. Но на сфере нет голоморфных дифференциалов, отличных от нуля.

Займемся доказательством утверждения о том, что голоморфный дифференциал $\partial_t S\,dz$ продолжается на всю внешность. Интеграл Коши позволяет представить любую функцию на гладком контуре в виде разности функции, голоморфной во внешности, и функции, голоморфной во внутренности. Пусть

$$\widehat{S}(z) = \oint_{\partial D} \frac{\partial_t S(w)\,dw}{z - w}.$$

Функция $\widehat{S}(z)$ голоморфна вне контура и продолжается на границу и $S^+ - S^- = \partial_t S$. Если начало координат находится внутри области и $|z| < |w|$, то

$$\widehat{S}(z) = \sum z^n \oint_{\partial D} \partial_t S(w) w^{-n}\,dw = \sum z^n \partial_t t_n,$$

потому что по теореме Стокса

$$t_n = \iint_D z^n\,dz\,d\bar{z} = \oint_{\partial D} z^{-n}\bar{z}\,dz.$$

Если не меняются моменты, то коэффициенты разложения S тождественно нулевые. Поэтому функция S^- тождественный нуль в некоторой

окрестности. Но эта функция аналитическая, поэтому она тождественно равна нулю. Чтобы $\partial_t S\,dz$ был голоморфным, нужно чтобы еще один коэффициент был нулевым. Потому что мы умножили эту функцию на dz, а этот дифференциал имеет полюс второго порядка.

Теперь ясно, что изменится, если мы будем дифференцировать по t_n. Это утверждение было для любой переменной. Коэффициенты разложения теперь будут не всюду тождественными нулями; один из коэффициентов будет ненулевым. Это означает, что, например, $\partial_{t_0} S\,dz$ — мероморфный дифференциал с простым полюсом на бесконечности. Такой дифференциал только один. Поэтому мы доказали, что $\partial_0 \bar z\,dz = \dfrac{dw}{w}$. Здесь мы дифференцируем $\bar z$ при постоянном z. Эквивалентная запись такова:

$$\{z(w, t_0), \bar z(w, t_0)\} = 1.$$

У нас получился мероморфный дифференциал с простым полюсом. Когда мы возьмем дубль, по принципу симметрии возникнет второй полюс. Возникнет дифференциал, который имеет вычет ± 1 в двух точках. Это глобальное свойство, как теорема Лиувилля. Если есть аналитическая функция на компактной поверхности, то она — константа. Эти два факта позволяют использовать глобальные свойства. Первый факт позволяет с области с границей переходить на компактную поверхность. А второй факт, и это уже требует специальных свойств, позволяет это аналитически продолжать до мероморфного объекта.

Это утверждение про нулевой момент. А утверждение про все остальные моменты выглядит следующим образом: $\partial_n \bar z\,dz = dz_+^n$. Здесь используются такие обозначения. Пусть $z(w) = w + \dots$ Тогда $z^n(w) = w^n + \dots$ Плюс означает, что берется только положительная часть (полином на сфере).

Меня могут спросить, почему дифференциал имеет полюс только в одной точке, хотя по принципу симметрии он должен иметь полюс и в симметричной точке. Но функции t_n не аналитические; это функции и от вещественной и от мнимой части: $t_n = x_n + iy_n$. При этом

$$\frac{\partial}{\partial t_n} = \frac{\partial}{\partial x_n} - i\,\frac{\partial}{\partial y_n}.$$

Следовательно,

$$\frac{\partial}{\partial x_n}\,\bar z\,dz = dz_+^n - d\bar z_+^n$$

и

$$\frac{\partial}{\partial y_n}\,\bar z\,dz = i(dz_+^n + d\bar z_+^n).$$

Дело в том, что можно писать иерархии по t_n и по комплексно сопряженной переменной $\bar t_n$. Получается линеаризованная цепочка Тодда.

Можно написать замечательную формулу:

$$F(t) = -\frac{t_0^2}{2} + \sum_{n \geqslant 0} (n-2)(t_n v_n + \bar{t}_n \bar{v}_n).$$

Эта формула содержит массу нетривиальных тождеств. Например, тождество $\partial_n F = v_n$ выглядит почти наивно. Но сами v_n зависят от t_n загадочным образом. Когда эти зависимости подставишь и продифференцируешь, то получается как раз v_n.

Функцию F можно явно вычислить для эллипса:

$$F = \frac{1}{2} t_0^2 \ln t_0 - \frac{3}{4} t_0^2 - \frac{1}{2} \ln(1 - 4|t_2|^2) + t_0 \frac{|t_1|^2 + t_1^2 \bar{t}_2 + \bar{t}_1^2 t_2}{1 - 4 t_2 \bar{t}_2}.$$

Этот пример показывает, как F зависит от первых трех моментов. (У дополнения к эллипсу только три ненулевых момента.)

Для неодносвязных областей первое утверждение (о производной функции Шварца) остается справедливым. Второй факт опирался на суммирование геометрической прогрессии для интеграла Коши. В конце 80-х годов мы с Новиковым, занимаясь оператором квантования бозонных струн, построили теорию Лорана—Фурье на произвольной римановой поверхности. Базис z^n заменяется на другой базис.

Написанная формула симметрична по t и v. Это наводит на мысль попытаться применить ее к старой классической задаче. Как по однолистному конформному отображению дополнений построить отображение самих областей? Например, отображение дополнения эллипса на дополнение круга — простая алгебраическая функция. А отображение внутренности эллипса на внутренность круга задается эллиптической функцией. Так что связь между этими отображениями нетривиальна.

В. Ю. Овсиенко

Проективная дифференциальная геометрия — старая и новая

Лекция 6 января 2000 года

1. Группы симметрий и дифференциальные инварианты

Я буду рассказывать о проективной и дифференциальной геометрии и некоторых классических теоремах из теории гладких кривых, которые принято сейчас называть теорией Штурма.

Пусть M — гладкое многообразие. Следуя Клейну (или Тёрстону), под дифференциальной геометрией на M будем понимать изучение инвариантов действия группы Ли G (это действие задано хотя бы локально). Я сразу приведу несколько примеров.

Пример 1. $M = \mathbb{R}^n$ с действием евклидовой группы $\mathscr{E}(n) = \mathrm{SO}(n) \ltimes \mathbb{R}^n$ (полупрямое произведение группы вращений на группу переносов).

Евклидова геометрия изучает инварианты действия группы $\mathscr{E}(n)$. Основной инвариант — метрика $g = \sum (dx^i)^2$. Кроме того, по метрике можно написать всевозможные кривизны (кривых, поверхностей и т. д.).

Пример 2. $M = \mathbb{R}^n$ с действием аффинной группы $\mathscr{A}(n) = \mathrm{GL}(n) \ltimes \mathbb{R}^n$.

В этом случае не существует $\mathscr{A}$-инвариантной метрики. Тем не менее, понятие кривизны может быть определено.

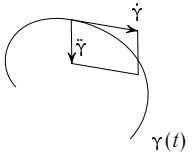


Рис. 1. Кривая γ

Рассмотрим для простоты случай $n = 2$. Пусть $\gamma(t)$ — локально выпуклая кривая, т. е. у нее нет перегибов. Тогда векторы $\dot\gamma$ и $\ddot\gamma$ не коллинеарны (рис. 1). В евклидовой геометрии кривизну можно определить как инвариант кривой, различающий кривые с точностью до евклидовых преобразований. Так же можно поступить и в аффинном случае. Мы предположили, что векторы $\dot\gamma$ и $\ddot\gamma$ не коллинеарны. Запишем вектор $\dddot\gamma$ в виде их линейной комбинации:

$$\dddot\gamma = a\dot\gamma + b\ddot\gamma.$$

Кривая $\gamma(t)$ определяется функциями $a(t)$ и $b(t)$. Выберем параметр t

так, чтобы избавиться от одной из этих функций. А именно, выберем t так, чтобы площадь параллелограмма, натянутого на векторы $\dot\gamma$ и $\ddot\gamma$ была постоянна: $[\dot\gamma, \ddot\gamma] = \text{const}$. Продифференцировав это равенство, получим $[\dot\gamma, \dddot\gamma] = 0$, т. е. $\dddot\gamma = k(t)\dot\gamma$. Функцию $k(t)$ называют *аффинной кривизной*. Можно доказать, что две кривые аффинно эквивалентны тогда и только тогда, когда соответствующие аффинные кривизны равны.

Отметим, что условие $[\dot\gamma, \ddot\gamma] = \text{const}$ аффинно инвариантно, поскольку форма площади аффинно инвариантна с точностью до умножения на константу.

Упражнение. Доказать, что

$$k(t) = \frac{[\dddot\gamma, \ddot\gamma]}{[\dot\gamma, \ddot\gamma]}.$$

Пример 3. $M = \mathbb{R}P^n$ — проективное пространство, т. е. множество прямых в $\mathbb{R}^{n+1}$, проходящих через начало координат.

Этот пример будет для нас основным.

Удобно выбирать аффинные координаты на произвольной аффинной гиперплоскости, не проходящей через начало координат. Почти все точки проективного будут параметризованы точками пересечения прямых с этой гиперплоскостью. Исключение составляют точки, принадлежащие проективному подпространству коразмерности 1.

На линейном пространстве $\mathbb{R}^{n+1}$ действует группа $G = \mathrm{SL}(n+1, \mathbb{R})$ (группа линейных преобразований, сохраняющих объем). Это действие переносится на $\mathbb{R}P^n$, потому что прямые переходят в прямые.

По нашему определению проективная геометрия должна изучать всевозможные инварианты геометрических объектов в проективном пространстве (кривых, подмногообразий, диффеоморфизмов) относительно действия этой группы.

Раньше проективная геометрия была необычайно популярна. Чем интересна проективная геометрия? Легко видеть, что в группу $\mathrm{SL}(n+1, \mathbb{R})$ вложены обе предыдущие группы — евклидова и аффинная. Оказывается, что в некотором смысле группа проективных симметрий максимальна. А именно, никакая другая группа Ли, содержащая $\mathrm{SL}(n+1, \mathbb{R})$, не может действовать на n-мерном многообразии даже локально. Поэтому если нам удастся определить проективные инварианты (кривизну и т. п.), то это будут самые сильные инварианты — инварианты по отношению к максимальной группе; их найти труднее всего.

Есть еще один случай, когда группа симметрий максимальна в этом же смысле. Это — группа конформных преобразований. Есть и другие примеры максимальных групп; классификации максимальных геометрий посвящено много работ.

В размерности 1 конформная геометрия и проективная геометрия совпадают. Мы начнем с этого простейшего случая — рассмотрим геометрию проективной прямой.

2. Производная Шварца

Фиксируем на проективной прямой $M = \mathbb{R}P^1$ координату $x = \frac{u}{v}$ (рис 2). Группа $G = \mathrm{SL}(2, \mathbb{R})$ действует на проективной прямой следующим образом:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} x = \frac{ax+b}{cx+d}.$$

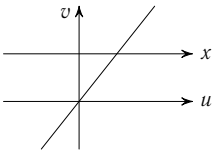


Рис. 2. Координаты на проективной прямой

Центр $\{\pm 1\}$ действует тривиально. Факторгруппа, которая действует на $\mathbb{R}P^1$ эффективно, — это группа

$$\mathrm{PSL}(2, \mathbb{R}) = \mathrm{SL}(2, \mathbb{R})/\{\pm 1\}.$$

В данном случае единственными естественными объектами, инварианты которых можно изучать, являются диффеоморфизмы $f\colon \mathbb{R}P^1 \to \mathbb{R}P^1$. Для такого диффеоморфизма определим *производную Шварца* следующей формулой:

$$S(f) = \frac{f'''}{f'} - \frac{3}{2}\left(\frac{f''}{f'}\right)^2.$$

Смысл этого сложного выражения состоит в следующей классической теореме.

Теорема 1. *Диффеоморфизмы f и g проективно эквивалентны* $\left(\text{т.е. } g(x) = \frac{af(x)+b}{cf(x)+d}\right)$ *тогда и только тогда, когда $S(f) = S(g)$.*

В одну сторону доказательство простое. А доказательство в другую сторону трудное.

Замечание 1. В одномерном случае евклидовы преобразования сводятся к переносам, поэтому два диффеоморфизма прямой f и g эквивалентны относительно евклидовой группы тогда и только тогда, когда $f(x) = g(x) + \mathrm{const}$. Эквивалентное условие: $f' = g'$. Поэтому в евклидовом случае диффеоморфизмы различает обычная производная.

Упражнение. Доказать, что на аффинной прямой роль производной играет логарифмическая производная f''/f'.

Выясним теперь, откуда возникает производная Шварца. Надо сказать, что этот объект необычайно универсален — существуют десятки или даже сотни мест, из которых он возникает. Я выбрал среди них два наиболее классических и наиболее простых.

Обычная производная измеряет, насколько диффеоморфизм f не сохраняет расстояния между двумя близкими точками x и $x+\varepsilon$:

$$f(x+\varepsilon)-f(x)=\varepsilon f'(x)+(\varepsilon^2).$$

В проективной геометрии расстояние между двумя точками не имеет никакого смысла, поскольку любые две (и даже три) точки проективным преобразованием можно перевести в любые другие. Но у четверки точек x_1, x_2, x_3, x_4 на проективной прямой существует (и притом единственный) инвариант — двойное отношение

$$[x_1, x_2, x_3, x_4]=\frac{(x_3-x_1)(x_4-x_2)}{(x_2-x_1)(x_4-x_3)}.$$

Традиционно проективная геометрия начинается с этого инварианта, который был известен еще древним грекам. В проективной геометрии двойное отношение играет такую же роль, как расстояние в евклидовой геометрии. И производная Шварца тоже возникает таким же способом, как обычная производная. Я нашел это определение у Э. Картана, но, вероятно, оно было известно и раньше. А именно, рассмотрим точки x, $x+\varepsilon$, $x+2\varepsilon$, $x+3\varepsilon$. Применим диффеоморфизм f и посчитаем разность двойных отношений. Оказывается, что

$$[f(x), f(x+\varepsilon), f(x+2\varepsilon), f(x+3\varepsilon)]-[x, x+\varepsilon, x+2\varepsilon, x+3\varepsilon]=-2\varepsilon^2 S(f)+(\varepsilon^3).$$

Но исторически производная Шварца возникла (в XIX в.) по-другому. Рассмотрим простейшее дифференциальное уравнение 2-го порядка:

$$y''(x)+u(x)\cdot y(x)=0,\quad u(x)\in C^\infty(\mathbb{R}).$$

Возьмем произвольные линейно независимые решения $y_1(x)$ и $y_2(x)$ и рассмотрим функцию $f(x)=\dfrac{y_1(x)}{y_2(x)}$ (в области, где $y_2(x)\neq 0$). Тогда $u(x)=\dfrac{1}{2}S(f)$. Это — второй способ определить производную Шварца.

Пространство решений двумерно. Поэтому если $\tilde{y}_1(x)$ и $\tilde{y}_2(x)$ — другая пара линейно независимых решений, то

$$\tilde{f}(x)=\frac{\tilde{y}_1(x)}{\tilde{y}_2(x)}=\frac{ay_1(x)+by_2(x)}{cy_1(x)+dy_2(x)}=\frac{af(x)+b}{cf(x)+d}.$$

Из этого можно получить доказательство теоремы 1 в другую сторону. Но доказательство того, что пространство решений двумерно, трудное, поэтому получается, что доказательство теоремы в обратную сторону трудное.

Это только два из многих способов получить производную Шварца. В действительности она обладает необыкновенным количеством разных свойств и используется во многих науках: в комплексном анализе, в теории динамических систем, математической физике.

3. Замечательное свойство кривизн

Классическая теорема о четырех вершинах утверждает следующее. «Пусть γ — выпуклая замкнутая кривая на плоскости. Тогда евклидова кривизна имеет по крайней мере 4 критических точки, т. е. функция кривизны имеет по крайней мере 4 экстремума.» Экстремальные точки кривизны называют *вершинами*.

Эту теорему доказал в 1909 г. индийский математик Mukhopadhyaya.

Давайте установим две геометрических способа понимать, что такое вершина кривой. Пусть кривая γ общего положения. Для каждой точки кривой можно построить соприкасающуюся окружность, которая приближает кривую с точностью до второго порядка. Радиус соприкасающейся окружности равен $1/k$, где k — кривизна. Поскольку окружность приближает кривую с точностью до второго порядка, кривая из внешней части окружности переходит во внутреннюю (рис. 3). Так будет во всех точках, которые не являются вершинами. И только в вершинах кривая локально будет расположена по одну сторону от окружности, потому что в вершинах окружность приближает кривую с точностью до третьего порядка.

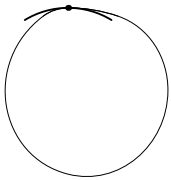


Рис. 3. Соприкасающаяся окружность

Другой способ геометрически определить вершины связан с так называемыми *каустиками* кривых. Каустика — это огибающая семейства нормалей к кривой (рис. 4). Точки возврата каустики соответствуют вершинам кривой. Каустика в случае эллипса была известна еще Аполлонию и Якоби.

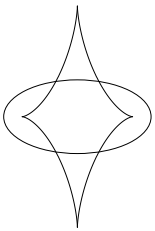


Рис. 4. Каустика

Теорема о четырех вершинах, несмотря на ее простоту, привлекает внимание до сих пор. Известны десятки ее различных доказательств и множество ее обобщений. Одно из обобщений — аффинная теорема о шести вершинах, которую Mukhopadhyaya доказал в той же самой работе 1909 г. Теорема о шести вершинах утверждает, что аффинная кривизна имеет по крайней мере 6 критических точек.

Точки экстремума аффинной кривизны (аффинные вершины) имеют такие же геометрические описания, как и точки экстремума обычной (евклидовой) кривизны, о которых шла речь выше. Во-первых, аффинная вершина — это точка необычайно тесного контакта кривой и коники. Во-вторых, аффинные вершины можно определить с помощью аффинной каустики. Аффинная каустика — это огибающая семейства аффинных нормалей, а аффинные нормали — это касательные к кривым, заметаемым серединами

хорд, параллельных касательной (рис. 5). Аффинные вершины соответствуют точкам возврата аффинной каустики.

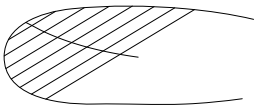


Рис. 5. Аффинная каустика

У эллипса все аффинные нормали проходят через центр. Это — вырожденный случай, как и окружность в евклидовой геометрии. Для эллипса соприкасающаяся коника совпадает с ним самим, так же как и для окружности соприкасающаяся окружность совпадает с ней самой.

Недавно В. И. Арнольд предложил доказательство теоремы о четырех вершинах, использующее симплектическую топологию.

4. Теорема Э. Жиса и нули производной Шварца

В 1995 г. Этьен Жис (E. Ghys) из Лиона обнаружил следующий удивительный факт.

Теорема 2. *Пусть f — произвольный диффеоморфизм проективной прямой $\mathbb{R}P^2$. Тогда производная Шварца $S(f)$ обращается в нуль по крайней мере в четырех различных точках.*

На первый взгляд ничего общего, кроме числа 4, между теоремой Жиса и теоремой о четырех вершинах нет. Но Жис открыл ее именно как аналог теоремы о четырех вершинах.

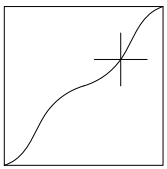


Рис. 6. График диффеоморфизма

Рассмотрим вместо диффеоморфизма f его график — замкнутую гладкую кривую на двумерном торе $\mathbb{T}^2 = \mathbb{R}P^1 \times \mathbb{R}P^1$. Эта кривая нигде не вертикальна и нигде не горизонтальна (рис. 6). Если $f \in \mathrm{PSL}(2, \mathbb{R})$, то график — гипербола. Оказывается, что в этой ситуации гиперболы играют ту же самую роль, что и соприкасающиеся окружности. А именно, для любой точки x существует единственное проективное преобразование $g_x \in \mathrm{PSL}(2, \mathbb{R})$, которое приближает диффеоморфизм f с точностью до 2-струи. Производная Шварца соответствует величине отклонения диффеоморфизма в данной точке от про-

ективного. Нули производной Шварца — это те точки, в которых проективное преобразование приближает f с точностью до 3-струи. Жису это дало способ получить первое доказательство его теоремы, которое следует классическому доказательству Кнезера теоремы о четырех вершинах, основанному на соприкасающихся окружностях.

Доказательство теоремы Жиса было опубликовано С. Табачниковым и автором.

5. Связь с лоренцевой геометрией

Теорема Жиса на самом деле не аналог теоремы о четырех вершинах, а в точности теорема о четырех вершинах, но в лоренцевой геометрии. Рассмотрим на торе плоскую лоренцеву метрику $g = dx\,dy$; конус изотропных векторов при этом — вертикальные и горизонтальные направления. Вычислим лоренцеву кривизну кривой $\gamma = (x, f(x))$ (эта кривая — тот самый график диффеоморфизма, который мы рассматривали). Параметризуем ее параметром t. Тогда $\dot\gamma = (\dot x, f'\dot x)$. Выберем параметр t так, что $\|\dot\gamma\| \equiv 1$, т. е. $\dot x = \frac{1}{\sqrt{f'}}$. Лоренцева кривизна равна длине вектора второй производной. Вычислим его:

$$\ddot\gamma = \left(\frac{1}{\sqrt{f'}}, \sqrt{f'}\right)^{\cdot} = \frac{1}{2}\left(-\frac{f''}{(f')^2}, -\frac{f''}{f'}\right).$$

Таким образом, лоренцева кривизна равна

$$\|\dot\gamma\| = \frac{1}{2}\,\frac{f''}{(f')^{3/2}} = k(\gamma).$$

Из этого мы получаем весьма неожиданное выражение

$$k' = \frac{1}{2\sqrt{f'}} S(f). \tag{1}$$

Оказалось, что производная Шварца связана с кривизной в лоренцевой метрике. Экстремумы кривизны — нули производной Шварца.

Замечание 2. Выражение (1) можно записать в более элегантном виде: $2dk\,dt = S(f)$ (как квадратичный дифференциал). Это — инвариантное выражение. Можно задать вопрос: «Для каких лоренцевых метрик выполняется такое соотношение?» Можно доказать, что оно выполняется в точности для лоренцевых метрик постоянной кривизны

$$g = \frac{dx\,dy}{(axy + bx + cy + d)^2}.$$

6. Решающее вмешательство проективной геометрии

Пока все результаты принадлежали разным геометриям — евклидовой, аффинной, лоренцевой. Как же все это связано с проективной геометрией? Сформулируем одно общее утверждение, из которого все будет следовать.

Рассмотрим гладкую замкнутую кривую $\gamma \subset \mathbb{R}P^n$. В 1956 г. M. Barner ввел понятие строго выпуклых кривых и доказал замечательную теорему. Назовем замкнутую кривую γ в проективном пространстве *строго выпуклой*, если через любую $n-1$ точку кривой γ можно провести гиперплоскость, которая пересекает γ только в этих точках. Пример строго выпуклой кривой в $\mathbb{R}P^2$ изображен на рис. 7.

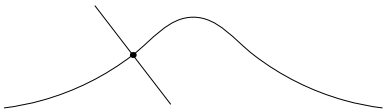


Рис. 7. Строго выпуклая кривая в $\mathbb{R}P^2$

Теорема Барнера состоит в следующем.

Теорема 3. *Любая строго выпуклая кривая в* $\mathbb{R}P^n$ *имеет по крайней мере* $n+1$ *точку уплощения.*

Напомню, что точка уплощения — это точка, в которой касательная гиперплоскость приближает с более высоким порядком, чем обычно, т. е. векторы $\dot\gamma$, $\ddot\gamma$, …, $\gamma^{(n)}$ линейно зависимы. В 2-мерном случае точка уплощения — это точка перегиба.

Покажем, что все предыдущие теоремы сводятся к теоремы Барнера.

Начнем с теоремы о четырех вершинах. Для этого рассмотрим отображение Веронезе $v\colon \mathbb{R}^2 \to \mathbb{R}P^3$, которое задается формулой

$$(x, y) \mapsto (x^2 + y^2 : x : y : 1).$$

Это отображение устанавливает взаимно однозначное соответствие между окружностями на плоскости $\mathbb{R}^2$ и гиперплоскостями в $\mathbb{R}P^3$. Вершины любой выпуклой кривой γ на плоскости $\mathbb{R}^2$ соответствуют точкам уплощения кривой $v(\gamma)$ в $\mathbb{R}P^3$. Действительно, рассмотрим для кривой γ соприкасающуюся окружность. Уравнение окружности — это в точности уравнение гиперплоскости в $\mathbb{R}P^3$. Образ $\mathbb{R}^2$ в $\mathbb{R}P^3$ — некоторая двумерная поверхность. Образ окружности — пересечение этой двумерной поверхности с гиперплоскостью. Вершина кривой (по крайней мере в случае общего положения) соответствует ситуации, когда кривая локально остается по одну

сторону от соприкасающейся окружности. В таком случае образ кривой остается локально по одну сторону от гиперплоскости, потому что образ $\mathbb{R}^2$ локально разделяет $\mathbb{R}P^3$ на две части. Это как раз соответствует точке уплощения.

Легко также проверить, что образ выпуклой кривой при отображении Веронезе является строго выпуклой кривой. Действительно, для выпуклой кривой на плоскости любая прямая, проходящая через две ее точки, не пересекает кривую в других точках. Прямая — частный случай окружности, поэтому ее образ тоже лежит в некоторой гиперплоскости. Эта гиперплоскость обладает требуемым свойством, т. е. пересекает $v(\gamma)$ ровно в двух точках.

Для доказательства теоремы о шести вершинах нужно рассмотреть отображение $\mathbb{R}^2 \to \mathbb{R}P^5$, которое задается формулой

$$(x, y) \mapsto (x^2 : xy : y^2 : x : y : 1).$$

Это отображение устанавливает взаимно однозначное соответствие междуониками и гиперплоскостями. Аффинные вершины — это точки, в которых соприкасающаяся коника приближает кривую лучше, чем обычно. Таким коникам соответствуют гиперплоскости, которые приближают образ кривой лучше, чем обычно. Такие гиперплоскости соответствуют точкам уплощения.

Для доказательства теоремы Жиса нужно рассмотреть отображение Сёгре $\mathbb{R}P^1 \times \mathbb{R}P^1 \to \mathbb{R}P^3$, которое задаётся формулой

$$\big((x : y), (y : z)\big) \mapsto (xz : xt : yz : yt).$$

Оказывается, что нули производной Шварца соответствуют точкам уплощения образа. Это легко понять, воспользовавшись интерпретацией нулей производной Шварца, основанной на соприкасающихся гиперболах.

Современное доказательство теоремы Барнера и вывод из нее четырех вершинах можно найти в недавней работе С. Табачникова.

7. Дискретизация

О. Мусин и В. Седых (1996) доказали дискретный аналог теоремы о четырех вершинах. Пусть P — выпуклый m-угольник на плоскости. Рассмотрим окружность, проходящую через три его последовательные вершины v_i, v_{i+1} и v_{i+2}. Вершины v_{i-1} и v_{i+3} могут лежать либо по разные стороны от этой окружности, либо по одну сторону. Если они лежат по одну сторону, то будем говорить, что тройка последовательных вершин *экстремальная*.

Теорема 4. *Для любого выпуклого m-угольника, где $m \geqslant 4$, найдутся по крайней мере 4 экстремальные тройки.*

Упражнение. Сформулировать дискретную версию теоремы о шести аффинных вершинах.

Сформулируем теперь дискретную версию теоремы Жиса. Заменим диффеоморфизм f на пару упорядоченных наборов точек в $\mathbb{R}P^1$; обозначим их $X=(x_1, \ldots, x_m)$ и $Y=(y_1, \ldots, y_m)$. График f при этом заменится на ломаную. Производную Шварца заменим на разность двойных отношений

$$[x_i, x_{i+1}, x_{i+2}, x_{i+3}]-[y_i, y_{i+1}, y_{i+2}, y_{i+3}].$$

Дискретная версия теоремы Жиса утверждает, что эта разность меняет знак по крайней мере 4 раза.

Теорема Барнера тоже дискретизируется. Определение строго выпуклого m-угольника то же самое. Доказательство дискретной версии теоремы Барнера можно провести индукцией по числу вершин многоугольника.

Интерес к дискретизации связан с тем, что дискретные теоремы сильнее гладких: гладкие теоремы получаются из дискретных предельным переходом. В то же время, доказательство многих дискретных теорем проще: оно может быть получено индукцией по числу вершин.

У теоремы о четырех вершинах есть две дискретные версии. Другая ее дискретная версия, которую предложил Wegner, отличается от теоремы Мусина и Седых тем, что описанные окружности заменяются на вписанные.

Самое поразительное, что существует дискретная версия теоремы о четырех вершинах, которая почти на 100 лет старше самой теоремы. Это — знаменитая лемма Коши (1813), которую он придумал специально для доказательства теоремы Коши о жесткости выпуклых многогранников.

Лемма (Коши). *Пусть P и P' — выпуклые m-угольники, $m \geqslant 4$. Предположим, что соответственные стороны этих многоугольников равны. Тогда разность соответствующих углов этих многоугольников меняет знак не менее четырех раз.*

Коши дал неверное доказательство этой леммы. Его доказательство исправил Адамар.

Автор искренне признателен В. Прасолову за подробный конспект и подготовку этой лекции к печати.

С. В. Матвеев

Метод нормальных поверхностей Хакена и его применения к классификации 3-мерных многообразий — история одной теоремы

Лекция 7 сентября 2000 года

В развитии математики есть верстовые столбы — проблемы, при решении которых мы приходим к новым открытиям и к пониманию природы математических объектов. В топологии одна из таких проблем — классификация трехмерных многообразий. В исследовании этой проблемы были следующие важнейшие этапы:

- 1960 Хакен (Haken): чтобы понять структуру трехмерного многообразия, нужно изучать в нем двумерные поверхности $F^2 \subset M^3$. Этот метод оказался очень продуктивным. Более половины работ в трехмерной топологии, появившихся с тех пор, основаны на этом методе. Классификация достаточно больших многообразий была получена именно с помощью этого метода.
- 1980 Gabai: нужно изучать слоения — это дает информацию о многообразиях.
- 1980 Thurston: нужно изучать геометрии на многообразиях — гиперболические многообразия, эллиптические многообразия, солвмногообразия и т. д., всего есть 8 однородных геометрий.
- 1990 Witten: нужно рассматривать статистические суммы и с их помощью строить инварианты трехмерных многообразий.
- В последнее время усилился интерес к алгоритмической топологии. Здесь метод Хакена играет ключевую роль.

В чем заключается метод Хакена? Пытаться рассматривать все поверхности в трехмерном многообразии безнадежно — их слишком много. Даже если рассматривать поверхности с точностью до шевеления (изотопии), то их все равно слишком много. Поэтому естественно выделить некоторый класс $\mathcal{N}$ поверхностей в многообразии так, чтобы он обладал двумя свойствами:

(1) класс $\mathcal{N}$ информативен (например, в таком смысле: этот класс содержит все интересные поверхности, расположенные в данном многообразии);

(2) класс $\mathcal{N}$ имеет явное описание.

Обозначение $\mathcal{N}$ связано с названием «нормальная поверхность». Теперь я объясню, что такое нормальная поверхность. В дальнейшем всегда предполагается, что многообразие M^3 триангулировано, т. е. разбито на тетраэдры так, что любые два тетраэдра либо не пересекаются, либо пересекаются по вершине, ребру или грани. Грубо говоря, нормальная поверхность (относительно фиксированной триангуляции) определяется так: она пересекает все тетраэдры триангуляции хорошим образом.

Сначала я скажу, как нормальная поверхность может пересекать тетраэдры, а потом перечислю запрещенные пересечения.

Поверхность может пересекать тетраэдр по треугольнику (или нескольким параллельным копиям треугольника) и по прямоугольнику (рис. 1). Для каждого тетраэдра есть 4 типа треугольников (каждый тип соответствует одной из граней) и 3 типа прямоугольников (каждый тип соответствует одной из пар противоположных ребер).

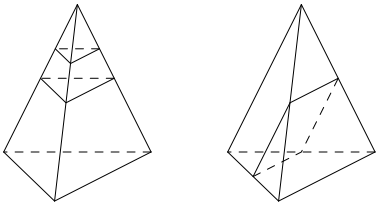


Рис. 1. Разрешенные пересечения

Поверхности запрещается входить в тетраэдр в виде трубочки и запрещены «возвраты» (рис. 2).

Требуется, чтобы каждая связная компонента поверхности пересекала ребро не более одного раза.

Более того, пересечение поверхности с каждым тетраэдром должно состоять из дисков, причем край каждого диска должен пересекать каждое ребро не более одного раза. Сама поверхность может при этом пересекать ребро много раз.

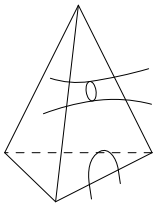


Рис. 2. Запрещенные пересечения

Теперь я могу объяснить, почему класс нормальных поверхностей информативен, т. е. содержит все интересные поверхности. Сначала нужно выяснить, какие поверхности неинтересные. К неинтересным поверхностям относятся все поверхности, расположенные в $\mathbb{R}^3$, потому что такие поверхности есть во всех трехмерных многообразиях.

Нужно запретить все поверхности «с трубочками» (рис. 3). Трубочка характеризуется тем, что есть диск, который пересекает поверхность в точности по своему краю, и это пересечение не ограничивает диска на поверхности. Такие диски называют *сжимающими*. Это название связано с тем, что сжимающий диск можно сильно сжать и тем самым упростить поверхность (перерезать ее по диску). Несжимаемая поверхность в M^3 — это поверхность, для которой нет сжимающих дисков.

Многообразие M^3 называют *неприводимым*, если любая вложенная сфера $S^2 \subset M^3$ ограничивает шар. Любое трехмерное многообразие является связной суммой неприводимых, кроме многообразия $S^2 \times S^1$, которое неприводимо, но не является нетривиальной связной суммой. Поэтому если мы поймем, как устроены неприводимые трехмерные многообразия, то мы поймем, как устроены все трехмерные многообразия.

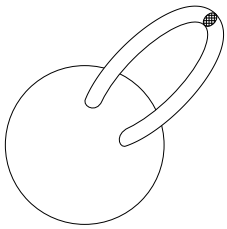


Рис. 3. Поверхность с трубочкой

Теорема 1. *Любая замкнутая несжимаемая поверхность в неприводимом многообразии M^3 изотопна нормальной поверхности.*

Доказательство. Возьмем произвольную замкнутую несжимаемую поверхность и приведем ее в общее положение относительно тетраэдров триангуляции. Если есть запрещенная ситуация (трубка, пересекающая грань тетраэдра по окружности), то тогда есть и сжимающий диск D. Поверхность несжимаемая, поэтому край ∂D этого диска ограничивает диск D' в поверхности, см. рис. 4. Так как многообразие неприводимо, то сфера $D \cup D'$ ограничивает шар, который можно втащить в тетраэдр или вытащить из него. После этого запрещенное пересечение будет уничтожено.

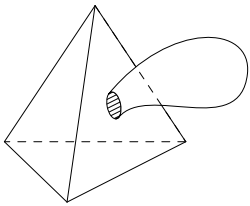


Рис. 4. Устранение пересечения с гранью

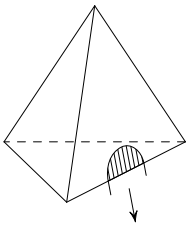


Рис. 5. Устранение пересечения с ребром

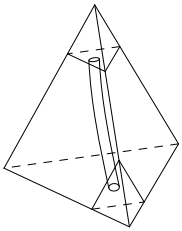


Рис. 6. Устранение «трубочки»

Если есть запрещенная ситуация, изображенная на рис. 5, то тогда пересечение такой компоненты сферы с тетраэдром устраняется очевидным образом, с помощью изотопии поверхности.

Далее, если внутри тетраэдра есть трубочка, соединяющая два треугольных сечения (рис. 6), то эту трубочку можно уничтожить, снова воспользовавшись несжимаемостью поверхности.

Все остальные случаи разбираются аналогично. □

Таким образом, первое условие выполнено: класс $\mathcal{N}$ информативен.

Теперь нужно проверить, что второе условие тоже выполняется. Для этого нужно явно описать класс всех нормальных поверхностей. Каждую нормальную поверхность можно задать целочисленным вектором. Это делается так. В каждом тетраэдре есть 7 разрешенных типов пересечений. Обозначим все возможные типы пересечений (со всеми тетраэдрами) $E_1, \ldots, E_n$; здесь $n = 7t$, где t — количество всех тетраэдров в триангуляции. Тогда поверхности F можно сопоставить целочисленный вектор $(x_1, x_2, \ldots, x_n)$, где x_i — количество треугольников или прямоугольников типа E_i в пересечении поверхности с тетраэдрами. Легко видеть, что если все числа известны, то по ним поверхность восстанавливается однозначно: участки поверхности можно состыковать лишь одним способом. Но при этом не любому набору чисел соответствует поверхность. Например, набору $(1, 0, \ldots, 0)$ не может соответствовать (замкнутая) поверхность.

Выясним, какие векторы реализуются поверхностями. Рассмотрим тетраэдр, выделим одну его грань, и у этой грани выделим один из трех углов. Проведем отрезок, соединяющий стороны этого угла. Этот отрезок может принадлежать одному треугольному сечению типа E_i и одному прямоугольному сечению типа E_j. Выделенная грань принадлежит еще ровно одному другому тетраэдру. В нем рассматриваемый отрезок может принадлежать сечениям типа E_k и типа E_m. В результате получаем соотношение $x_i + x_j = x_k + x_m$. Кроме того, если поверхность вложенная, то ни в одном тетраэдре не может быть прямоугольных сечений разного типа (такие сечения всегда пересекаются); решения полученной системы уравнений будем называть допустимыми, если у него нет одновременно двух ненулевых x_i, относящихся к разным типам прямоугольных сечений одного и того же тетраэдра.

Теорема 2. *Множество всех нормальных поверхностей находится во взаимно однозначном соответствии с множеством допустимых целых неотрицательных решений описанной выше системы уравнений.*

Доказательство этой теоремы очень простое; я не буду на нем останавливаться.

Количество уравнений в системе равно $6t$, поэтому система недоопределенная. У нее должно быть много решений. Важное наблюдение Хакена заключается в том, что множество решений этой системы имеет конечный базис фундаментальных решений.

Решение $\bar{x}$ называют фундаментальным (неразложимым), если из равенства $\bar{x} = \bar{y} + \bar{z}$, где y, z — неотрицательные целые решения, следует, что $\bar{y} = 0$ или $\bar{z} = 0$.

Теорема 3 (Хакен). *Множество фундаментальных решений конечно.*

В линейной алгебре хорошо известна аналогичная теорема, но там разрешены любые коэффициенты: целые и не целые, положительные и отрицательные. Здесь же коэффициенты должны быть целыми неотрицательными.

Доказательство теоремы Хакена достаточно просто, но я вместо доказательства в общем случае рассмотрю простой пример.

Прежде всего я отмечу следующее. Мы ищем неотрицательные решения; тем самым в систему мы включаем неравенства. Можно считать, что вся система состоит только из однородных неравенств. Действительно, уравнение $x = 0$ эквивалентно системе двух неравенств $x \geqslant 0$ и $x \leqslant 0$.

Рассмотрим систему неравенств $-x + 4y \geqslant 0$ и $2x - y \geqslant 0$ (рис. 7). Сложим координаты фундаментальных решений, принадлежащих прямым: $(5, 3) = (4, 1) + (1, 2)$. Ясно, что все фундаментальные решения содержатся в области $x \leqslant 5$, $y \leqslant 3$. Действительно, если решение лежит вне этой области, то из этого решения можно вычесть одно из двух фундаментальных решений (4, 1) и (1, 2). В общем случае рассуждения те же самые.

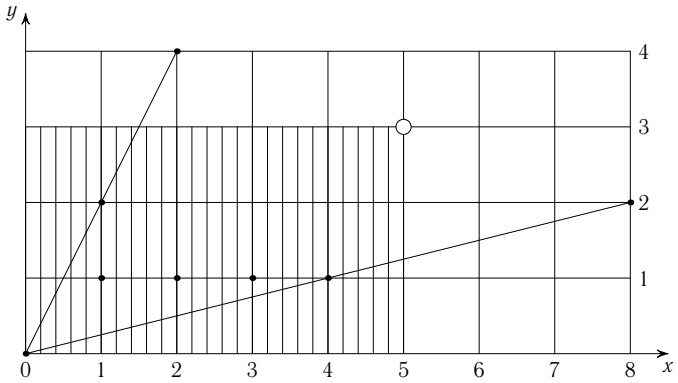


Рис. 7. Система неравенств

Теперь я могу выполнить одну из своих целей — рассказать алгоритм распознавания тривиального узла. Тривиальный узел характеризуется тем, что он ограничивает вложенный диск. Это вопрос о поверхности. (Существует ли поверхность данного типа?) Правда, здесь поверхность незамкнутая. Но все, что я говорил, проходит и для незамкнутых поверхностей, правда число уравнений уменьшается: отрезки в треугольниках на крае многообразия никаких уравнений не дают. Нужно только не включать в

систему уравнений те треугольники и прямоугольники, которые примыкают к краю поверхности.

Для точности узел нужно заменить на полноторие — трубчатую окрестность узла. Узел тривиален тогда и только тогда, когда на какую-нибудь параллель этого полнотория можно снаружи натянуть диск.

Несложно показать, что если существует хотя бы один диск, натянутый на параллель, то такой диск найдется и среди фундаментальных поверхностей. Если это известно, то алгоритм распознавания узла, по крайней мере теоретически, чрезвычайно прост. Удаляем трубчатую окрестность узла и триангулируем полученное многообразие. Составляем систему уравнений (неравенств) и находим фундаментальные решения. Затем для каждого фундаментального решения проверяем, дает ли оно диск, край которого является параллелью. Это легко сделать, вычислив эйлерову характеристику поверхности и проверив, что ее край состоит из окружности, пересекающей медиан один раз.

Теперь я хочу рассказать о классификации достаточно больших трехмерных многообразий. Неприводимое многообразие M^3 называют *достаточно большим*, если оно содержит несжимаемую поверхность, отличную от S^2, $\mathbb{R}P^2$ и D^2. Например, любое неприводимое многообразие с краем либо достаточно большое, либо гомеоморфно полному кренделю некоторого рода. Многообразия, у которых одномерная группа гомологий бесконечна, тоже достаточно большие.

Теорема 4 (Хакен, Вальдхаузен, Иогансен, Хемион). *Существует алгоритм распознавания достаточно больших трехмерных многообразий.* (*Для любых двух достаточно больших многообразий этот алгоритм позволяет выяснить, гомеоморфны они или нет.*)

Из этой теоремы легко выводится существование алгоритмической классификации достаточно больших трехмерных многообразий. А именно, сначала строится алгоритм, перечисляющий все трехмерные многообразия. Для этого нужно среди 3-мерных симплициальных комплексов выбрать те, которые являются многообразиями. В полученном списке будут дубликаты. От дубликатов можно избавиться, применив алгоритм распознавания.

История теоремы классификации достаточно больших трехмерных многообразий такова. Впервые ее доказал Хакен, после того как он построил теорию нормальных поверхностей в 1962 г. Но вскоре в его доказательстве обнаружили серьезный пробел. Долгое время различные математики (Вальдсхаузен, Иоганнсон, Джейко, Шэлен и др.) пытались его ликвидировать. При этом была проделана большая и важная работа, построена теория характеристических подмногообразий.

В конце концов было выделено ключевое препятствие: для некоторого очень специального класса многообразий метод Хакена не работает. Это

препятствие — так называемые многообразия Столлингса. Позже я объясню, почему для них метод Хакена не работает.

Препятствие, связанное с многообразиями Столлингса, преодолел Хемион в 1976 г. Он сумел решить задачу распознавания для многообразий Столлингса независимым методом, который не имел никакого отношения к теории Хакена.

После этого было объявлено, что теорема классификации достаточно больших многообразий доказана. На эту тему появилось несколько обзорных статей, в 1991 г. вышла книга Хемиона. Теорема важная, на нее ссылаются во многих работах.

Все публикации по этой теме были построены по одной и той же схеме. Сначала говорилось о доказательстве Хакена и возникающем там препятствии, затем о том, как Хемион это препятствие преодолел, после чего делалось заключение, что теорема доказана. Однако, никакого претендующего на полноту текста так и не появилось. Возникает естественный вопрос: «Почему нет других препятствий?» Я решил в этом разобраться и написать полное доказательство. Оказалось, что действительно есть еще одно препятствие, которое я назвал квазимногообразиями Столлингса и которое не может быть преодолено с помощью метода Хемиона. Других препятствий нет. Чтобы преодолеть второе препятствие, нужен очень мощный аппарат теории гомеоморфизмов поверхностей, разработанный Тёрстоном гораздо позже 1976 г., когда было объявлено о доказательстве теоремы. Работы Тёрстона появились в 80-е годы, а нам нужна алгоритмическая версия этого аппарата, так называемые train tracks (железнодорожные пути), которую предложили Бествина и Хэндел еще позже — в 1995 г.

Таким образом, до 1998 г., когда была опубликована моя статья в «Успехах математических наук», теорема оставалась не доказанной. Полное доказательство заняло 20 лет.

Из теоремы о классификации достаточно больших многообразий сразу вытекает существование алгоритма классификации узлов. Точнее говоря, не из самой теоремы, а из метода ее доказательства: нужно повторить доказательство теоремы Хакена, следя за тем, как себя ведет меридиан трубчатой окрестности узла.

Теорема классификации доказывается с помощью конструкции, называемой *иерархией*, или *скелетом*. Возьмем достаточно большое трехмерное многообразие. В нем есть несжимаемая поверхность. Разрежем многообразие по этой поверхности. В результате получим либо одно многообразие, либо два. Эти многообразия будем называть ячейками. Ячейки имеют край, поэтому они достаточно большие. Снова сделаем разрезы по несжимаемым поверхностям и т. д. Останавливаемся тогда, когда полученные ячейки — шары. При этом нужно следить за тем, чтобы поверхности были в общем положении. Например, не должно быть троек поверхностей,

пересекающихся в одной точке. Технически удобнее не разрезать многообразие, а возводить перегородки. Тогда в многообразии получаем некоторый двумерный полиэдр P^2, который я и называю скелетом. Если проводить разрезы, то результат называют иерархией. Но при этом очень сложно следить за тем, что происходит. Рассмотрение скелета (объекта), а не иерархии (процесса), существенно упрощает доказательство теоремы классификации.

Доказательство того, что мы рано или поздно остановимся, основано на понятии *сложности* трехмерного многообразия. При разрезании сложность всегда уменьшается (если только она отлична от нуля).

Возьмем другое многообразие и построим для него скелет. Первое важное наблюдение таково: если скелеты гомеоморфны, то многообразия тоже гомеоморфны. Действительно, гомеоморфизм границ двух шаров с помощью конической конструкции можно продолжить до гомеоморфизма шаров.

Второе наблюдение таково. Предположим, что нам удалось ввести на допустимые вставки такие жесткие ограничения, что на каждом шаге можно делать только конечное число вставок (с точностью до гомеоморфизма многообразия на себя). Тогда для каждого многообразия получается конечный набор скелетов, причем многообразия гомеоморфны тогда и только тогда, когда наборы скелетов попарно гомеоморфны. Действительно, с одной стороны, если какие-то два скелета гомеоморфны, то, как уже было показано, гомеоморфны и сами многообразия. С другой стороны, если многообразия гомеоморфны, то наборы скелетов для них тоже окажутся гомеоморфными, потому что конструкция определена с точностью до гомеоморфизма.

Все сводится к тому, чтобы добиться конечности допустимых вставок. Будем вставлять поверхности минимальной сложности. *Сложностью* поверхности F назовем число $c(F) = -\chi(F) + N$, где N — число точек пересечения с сингулярностями полиэдра, полученного на предыдущем шаге. Теорема о конечности числа поверхностей минимальной сложности — это вариант теории нормальных поверхностей, учитывающий сложность поверхности. Количество фундаментальных поверхностей с учетом сложности тоже остается конечным.

Хакен отметил следующее. Предположим, что на каждом шаге все вставляемые фундаментальные поверхности имеют положительную сложность. Тогда поверхности минимальной сложности — часть фундаментальных поверхностей; в частности, поверхностей минимальной сложности конечное число. Утверждение очевидное, так как сложность аддитивна по отношению к суммированию поверхностей. Поэтому, пусть некоторая поверхность F является суммой двух поверхностей F_1 и F_2 положительной сложности. Тогда $c(F) > c(F_2)$ и $c(F) > c(F_1)$. Поэто-

му любая поверхность минимальной сложности является фундаментальной.

Это — легкий случай. Хакен дальше него не продвинулся. Здесь возникает вопрос: как быть в том случае, когда есть фундаментальные поверхности нулевой сложности? Хакен посчитал, что с этим случаем легко можно справиться, но оказалось, что это не так.

Какие бывают поверхности нулевой сложности? Другими словами, когда выполняется равенство $-\chi(F)+N=0$? Эйлерова характеристика может быть положительной лишь в случае сферы, проективной плоскости $\mathbb{R}P^2$ и диска. Но из-за неприводимости сферы исключены. Поэтому проективные плоскости тоже исключены, потому что край трубчатой окрестности проективной плоскости — сфера. Диски мы исключаем, пользуясь несжимаемостью: поверхности вставляем так, чтобы они были несжимаемыми; тогда дисков не будет.

Таким образом, поверхность нулевой сложности не содержит сингулярных точек и имеет нулевую эйлерову характеристику. Нулевую эйлерову характеристику имеют кольцо, лист Мёбиуса, тор и бутылка Клейна. Торы и бутылки Клейна нам не мешают, так как любая ячейка содержит только конечное число таких поверхностей (с точностью до гомеоморфизмов ячейки, неподвижных на ее крае).

Случай колец и листов Мёбиуса гораздо сложнее, поскольку число таких поверхностей в ячейке может быть бесконечным, даже с точностью до неподвижных на крае гомеоморфизмов.

Например, многократно скручивая одно кольцо вдоль другого (см. рис. 8), можно получить бесконечное число неэквивалентных колец, поскольку скручивания ячейки вдоль кольца сдвигают край.

Правильный путь состоит в разбиении всех колец на два типа: продольные и поперечные (случай листов Мёбиуса рассматривается аналогично, и его можно опустить). Здесь я имею в виду такую ситуацию. Все кольцо лежит внутри одной ячейки, и край кольца принадлежит краю ячейки. Кольцо A называют *продольным*, если любое другое кольцо A_1 можно сдвинуть так, чтобы для нового кольца A_1' выполнялось условие: пересечение $A\cap A_1'$ либо пусто, либо состоит из окружностей, параллельных средней линии кольца A (рис. 9). В противном случае кольцо называют *поперечным*.

Рис. 8

Окружность на кольце, ограничивающую диск, можно убрать из-за несжимаемости. Поэтому пересечение любого кольца с поперечным кольцом можно привести к виду, изображенному на рис. 10.

Оказалось, что продольные кольца намного приятнее, чем поперечные. Наличие колец нам мешало тем, что скручивание вдоль кольца могло приводить к гомеоморфизмам ячейки, которые не продолжаются до гомеоморфизма многообразия. Но при скручивании вдоль любого другого кольца продольное кольцо остается инвариантным, потому что пересечение этих колец состоит из окружностей, параллельных средней линии кольца. Эти окружности инвариантны, поэтому продольное кольцо нечувствительно к скручиваниям. Отсюда следует, что продольных колец конечное число.

Продольных колец конечное число, потому что добавление к ним других колец равносильно скручиваниям, а продольные кольца к скручиваниям нечувствительны.

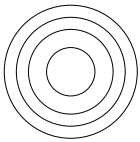


Рис. 9. Продольное кольцо

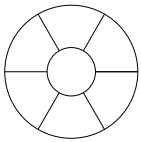


Рис. 10. Поперечное кольцо

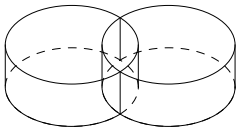


Рис. 11. Структура прямого произведения

Теперь можно сформулировать окончательный вариант способа вставления поверхностей. Если колец вообще нет, то работает предыдущий метод. Если есть хотя бы одно продольное кольцо, то мы его вставляем; продольных колец конечное число. Остановимся мы тогда, когда останутся только шары и ячейки, в которых нет продольных колец, но есть поперечные кольца. Оказывается, что такие ячейки легко описать: они имеют вид $F \times I$ или $F\tilde{\times}I$ (нетривиальное расслоение со слоем отрезок).

Доказывается это весьма просто. Нарисуем поперечное кольцо в виде цилиндра. Из того, что оно поперечное, следует, что существует другое кольцо, которое пересекает его по двум вертикальным отрезкам (рис. 11). Тогда структура прямого произведения есть уже на окрестности объединения этих двух колец. Конструкция расширяется и каждый раз сохраняется структура прямого произведения, точнее говоря, расслоения со слоем отрезок (при распространении отрезки могут перевернуться, т. е. может получиться косое произведение).

На шары я не буду обращать внимание. Если есть поверхность $F \times I$, то к ней тоже что-то приклеивается и т. д. В результате получается многообразие Столлингса. Для него алгоритм Хакена не работает, потому что после первого разреза получаем многообразие $F \times I$, в котором очень много разных колец. Кольца соответствуют кривым на поверхности, а число кривых на поверхности бесконечно. Поэтому нет шансов сделать процедуру выбора какого-нибудь кольца конечной.

Для многообразий, содержащих многообразия Столлингса, задачу классификации нужно решать отдельно. Это сделал Хемион. После этого было объявлено, что задача классификации трехмерных многообразий решена. Но есть еще и квазимногообразия Столлингса, которые склеиваются из $F \tilde{\times} I$ и $F\tilde{I}$. Для них процедура Хакена не работает по тем же самым причинам.

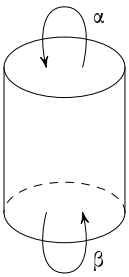


Рис. 12. Две инволюции

Возьмем поверхность $F \times I$ и на каждом ее крае рассмотрим некоторую обращающую ориентацию инволюцию (гомеоморфизм периода 2) без неподвижных точек. Произведем по этим инволюциям склейку (рис. 12). Полученное в результате многообразие — это и есть квазимногообразие Столлингса. Для таких многообразий метод не работает по тем же причинам, что и раньше. Для них надо решать проблему отдельно. Несложная редукция сводит проблему распознавания квазимногообразий Столлингса к следующей проблеме о гомеоморфизмах поверхностей. Пусть заданы два гомеоморфизма $f, g\colon F \to F$ поверхности на себя. Требуется выяснить, существует ли такое число n, что $f^n = g$ (равенство с точностью до изотопии). Гомеоморфизмы поверхностей, рассматриваемые с точностью до изотопии, описываются в терминах гомоморфизмов фундаментальной группы. Поэтому задача чисто алгебраическая. Она трудная лишь потому, что число n не ограничено. Если же получить оценку для числа n, то задача сразу становится легкой.

Оценку для числа n можно получить с помощью теории Тёрстона так называемых *растягивающих факторов*. Наиболее проста эта теория для тора. Гомеоморфизмы тора задаются матрицами порядка 2 с определителем, равным 1. Мнимые собственные значения бывают очень редко; этот случай неинтересный. Если же собственные значения действительные, то они имеют вид λ и λ^{-1}. Таким образом, в одном направлении происходит растяжение, а в другом сжатие. Для тора это было известно давно, а Тёрстон доказал, что так устроены не только гомеоморфизмы тора, но и гомеоморфизмы всех поверхностей. Для каждого гомеоморфизма поверхности определен растягивающий фактор $\lambda > 1$. Тогда оценить число n можно так: оно не превосходит любого такого N, что $\lambda(f)^N \geqslant \lambda(g)$. Это завершает доказательство теоремы об алгоритмическом распознавании достаточно больших многообразий и теоремы об их алгоритмической классификации.

Тогда неравенство $\lambda(f)^n > \lambda(g)$ позволяет получить оценку для n.

Это преодолевает второе неравенство. Я доказал, что больше никаких других препятствий нет.

Оглавление

Студенческие чтения НМУ
Выпуск 2

Научный редактор *В. Прасолов*
Редактор *Ю. Торхов*
Обложка *М. Панов*

Издательство Московского Центра
непрерывного математического образования
Лицензия ИД №01335 от 24.03.2000 г.

Подписано в печать 30.11.2001 г. Формат $60 \times 88^1/_{16}$. Бумага офсетная №1.
Печать офсетная. Печ. л. 9,0. Тираж 1000 экз. Заказ №

МЦНМО
121002, Москва, Большой Власьевский пер., 11

Отпечатано с готовых диапозитивов в АООТ «Политех-4»
129110, Москва, ул. Большая Переславская, 46.

Книги издательства МЦНМО можно приобрести в магазине «Математическая книга»,
Большой Власьевский пер., д. 11. Тел. (095) 241–72–85. E-mail: biblio@mccme.ru