Роджер Пенроуз

МΩДА

и новая

ВЕРА

физика

ФАНТАЗИЯ

Вселенной

Роджер
Пенроуз

МОДА

и новая

ВЕРА

физика

ФАНТАЗИЯ

Вселенной

ПИТЕР®

Санкт-Петербург • Москва • Екатеринбург • Воронеж
Нижний Новгород • Ростов-на-Дону
Самара Минск
2020

Роджер Пенроуз

Мода, вера, фантазия и новая физика Вселенной

Серия «New Science»

Перевели с английского *А. Пасечник, О. Сивченко*

Заведующая редакцией	*Ю. Сергиенко*
Ведущий редактор	*Н. Римицан*
Научный редактор	*А. Пасечник*
Литературный редактор	*О. Лапина*
Художественный редактор	*В. Мостипан*
Корректоры	*Н. Викторова, Н. Сидорова, Г. Шкатова*
Верстка	*Л. Егорова*

ББК 22.6
УДК 524

Пенроуз Роджер

П25 Мода, вера, фантазия и новая физика Вселенной. — СПб.: Питер, 2020. — 512 с.: ил. — (Серия «New Science»).

ISBN 978-5-4461-1598-3

Можно ли говорить о моде, вере или фантазии в фундаментальной науке?

Вселенной не интересна человеческая мода. Науку невозможно трактовать как веру, ведь научные постулаты постоянно подвергаются строгой экспериментальной проверке и отбрасываются, как только догма начинает конфликтовать с объективной реальностью. А фантазия вообще пренебрегает и фактами, и логикой. Тем не менее великий Роджер Пенроуз не желает полностью отвергать эти феномены, ведь научная мода может оказаться двигателем прогресса, вера появляется, когда теория подтверждается реальными экспериментами, а без полета фантазии не постичь все странности нашей Вселенной.

В главе «Мода» вы узнаете о теории струн — самой модной теории последних десятилетий. «Вера» посвящена догматам, на которых стоит квантовая механика. А «Фантазия» касается ни много ни мало — теорий происхождения известной нам Вселенной.

16+ (В соответствии с Федеральным законом от 29 декабря 2010 г. № 436-ФЗ.)

ISBN 978-0691119793 англ.
ISBN 978-5-4461-1598-3

© 2016 by Roger Penrose
© Перевод на русский язык ООО Издательство «Питер», 2020
© Издание на русском языке, оформление ООО Издательство «Питер», 2020
© Серия «New Science», 2020

Права на издание получены по соглашению с Synopsis. Все права защищены. Никакая часть данной книги не может быть воспроизведена в какой бы то ни было форме без письменного разрешения владельцев авторских прав.

Информация, содержащаяся в данной книге, получена из источников, рассматриваемых издательством как надежные. Тем не менее, имея в виду возможные человеческие или технические ошибки, издательство не может гарантировать абсолютную точность и полноту приводимых сведений и не несет ответственности за возможные ошибки, связанные с использованием книги. Издательство не несет ответственности за доступность материалов, ссылки на которые вы можете найти в этой книге. На момент подготовки книги к изданию все ссылки на интернет-ресурсы были действующими.

Изготовлено в России. Изготовитель: ООО «Прогресс книга».
Место нахождения и фактический адрес: 194044, Россия, г. Санкт-Петербург,
Б. Сампсониевский пр., д. 29А, пом. 52. Тел.: +78127037373.

Дата изготовления: 11.2019. Наименование: книжная продукция. Срок годности: не ограничен.

Налоговая льгота — общероссийский классификатор продукции ОК 034-2014, 58.11.12.000 —
Книги печатные профессиональные, технические и научные.

Импортер в Беларусь: ООО «ПИТЕР М», 220020, РБ, г. Минск, ул. Тимирязева, д. 121/3, к. 214, тел./факс: 208 80 01.

Подписано в печать 07.11.19. Формат 70x100/16. Бумага офсетная. Усл. п. л. 41,280. Тираж 3000. Заказ 0000.

Оглавление

Иллюстрации. Благодарности 5

Предисловие. Уместны ли мода, вера или фантазия в фундаментальной науке? 6

Глава 1. Мода 13

1.1. Математическая красота как движущая сила 13
1.2. Некоторые модные идеи прошлого 23
1.3. Физика частиц как основание для теории струн 30
1.4. Принцип суперпозиции в КТП 33
1.5. Сила фейнмановских диаграмм 38
1.6. Исходные ключевые идеи теории струн 46
1.7. Время в общей теории относительности Эйнштейна 57
1.8. Вейль и его калибровочная теория электромагнетизма 68
1.9. Функциональная свобода в модели Калуцы — Клейна и струнной модели 75
1.10. Квантовые препятствия для функциональной свободы 86
1.11. Классическая нестабильность теории струн для высших измерений 95
1.12. Модность теории струн 101
1.13. М-теория 109
1.14. Суперсимметрия 114
1.15. AdS/CFT-соответствие 124
1.16. Миры на бранах и ландшафт 137

Глава 2. Вера 141

2.1. Квантовое откровение 141
2.2. Уравнение Макса Планка E = hv 146
2.3. Корпускулярно-волновой парадокс 154
2.4. Квантовые и классические уровни: C, U и R 159
2.5. Волновая функция точечной частицы 166
2.6. Волновая функция фотона 174
2.7. Квантовая линейность 180
2.8. Квантовое измерение 187
2.9. Геометрия квантового спина 197
2.10. Квантовая запутанность и ЭПР-эффекты 206
2.11. Квантовая функциональная свобода 213
2.12. Квантовая реальность 224
2.13. Объективная редукция квантового состояния: предел квантовой веры? 232

Глава 3. Фантазия 244
3.1. Большой взрыв и вселенные Фридмана 244
3.2. Черные дыры и локальные нерегулярности 259
3.3. Второй закон термодинамики 272
3.4. Парадокс Большого взрыва 281
3.5. Горизонты, сопутствующие объемы и конформные диаграммы 290
3.6. Феноменальная точность Большого взрыва 302
3.7. Космологическая энтропия? 309
3.8. Энергия вакуума 318
3.9. Инфляционная космология 328
3.10. Антропный принцип 345
3.11. Некоторые более фантастические космологии 359

Глава 4. Новая физика Вселенной 371
4.1. Теория твисторов: альтернатива теории струн? 371
4.2. Откуда берутся квантовые основания? 392
4.3. Конформная невообразимая космология? 411
4.4. Личное послесловие 432

Приложение А. Математическое приложение 438
A.1. Итерационные степени 438
A.2. Функциональная свобода полей 442
A.3. Векторные пространства 449
A.4. Векторные базисы, координаты и дуальность 454
A.5. Математика многообразий 458
A.6. Многообразия в физике 468
A.7. Расслоения 474
A.8. Функциональная свобода на уровне расслоений 482
A.9. Комплексные числа 488
A.10. Комплексная геометрия 492
A.11. Гармонический анализ 502

Список литературы ..... https://clck.ru/Jojns

Иллюстрации. Благодарности

Автор искренне благодарит правообладателей следующих иллюстраций:

Рис. 1.38: М. К. Эшер, «Предел круга I», © 2016, Компания М. К. Эшера — Нидерланды. Все права защищены, www.mcescher.com.

Рис. 3.1: (а) М. К. Эшер, «Фотография сферы», (б) «Симметрия. Рисунок E45» и (в) «Предел круга IV», © 2016, Компания М. К. Эшера — Нидерланды. Все права защищены, www.mcescher.com.

Рис. 3.38 (а) и (б): Лекция «Космическая инфляция» Андреаса Альбрехта из сборника «Формирование структур во Вселенной» под ред. Р. Криттендена и Н. Турока (Cosmic Inflation by Andreas Albrecht, in Structure Formation in the Universe, ed. R. Crittenden and N. Turok). Публикуется с разрешения «Springer Science and Business Media».

Рис. 3.38 (в): «Инфляция для астрономов» Дж. В. Нарликара и Т. Падманабхана в обработке Итана Сигела из работы «Почему мы считаем, что существует не одна Вселенная, а целая Мультивселенная» (Inflation for Astronomers by J.V. Narlikar and T. Padmanabhan as modified by Ethan Siegel in “Why we think there’s a Multiverse, not just our Universe”, https://medium.com/starts-with-a-bang/why-we-think-theres-a-multiverse-not-just-our-universe-23d5ecd33707#.3iib9ejum). Воспроизведено с разрешения Annual Review of Astronomy and Astrophysics, 1 September 1991, Volume 29_c by Annual Reviews, http://www.annualreviews.org.

Рис. 3.38 (г): Статья «Вечная инфляция, прошлое и будущее» Энтони Агирре из книги «По ту сторону Большого взрыва: конкурирующие модели вечной Вселенной» под ред. Руди Ваааса (Eternal Inflation, Past and Future by Anthony Aguirre, in Beyond the Big Bang: Competing Scenarios for an Eternal Universe (The Frontiers Collection) (ed. RudyVaas)). Публикуется с разрешения «Springer Science and Business Media».

Рис. 3.43: Публикуется с разрешения Европейского Космического агентства и Коллаборации проекта Planck.

ПРЕДИСЛОВИЕ

Уместны ли мода, вера или фантазия в фундаментальной науке?

Эта книга основана на материале трех лекций, прочитанных мною в Принстонском университете в октябре 2003 года по приглашению издательства *Princeton University Press*. Я предложил издательству озаглавить лекции «Мода, вера, фантазия и новая физика Вселенной» — это и стало названием книги, хотя, возможно, оно оказалось несколько опрометчивым. Однако оно точно отражает, насколько непросто мне было выразить некоторые тенденции, над которыми я тогда размышлял. Они связаны с физическими законами, управляющими той Вселенной, в которой мы живем. С тех пор прошло уже более десяти лет, но многое из того, что я тогда говорил, не потеряло своей актуальности. Могу добавить, что, читая эти лекции, я испытывал некоторую тревогу, поскольку пытался озвучить кое-какие точки зрения, которые, как мне казалось, будут весьма прохладно восприняты многими присутствовавшими на лекциях заслуженными экспертами.

Каждое из слов, вошедших в заглавие, — «мода», «вера» и «фантазия» — описывает феномен, который на первый взгляд очень плохо вписывается в круг процедур, обычно применяемых при поиске глубочайших принципов, лежащих в фундаменте нашей Вселенной. Действительно, в идеале было бы весьма разумно утверждать, что такие факторы, как мода, вера и фантазия, абсолютно не должны волновать людей, всерьез занимающихся поиском первооснов мироздания. В конце концов, Природа как таковая определенно не выказывает заметного интереса к мимолетным переменам человеческой моды. Точно так же науку не следует трактовать как веру, поскольку научные догмы постоянно пересматриваются и подвергаются строгой экспериментальной проверке. Они отметаются сразу же, как только обнаруживается очевидный конфликт между такой догмой и объективной реальностью. А фантазия определенно является уделом тех творческих и развлекательных жанров, где не считается важным принимать во внимание ни наблюдаемые факты, ни строгую логику, ни даже

обычный здравый смысл. Действительно, если будет обнаружено, что предлагаемая научная теория подчиняется существующей моде либо слепо следует какой-либо вере, не имеющей под собой экспериментального подтверждения, либо связана с романтическими фантазиями, то наш долг — погрозить пальцем тем, кто, пусть даже неосознанно, следует подобным веяниям.

Тем не менее я не желаю полностью отвергать эти феномены. Ведь можно утверждать, что у каждого из них есть очевидная положительная сторона. В конце концов, модная теория вряд ли приобрела бы свой статус по чисто социальным причинам. Она должна обладать множеством ценных качеств, чтобы столь многие исследователи продолжали работать в столь модной научной сфере. При этом вряд ли ученые разрабатывали бы крайне сложную научную дисциплину из одного лишь желания быть как все — зачастую сама эта сложность проистекает из конкурентной природы наиболее модных изысканий.

Далее здесь следует отметить, что исследования в теоретической физике, будучи модными, могут отнюдь не давать правдоподобного описания мира — действительно, зачастую они возмутительно противоречат реальным наблюдениям. При том что ученые, работающие в этих областях, были бы изрядно польщены, если бы наблюдаемые факты лучше согласовывались с получающейся у них картиной мира, обычно они относительно спокойно воспринимают те факты, которые оказываются упрямее, чем хотелось бы. И это не лишено смысла, так как в значительной степени эти ученые заняты именно *исследовательской* работой; считается, что такая работа помогает приобрести некоторый опыт, и этот опыт в конечном итоге пригодится при разработке более качественных теорий, которые, в конечном итоге, будут точнее соответствовать реальному устройству известной нам Вселенной.

Когда дело доходит до исключительной веры в определенные научные догмы, которую зачастую демонстрируют ученые, то она также, вероятно, имеет право на существование, даже если эта вера подразумевает применимость такой догмы в обстоятельствах, выходящих далеко за рамки исходных ситуаций, где на базе убедительных результатов наблюдений были изначально заложены основы этой догмы. Выдающиеся физические теории прошлого по-прежнему заслуживают доверия и работают с исключительной точностью, пусть даже в некоторых областях им на смену пришли более совершенные теории, обладающие либо более высокой точностью, либо более широкой областью применения. Именно так случилось, когда величественная ньютоновская теория тяготения уступила место эйнштейновской и когда красивая максвелловская электромагнитная теория света была заменена квантовой теорией, позволяющей понять некоторые свойства света (фотонов). В каждом случае каждая отдельная теория сохраняет

достоверность при условии, что исследователи не забывают о присущих ей недостатках.

Но что насчет фантазии? Определенно, в науке мы стремимся к чему-то прямо противоположному. Однако мы увидим, что нашей реальной Вселенной присущи некоторые столь странные аспекты (хотя эта странность не всегда полностью осознается), что если не отдаться на волю самым, на первый взгляд, безудержным фантазиям, то мы никогда не сможем постичь фундаментальную истину, которая при этом покажется нам исключительно фантастичной.

В первых трех главах я проиллюстрирую три феномена, вынесенных в заглавие, на примере трех хорошо известных теорий либо их совокупности. Я не стану вдаваться во второстепенные детали, а сразу начну с «крупной рыбы», обитающей в океанах современной теоретической физики. В главе 1 я решил рассказать о теории струн, по-прежнему очень модной (а также о теории суперструн либо о ее обобщениях, например об М-теории или о наиболее модном в настоящее время аспекте этого направления исследований — о научной концепции под названием *AdS/CFT-соответствие*). Вера, которой я займусь в главе 2, — еще более «крупная рыба», и вот почему: методы, связанные с квантовой механикой, неукоснительно подчиняются догме, независимо от того, насколько крупны и массивны объекты, которыми они оперируют. А тема главы 3 в некоторых отношениях — самая огромная «рыбина», причем эта тема касается ни много ни мало происхождения известной нам Вселенной. Здесь нам придется рассмотреть некоторые гипотезы, кажущиеся сущей фантазией. Они были сформулированы ради объяснения некоторых подлинно неудобных деталей, обнаруженных в ходе тщательно выверенных наблюдений за самыми первыми этапами развития всей Вселенной.

Наконец, в главе 4 я изложу некоторые собственные воззрения, чтобы показать, какими альтернативными путями мы могли бы пойти. Здесь мы обнаружим, что возможность свернуть на любой из предложенных мною путей можно рассматривать лишь с некоторой долей иронии. Действительно, мой излюбленный подход к пониманию фундаментальной физики не лишен иронии и содержит дань моде — я кратко познакомлю с ним читателя в разделе 4.1. Это путь, проторенный *теорией твисторов*, в разработке которой я сам сыграл ключевую роль, причем на протяжении 40 лет эта теория не привлекала особого внимания физического сообщества. Но мы убедимся, что сегодня теория твисторов все больше входит в моду, причем именно в том же ключе, что и теория струн.

Кстати, всепоглощающая непоколебимая вера в квантовую механику, которую, по-видимому, разделяет большинство физиков, была подтверждена

примечательными экспериментами, например теми, что поставили Серж Арош и Дэвид Вайнленд, причем за эти эксперименты в 2012 году им была присуждена Нобелевская премия по физике. Более того, в 2013 году Нобелевская премия по физике досталась Питеру Хиггсу и Франсуа Энглеру за предсказание существования так называемого *бозона Хиггса*. Открытие этого бозона поразительным образом подтверждает не только некоторые идеи относительно происхождения масс частиц, высказанные этими учеными (а также некоторыми другими, и в частности Томом Кибблом, Джералдом Гуральником, Карлом Хейгеном и Робертом Браутом), но и многие фундаментальные аспекты квантовой теории поля как таковой. Однако, как я покажу в разделе 4.2, все эти крайне тонкие эксперименты, которые уже были поставлены, не всегда показывают достаточный сдвиг массы (об этом пойдет речь в разделе 2.13), который позволил бы нам всерьез усомниться в нашей вере в квантовую механику. В настоящее время разрабатываются и другие эксперименты, но они призваны достичь такого сдвига массы, который, на мой взгляд, помог бы нам разрешить некоторые глубокие конфликты между современными представлениями о квантовой механике и другими признанными физическими принципами, например эйнштейновской теорией относительности. Раздел 4.2 посвящен серьезному конфликту между современной квантовой механикой и фундаментальным эйнштейновским принципом эквивалентности — между гравитационными полями и ускорениями. Вы вправе спросить: «Почему эйнштейновский принцип эквивалентности заслуживает большего доверия, нежели основополагающие принципы квантовой механики, подтвержденные несоизмеримо большим количеством экспериментов?» Действительно, хороший вопрос. С полным правом можно утверждать, что принципы, сформулированные Эйнштейном, должны восприниматься как минимум с неменьшим доверием, чем принципы квантовой механики. Эта проблема вполне может быть разрешена экспериментально в недалеком будущем.

Что касается того, в какой степени современная космология погрязла в фантазиях, в разделе 4.3 я напомню (иронически), что еще в 2005 году сам же предложил одну картину мира — конформную циклическую космологию (КЦК), — которая в некоторых аспектах даже более фантастична, чем экстраординарные гипотезы, обсуждаемые в главе 3; на сегодняшний день некоторые из них обсуждаются в рамках практически любых дискуссий о самых ранних этапах развития Вселенной. Однако в настоящее время КЦК постепенно начинает подтверждаться на материале анализа наблюдений — по-видимому, существуют реальные физические факты, на которых она может быть основана. Определенно стоит надеяться, что вскоре у нас будут бесспорные эмпирические доказательства, которые позволят превратить вчерашнюю чистейшую

(на чей-то взгляд) фантазию в убедительную фактическую картину мира, отражающую устройство реальной Вселенной. При этом действительно можно отметить, что, в отличие от модной теории струн либо большинства теоретических построений, призванных подорвать наше общее доверие к принципам квантовой механики, вышеупомянутые фантастические предположения выдвигались для описания происхождения Вселенной и уже подвергаются тщательной наблюдательной проверке: так, обширная информация поступает от спутников COBE, WMAP, космической платформы «Планк». Также учитываются результаты наблюдений аппарата BICEP2, расположенного на Южном полюсе, опубликованные в марте 2014 года. На момент написания книги интерпретация данных последнего эксперимента вызывает большие проблемы, но рано или поздно загадка должна разрешиться. Возможно, вскоре у нас будут гораздо более четкие доказательства, которые позволят выбрать между конкурирующими фантастическими теориями либо остановиться на теории, которая пока не придумана.

Пытаясь обсудить все эти проблемы адекватным (но не слишком научным) языком, я столкнулся с одним фундаментальным препятствием. Это проблема математики и ее центральной роли в любой физической теории, всерьез претендующей на достаточно полное описание природы. Важнейшие аргументы, которые я привожу в этой книге, призваны показать, что мода, вера и фантазия действительно негативно влияют на прогресс фундаментальной науки, поэтому в разумной степени мы должны опираться на технические соображения, а не на эмоциональные предпочтения, а для этого мода, вера и фантазия должны содержать некоторый объем серьезной математики. Однако мой рассказ не задумывался как технические выкладки, понятные лишь опытным математикам или физикам, — нет, я стремился, чтобы книгу могли с пользой для себя прочесть неспециалисты. И я постарался свести технические детали к разумному минимуму. Однако существуют математические концепции, которые очень пригодятся для понимания тех важнейших проблем, что я хочу здесь затронуть. Поэтому в приложении я поместил одиннадцать довольно простых математических разделов. Там приведена не самая сложная математика, но при необходимости она вполне поможет неспециалистам глубже понять многие из основных проблем.

В первых двух из этих разделов (A1 и A2) рассмотрены очень простые идеи, правда, не слишком известные, и нет никаких сложных формул. Однако эти идеи играют особую роль в аргументации, приведенной в этой книге, и особенно это касается модных гипотез, рассматриваемых в главе 1. Любой читатель, желающий понять обсуждаемую в ней центральную и важную проблему, на том или ином этапе должен ознакомиться с материалом разделов A.1 и A.2, где я излагаю

ключевые аргументы против того, что в нашей физической Вселенной якобы существуют дополнительные пространственные измерения. Такая надразмерность — важнейший предмет обсуждения в большинстве современных теорий струн и их основных вариантах. Мои критические аргументы направлены против нынешнего «струнно-обусловленного» воззрения, которое заключается в том, что в физическом пространстве должно насчитываться больше трех измерений, доступных для нашего непосредственного восприятия. Здесь я обращаюсь к ключевой проблеме функциональной свободы, а в разделе A.8 даю более развернутые выкладки, поясняющие основную идею. Корни рассматриваемой в данном случае математической проблемы восходят к работам великого французского математика Эли Картана, а сама идея впервые появляется на рубеже XII века. Однако она почти не пользуется вниманием современных физиков-теоретиков, хотя и очень важна с точки зрения правдоподобия актуальных физических идей, касающихся надразмерности.

Теория струн и ее современные разновидности далеко продвинулись за последние годы. С тех пор как были прочитаны упоминаемые здесь принстонские лекции, они обогатились многими новыми техническими деталями. Я ни в коем случае не претендую на экспертное владение этой темой, хотя и перелопатил немало соответствующего материала. Меня волнуют не столько детали, сколько сам вопрос о том, на самом ли деле эта работа помогает нам куда-то двигаться, полнее понимать реальный физический мир, в котором мы живем. Наиболее важно, что я почти не вижу попыток решить вопрос избыточной функциональной свободы, обусловленной предполагаемой пространственной надразмерностью. На самом деле ни в одной работе по теории струн, которые мне доводилось видеть, эта проблема даже не упоминается. Меня это порядком удивляет, и не только потому, что данной проблеме было уделено основное внимание в первой из трех моих принстонских лекций, прочитанных более 10 лет тому назад. Ранее я выступил с лекцией на конференции, состоявшейся в Кембриджском университете в январе 2002 года и приуроченной к шестидесятилетию Стивена Хокинга, и упоминал в ней эту проблему. В зале присутствовали несколько ведущих исследователей, занимающихся теорией струн. Впоследствии я видел и письменные источники, в которых была зафиксирована эта лекция.

Здесь я должен сделать важное замечание. Квантовые физики зачастую отбрасывают проблему функциональной свободы, считая, что она применима лишь к классической физике. Те сложности, которые она представляет для надразмерных теорий, обычно огульно отвергаются, причем нам стараются продемонстрировать несущественность этих феноменов в квантово-механических ситуациях. В разделе 1.10 я излагаю основные возражения против этого базового

аргумента, поэтому особенно рекомендую прочитать этот раздел сторонникам пространственной надразмерности. Надеюсь, что, повторив эти аргументы здесь и разработав их в некоторых дополнительных физических контекстах (см. разделы 1.10, 1.11, 2.11 и A.11), я подвигну кого-нибудь как следует рассмотреть их в последующих работах.

В оставшихся разделах приложения читатель кратко ознакомится с векторными пространствами, многообразиями, расслоениями, гармоническим анализом, комплексными числами и их геометрией. Эти темы определенно хорошо знакомы экспертам, но этот самодостаточный образовательный материал может помочь непрофессионалу разобраться с более специальными разделами книги. Во всех моих описаниях я воздерживался от какого-либо серьезного введения в идеи дифференциального (или интегрального) исчисления. Без сомнения, читателям полезно иметь понятие о математическом анализе, но те, кто такого понятия не имеют, ничего не поймут из столь поверхностно изложенного материала по данной теме. Однако в разделе A.11 я все-таки слегка затронул проблемы дифференциальных операторов и дифференциальных уравнений, чтобы пояснить некоторые вопросы, важные для понимания всего текста.

ГЛАВА 1

Мода

1.1. Математическая красота как движущая сила

Как упоминалось в предисловии, проблемы, рассматриваемые в этой книге, были вкратце очерчены в трех лекциях, которые я прочитал в Принстонском университете в октябре 2003 года по приглашению издательства *Princeton University Press*. Пожалуй, на протяжении этих лекций, прочитанных столь сведущей аудитории, как принстонское научное сообщество, я больше всего волновался, рассказывая о моде. Ведь я решил проиллюстрировать проблему моды именно на примере теории струн и различных ее дочерних теорий. А эти исследования как раз в Принстоне достигли таких высот, как, пожалуй, более нигде в мире. Кроме того, сама тема была глубоко специальной, а я не мог претендовать на полную компетентность во многих ее важных деталях. Мои представления об этих тонкостях были достаточно ограниченны, учитывая, что в этой сфере я считался чужаком. Однако мне казалось, что я не вправе пасовать перед подобными трудностями, поскольку если считать, что важные комментарии могут делать лишь те, кто полностью в теме, то критика, вероятно, ограничится относительно малосущественными деталями, а более серьезные и обширные темы, несомненно, в значительной степени будут проигнорированы.

С тех пор как были прочитаны упомянутые лекции, вышли три исключительно важные книги по теории струн: «Даже не неправильно» Питера Войта (*Peter Woit*. «Not Even Wrong»), «Неприятности с физикой» Ли Смолина (*Lee Smolin*. «The Trouble with Physics») и «Прощание с реальностью: как сказочная физика предает поиск научной истины» Джима Бэггота (*Jim Baggot*. «Farewell to Reality: How Fairytale Physics Betrays the Search for Scientific Truth»). Разумеется, Войт и Смолин гораздо лучше меня знакомы с сообществом физиков-струнников и представляют себе их ультрамодный статус. Но я сформулировал критические замечания к теории струн, вошедшие в главы 31 и 34 моей книги «Путь к реальности», опубликованной еще раньше трех вышеупомянутых работ.

Однако моими ремарками относительно физической роли теории струн, пожалуй, можно пренебречь по сравнению с замечаниями этих ученых. На самом деле я в основном буду давать общие комментарии, которые почти не касаются узкоспециальных деталей.

Давайте, как и подобает, начнем с общего (и, пожалуй, очевидного) утверждения. Отметим тот факт, что впечатляющий прогресс, в самом деле достигнутый физической наукой за несколько последних веков, зиждется на исключительно точных и изощренных математических построениях. Это очевидно, а значит, любой дальнейший значительный прогресс также должен базироваться на некоем четком математическом аппарате. Чтобы любая физическая теория могла совершенствоваться на базе уже имеющихся достижений, давать точные и недвусмысленные прогнозы, превосходящие наши прошлые достижения, она также должна быть основана на строгой математической схеме. Более того, предполагается, что качественная математическая теория должна быть математически осмысленной, то есть *математически непротиворечивой*. Из внутренне противоречивой схемы, в принципе, кто угодно может вывести любой устраивающий его ответ.

Тем не менее внутренняя непротиворечивость — достаточно сильный критерий, и оказывается, что лишь немногие версии физических теорий — даже тех, что ранее считались очень успешными, — действительно обладают полной внутренней непротиворечивостью. Иногда для правильного и однозначного применения теории приходится дополнительно к математике привлекать и достаточно сложные физические соображения. Разумеется, эксперименты играют ключевую роль для физической теории, и экспериментальная проверка теории очень отличается от проверки теории на логическую непротиворечивость. Обе проверки важны, но практика показывает, что физики не слишком стараются добиться полной математической непротиворечивости теории, если на вид она согласуется с физическими фактами. Отчасти как раз такая ситуация сложилась и с исключительно успешной квантово-механической теорией, в чем мы убедимся на страницах главы 2 (и в разделе 1.3). Самая первая работа на эту тему, а именно эпохальная попытка Макса Планка объяснить частотный спектр электромагнитного излучения, находящегося в равновесии с веществом при заданной температуре (спектр абсолютно черного тела; см. разделы 2.2 и 2.11), потребовала построить отчасти гибридную картину, которая на самом деле не была внутренне непротиворечивой [Pais, 1982]. Точно так же нельзя сказать, что старая квантовая теория атома, блестяще предложенная Нильсом Бором в 1913 году, обладала полной внутренней согласованностью. При последующей разработке квантовой теории были возведены крайне замысловатые математические построения — движущей силой всей этой работы было стремление к математиче-

ской непротиворечивости. Однако современная квантовая теория по-прежнему до конца не убирает некоторые противоречия, и в этом мы убедимся далее, в частности в разделе 2.13. Тем не менее именно *экспериментальное* подтверждение является железобетонным фундаментом квантовой теории, причем оно обнаруживается при исследовании самых разнообразных физических явлений. Физики не слишком заморачиваются ювелирными тонкостями математической или онтологической несогласованности, если теория, в которой применяются адекватные суждения и тщательные вычисления, по-прежнему дает ответы, превосходно согласующиеся с результатами наблюдений — иногда с поразительной точностью — на материале тонких и филигранных экспериментов.

С теорией струн складывается принципиально иная ситуация. В данном случае нет вообще *никаких* результатов, которые обеспечивали бы этой теории экспериментальную поддержку. Существует мнение, что это и неудивительно, так как теория струн, на данный момент формулируемая в общем и целом как *теория квантовой гравитации*, фундаментальным образом сопряжена с *планковскими масштабами*, то есть с очень малыми расстояниями, которые составляют примерно 10^{-15} или 10^{-16} от тех, что доступны для современных экспериментальных исследований. Следовательно, на таких расстояниях действуют энергии примерно в 10^{15} или 10^{16} раз выше, чем доступны в современных экспериментах. При этом следует отметить, что, согласно базовым принципам теории относительности, малое расстояние, в сущности, эквивалентно малому времени, и это связано со скоростью света; кроме того, согласно базовым принципам квантовой механики, малое время в сущности тождественно большой энергии, и это связано с постоянной Планка (см. разделы 2.2 и 2.11). Приходится признать очевидный факт: при всей мощности современных ускорителей частиц потенциально достижимые на них в настоящее время энергии и близко не дотягивают до тех, которые могли бы иметь непосредственное значение для таких теорий, как современная теория струн, — речь идет о теориях, в которых мы пытаемся применять принципы квантовой механики к гравитационным феноменам. Однако такое состояние физической теории едва ли можно считать удовлетворительным, поскольку окончательный критерий, позволяющий судить о жизнеспособности физической теории, это экспериментальное подтверждение.

Разумеется, возможно, что мы как раз подходим к новому этапу базовых исследований в фундаментальной физике, где требования математической непротиворечивости приобретают первостепенное значение, и в ситуациях, когда такие требования (плюс согласованность с ранее установленными принципами) оказываются недостаточными, следует привлекать дополнительные критерии *математической красоты* и простоты. Хотя и может показаться, что ненаучно апеллировать к таким эстетическим пожеланиям при совершенно объективном

поиске физических законов, лежащих в основе Вселенной, примечательно, сколь плодотворными — принципиально плодотворными на самом деле — зачастую оказываются такие эстетические суждения. Физике известно много примеров, в которых красивая математическая идея вела к фундаментальному прорыву в понимании чего-либо. Великий физик-теоретик Поль Дирак совершенно недвусмысленно высказывался о важности эстетических соображений, описывая, как открыл уравнение для движения электрона и спрогнозировал существование античастиц. Определенно, уравнение Дирака оказалось абсолютно фундаментальным для оснований физики, и эстетическая привлекательность этого уравнения получила очень широкое признание. Идея античастиц интересна и тем, что она возникла в результате глубокого анализа, которому Дирак подверг собственное уравнение, описывавшее электрон.

Однако роль эстетических соображений — это проблема, о которой очень сложно говорить объективно. Зачастую бывает, что то или иное построение, кажущееся некоему физику очень красивым, будет с жаром *отвергнуто* другим физиком! Говоря об эстетических соображениях, следует отметить, что они могут содержать изрядный «модный» компонент — этим физика ничуть не отличается от искусства или дизайна одежды.

Следует четко отметить, что вопрос эстетических соображений в физике тоньше принципа, зачастую именуемого *бритвой Оккама* — то есть устранения излишней сложности. Действительно, вынесение решения о том, какая из двух противоборствующих теорий проще и, соответственно, возможно, элегантнее, — отнюдь не простая задача. Например, проста ли эйнштейновская общая теория относительности? Она проще или сложнее ньютоновской теории тяготения? Или поставим вопрос так: теория Эйнштейна проще или сложнее, чем теория, выдвинутая в 1894 году Аспетом Холлом, за 21 год до того, как Эйнштейн представил общую теорию относительности? Эта теория была очень похожа на ньютоновскую — с той оговоркой, что важный для тяготения закон обратных квадратов был заменен другим, где гравитационное притяжение между массой M и массой m равно $G\,m\,M\,r^{-2{,}00000016}$, а не $G\,m\,M\,r^{-2}$, как у Ньютона. Холл предложил свою теорию, чтобы объяснить небольшое отклонение от предсказания теории Ньютона, наблюдавшееся в смещении перигелия планеты Меркурий; это смещение было известно с 1843 года. (Перигелий — это точка на орбите планеты, в которой она находится ближе всего к Солнцу [Roseveare, 1982].) Кроме того, эта теория чуть лучше описывала движение Венеры, нежели теория Ньютона. В некотором смысле теория Холла была лишь незначительно сложнее ньютоновской, однако все зависит от того, какую сложность мы усматриваем в замене красивого простого числа 2 на 2,00000016. Несомненно, при такой замене математическая красота теряется, но, как было сказано выше, таким суждениям присущ сильный

субъективный элемент. Можно еще точнее выразиться, что из закона обратных квадратов следует ряд красивых математических свойств (в целом выражающих сохранение потока гравитационного поля, что нарушается в теории Холла). Но опять же, эту проблему можно счесть эстетической — физическое значение ее не стоит переоценивать.

А как насчет эйнштейновской общей теории относительности? Разумеется, как только речь заходит об обсуждении деталей теорий, применение теории Эйнштейна к конкретным физическим системам — гораздо более сложный процесс, чем применение теории Ньютона (и даже Холла). Если просто взять и записать уравнения, то для теории Эйнштейна они будут несравнимо сложнее — их даже нелегко выписать во всех деталях. Более того, решать их неизмеримо труднее, причем в теории Эйнштейна присутствует много *нелинейностей*, нехарактерных для теории Ньютона, и тут уже не работают простые аргументы из разряда сохранения потока, которые несостоятельны уже в рамках теории Холла. (Смысл *линейности* подробнее описан в разделах A.4 и A.11, а ее особая роль в квантовой механике — в разделе 2.4). Кроме того, физическая интерпретация эйнштейновской теории напрямую зависит от устранения фиктивных эффектов, возникающих при выборе конкретной системы координат, тогда как считается, что выбор координатной системы не должен иметь физического значения. На практике обращаться с теорией Эйнштейна гораздо сложнее, чем с ньютоновской (и даже холловской) теорией тяготения.

Однако в одном отношении теория Эйнштейна все-таки очень проста — возможно, даже проще (или «естественнее»), чем ньютоновская. Теория Эйнштейна базируется на математической теории римановой (вернее, как мы увидим в разделе 1.7, *псевдоримановой*) геометрии произвольно искривленных четырехмерных многообразий (см. также раздел A.5). Это отнюдь не простой для усвоения корпус математического материала, поскольку здесь требуется понимать, что такое тензор и каковы его свойства, а также как построить конкретный тензорный объект $\mathbf{R}$ — так называемый *риманов тензор кривизны* — из *метрического тензора* $\mathbf{g}$, определяющего геометрию. Затем путем свертки и ряда математических операций мы строим *тензор Эйнштейна* $\mathbf{G}$. Тем не менее общие геометрические идеи, выходящие за рамки этих формулировок, достаточно просто уловить, и стоит вам как следует понять детали такой криволинейной геометрии, как оказывается, что на самом деле можно записать весьма ограниченное семейство возможных (или вероятных) уравнений, которые удовлетворяли бы предложенным общефизическим и геометрическим требованиям. Простейшая из таких возможностей дает нам знаменитое эйнштейновское уравнение поля $\mathbf{G} = 8\pi\gamma\mathbf{T}$, соответствующее общей теории относительности (где $\mathbf{T}$ — тензор энергии-импульса материи, а γ — гравитационная постоянная Ньютона, опре-

деляемая согласно конкретной ньютоновской формулировке, поэтому даже 8π — никакое не осложнение, а просто вариант определения γ).

Можно произвести небольшую и при этом еще более простую модификацию эйнштейновского уравнения поля, при которой все основные требования описанной схемы не изменяются. Речь идет о включении постоянной Λ — так называемой *космологической постоянной* (которую Эйнштейн ввел в 1917 году по причинам, впоследствии им же забракованным). С постоянной Λ уравнение Эйнштейна примет вид $\mathbf{G} = 8\pi\gamma\mathbf{T} + \Lambda\mathbf{g}$. В настоящее время величину Λ часто именуют *темной энергией* — она вводится с целью обобщения теории Эйнштейна таким образом, чтобы значение Λ могло варьироваться. Однако существуют строгие математические ограничения на возможные значения космологической постоянной, и в разделах 3.1, 3.7, 3.8 и 4.3, где величина Λ сыграет важную роль, я ограничусь рассмотрением таких ситуаций, в которых она действительно постоянна. Космологическая постоянная будет для нас важна в главе 3 (а также в разделе 1.15). Действительно, результаты самых последних наблюдений убедительно указывают, что Λ физически существует в реальности и имеет очень небольшое, по-видимому, положительное значение. Свидетельства в пользу $\Lambda > 0$ — или, возможно, в пользу существования какой-либо более универсальной «темной энергии» — сейчас выглядят очень впечатляюще, и их число постоянно растет со времен первых наблюдений, проведенных Перлмуттером и др. [1999], Риссом и др. [1998] и их коллегами. За эти работы Сол Перлмуттер, Брайан П. Шмидт и Адам Г. Рисс были удостоены Нобелевской премии по физике за 2011 год. Выражение $\Lambda > 0$ дает ощутимый эффект только в огромнейших космологических масштабах, а более локальные движения небесных тел можно адекватно описывать при помощи исходного и более простого эйнштейновского уравнения $\mathbf{G} = 8\pi\gamma\mathbf{T}$. Сегодня установлено, что это уравнение с беспрецедентной точностью моделирует движение небесных тел под действием силы тяжести, и значение Λ не оказывает существенного влияния на такую локальную динамику.

Исторически в данном контексте наиболее важную роль сыграла система двойной звезды PSR1913+16, в которой присутствует *пульсар*, испускающий чрезвычайно регулярные электромагнитные сигналы, улавливаемые на Земле. Вращение одной звезды вокруг другой — это явление, имеющее практически чисто гравитационную природу, и общая теория относительности моделирует такое явление со столь огромной точностью, что погрешность составляет не более 10^{-14} за 40 лет. В 40 годах около 10^9 секунд, поэтому точность 10^{-14} означает, что соответствие наблюдений и теории прослеживается до 10^{-5} (одной стотысячной) секунды за весь этот период — что, собственно, и подтверждается. Не так давно были обнаружены и другие системы [Kramer et al., 2006],

в которых присутствует пульсар или даже два пульсара, и в них точность предсказаний может оказаться даже выше, если наблюдать их примерно так же долго, как и PSR19+16.

Однако если мы скажем, что 10^{-14} — это наблюдаемая степень точности общей теории относительности, то закономерен один вопрос. Действительно, конкретные массы и орбитальные характеристики должны вычисляться на основе наблюдаемых движений, это не должны быть числа, полученные из теории или независимых наблюдений. Более того, такая экстраординарная точность во многом присутствует уже в ньютоновской теории тяготения.

Тем не менее здесь нас интересует вся теория тяготения в целом, а эйнштейновская теория содержит дедуктивные выводы из ньютоновской (описывающей кеплеровские орбиты и т. д.) в первом приближении, но дает различные поправки относительно кеплеровских орбит (в частности, речь идет о смещении перигелия) и, наконец, предсказывает потерю энергии в системе: любая масса, движущаяся с ускорением, должна терять энергию, испуская гравитационные волны. Это «рябь» пространства-времени, гравитационный аналог электромагнитных волн (то есть света), излучаемых ускоренно движущимися электрически заряженными телами. Дополнительное подтверждение существования гравитационного излучения и правильной формы этих волн было получено в 2016 году [Abbott et al., 2016], когда их обнаружил детектор LIGO, причем это открытие великолепно подтверждает еще один прогноз общей теории относительности: существование черных дыр, о чем мы поговорим в разделе 3.2, в последующих разделах главы 3 и в разделе 4.3.

Следует подчеркнуть, что точность, зафиксированная в этом эксперименте, несравнимо выше — еще в 10^8 (сто миллионов) раз или более — точности тех экспериментов, на которые опирался Эйнштейн, когда только формулировал теорию относительности. Погрешность ньютоновской теории тяготения составляет около 10^{-7}. Таким образом, можно сказать, что общая теория относительности с ее точностью предсказаний 10^{-14} уже «существовала» в природе до того, как Эйнштейн ее сформулировал. Дополнительная точность порядка 10^{-8}, неизвестная Эйнштейну, никак не могла отразиться на теории относительности в эйнштейновской трактовке. Соответственно, такая новая математическая модель природы — это не человеческий конструкт, изобретенный просто при попытке найти теорию, которая бы лучше всего соответствовала фактам; эта математическая схема уже в буквальном смысле была вплетена в устройство природы. Эта математическая простота или красота — как бы вы ее ни назвали — подлинная составляющая природы, и дело не в том, что наш разум просто склонен впечатляться такой математической красотой.

С другой стороны, если мы будем сознательно ориентироваться на критерий математической красоты, формулируя наши теории, то легко свернем не туда. Общая теория относительности и в самом деле очень красива, но как судить о красоте физических теорий? Эстетические предпочтения у разных людей очень различаются. Далеко не очевидно, что нечто красивое с точки зрения одного человека покажется столь же красивым другому, и могут ли чьи-то эстетические суждения считаться весомее суждений другого при формулировке успешной физической теории? Более того, красота, присущая теории, на первый взгляд не очевидна — она может быть выявлена лишь позднее, когда новые технические наработки позволят прояснить математическую структуру этой теории. Типичный пример — ньютоновская динамика. Многие, несомненно, красивые черты ньютоновских построений были обнаружены лишь гораздо позднее, по результатам блестящих трудов таких великих математиков, как Эйлер, Лагранж, Лаплас и Гамильтон (о чем свидетельствуют, например, такие названия, как *уравнение Эйлера — Лагранжа*, *оператор Лапласа*, *лагранжиан* и *гамильтониан* — ключевые элементы современной теоретической физики). Например, третий закон Ньютона, согласно которому на каждое действие существует равносильное противодействие, направленное в противоположную сторону, играет центральную роль в лангранжевой формулировке современной физики. Я бы не удивился, если бы оказалось, что вся та красота, которую приписывают современным успешным физическим теориям, обнаруживалась в них *постфактум*. Сама успешность физической теории — как на уровне математики, так и на уровне наблюдений — может значительно влиять на эстетические качества, которые впоследствии ей приписываются. Из всего вышесказанного следует, что суждения о достоинствах выдвигаемой физической теории, связанные с ее эстетической привлекательностью, скорее всего, окажутся неочевидными или по меньшей мере неоднозначными. Бесспорно, суждения о новой теории будут более надежными, если формулировать их в зависимости от того, насколько теория согласуется с результатами наблюдений и какова ее прогностическая сила.

Однако что касается экспериментального подтверждения, зачастую важнейшие эксперименты поставить не удается. Например, такова ситуация с недостижимо высокой энергией, которую хотелось бы сообщить отдельно взятой частице. Эта энергия выше значений, достижимых в современных ускорителях частиц (см. раздел 1.10), и тем не менее часто утверждается, что как раз такие энергии потребовались бы для адекватной наблюдательной проверки любой теории квантовой гравитации. Более скромные экспериментальные запросы также могут оказаться нереализуемыми — отчасти из-за дороговизны самих экспериментов, отчасти из-за их неотъемлемой сложности. Даже с исключительно успешными экспериментами зачастую бывает так, что они позволяют собрать огромный объем данных, а проблема заключается в том, чтобы извлечь

ключевую информационную составляющую из этого вороха сведений. Определенно, как раз такая ситуация складывается в физике частиц, где мощные ускорители и коллайдеры сегодня позволяют собирать колоссальные объемы информации. Проблема также приобретает актуальность и в космологии, где современные наблюдения космического фонового микроволнового излучения дают очень большие объемы информации (см. разделы 3.4, 3.9 и 4.3). Многие из этих данных считаются не слишком полезными, поскольку просто подтверждают то, что было выяснено ранее на материале предыдущих экспериментов. Требуется трудоемкая статистическая обработка, позволяющая извлечь некий исчезающе малый осадок — новую характеристику, интересующую экспериментаторов, — что позволило бы подтвердить или опровергнуть некое теоретическое предположение.

Здесь следует отметить, что статистическая обработка обычно очень специфична для конкретной теории, она специально рассчитана на поиск тех малозаметных дополнительных эффектов, которые может прогнозировать теория. Вполне возможно, что какая-либо радикально иная совокупность идей, совсем не похожих на те, что сегодня в моде, до сих пор не проверена, хотя окончательный ответ может скрываться в уже собранных данных и при этом ускользать от физиков, так как применяемые статистические процедуры слишком жестко настроены на актуальную теорию. Поразительный пример такого рода мы рассмотрим в разделе 4.3. Но даже когда понятно, как именно извлечь ключевую информацию из имеющегося множества надежных данных, непреодолимым препятствием становится то огромное количество компьютерного времени, которое требуется для выполнения анализа, — особенно если в очереди также стоят более модные изыскания.

Еще важнее тот факт, что сами эксперименты обычно невероятно дороги, и уже на этапе проектирования их подстраивают под проверку тех теорий, которые вписываются в картину общепринятых идей. Любая теоретическая схема, которая слишком далека от принятого мнения, с трудом привлечет достаточные средства, необходимые для ее проверки. В конце концов, чтобы сконструировать очень дорогой экспериментальный аппарат, требуется одобрение множества комитетов, состоящих из признанных экспертов, а каждый их этих экспертов, вероятно, внес свой вклад в формирование нынешней научной парадигмы.

Говоря об этой проблеме, можно обсудить Большой адронный коллайдер (БАК), расположенный в швейцарском городе Женева. Работа над коллайдером была закончена в 2008 году. В это устройство входит 27-километровый туннель, проложенный на территории двух государств — Франции и Швейцарии. Впервые

запущенный в эксплуатацию в 2010 году, в настоящее время БАК известен как ускоритель, на котором удалось зафиксировать прежде неуловимый бозон Хиггса, имеющий огромную важность для всей физики частиц, особенно в контексте той роли, которую он играет при возникновении массы у частиц, участвующих в слабых взаимодействиях. В 2013 году Нобелевская премия по физике была присуждена Питеру Хиггсу и Франсуа Энглеру за их роль в эпохальной работе, в ходе которой было предсказано существование этой частицы, а также спрогнозированы ее свойства.

Несомненно, это великое достижение, и я ни в коем случае не пытаюсь преуменьшить его важность. Тем не менее случай БАК кажется очень характерным. Тот способ, которым анализируются столкновения частиц высоких энергий, предполагает использование очень дорогих детекторов, настроенных именно с расчетом на поиск информации, связанной с господствующей теорией физики частиц. В связи с этим совсем непросто добыть информацию, важную в контексте нетрадиционных идей и связанную с базовой природой основополагающих частиц и их взаимодействий. В целом те гипотезы, что резко противоречат господствующей точке зрения, почти не имеют шансов на адекватное финансирование, а кроме того, вполне могут вообще не дойти до проверки в убедительных экспериментах.

Следующий важный факт заключается в том, что аспиранты, подыскивающие тему для диссертации, обычно сильно ограничены в выборе подходящих направлений для исследования. Студенты, работающие в немодных областях, конечно, могут успешно защитить диссертацию, но после этого, вероятно, столкнутся с огромными сложностями при поиске академической должности — независимо от того, насколько этот соискатель талантлив, сведущ или оригинален. Число академических постов ограниченно, а выбить финансирование для исследований сложно. Скорее всего, научный руководитель будет заинтересован в разработке прежде всего тех идей, которые сам и продвигает, а эти идеи, вероятно, будут относиться к областям, уже ставшими модными. Более того, научный руководитель, заинтересованный в разработке «немейнстримовой» идеи, вполне возможно, не решится советовать потенциальному студенту работать в этой сфере, понимая, насколько сложно будет тому впоследствии конкурировать на рынке труда — там, где компетентность в модных научных областях является очевидным преимуществом.

Те же проблемы возникают при финансировании исследовательских проектов. Проекты из модных областей имеют гораздо больше шансов на одобрение (см. также раздел 1.12). Опять же, судить о предложениях будут признанные эксперты, которые (с огромной вероятностью) работают в уже ставших модными об-

ластях, поэтому и они сами могут принимать активное участие в распределении финансирования. Если проект слишком сильно отклоняется от общепринятых норм, то, скорее всего, он останется без поддержки, даже если будет хорошо продуман и в высшей степени оригинален. Кроме того, дело не только в ограниченности доступных средств: влияние моды сказывается даже в США, где на научные исследования по-прежнему тратятся относительно крупные суммы.

Разумеется, следует отметить, что в большинстве немодных исследовательских областей с гораздо меньшей вероятностью будут развиты успешные теории по сравнению с теми областями, которые уже стали модными. В огромном большинстве случаев радикально новая точка зрения почти не имеет шансов превратиться в жизнеспособную гипотезу. Стоит ли говорить, что любая столь же революционная идея, как общая теория относительности Эйнштейна, сама по себе должна согласовываться с уже имеющимися экспериментальными данными, а в противном случае дорогостоящие эксперименты могут и не понадобиться — не выдерживающие критики идеи будут отвергнуты и без них. Однако если теоретические версии не противоречат ни одному из поставленных экспериментов, но на данный момент отсутствуют реальные перспективы их подтвердить или опровергнуть (по вышеописанным причинам), то, полагаю, мы вновь должны прибегать к критериям математической непротиворечивости, общей применимости и красоты, когда беремся судить о правдоподобии и значимости предлагаемой физической теории. Именно в таких обстоятельствах, когда мода начинает играть довлеющую роль, мы должны быть очень внимательны и не поддаваться на модность той или иной теории, поскольку из-за этого можем необъективно судить о физическом правдоподобии такой теории.

1.2. Некоторые модные идеи прошлого

Все это особенно важно для тех теорий, которые претендуют на изучение самых основ физической реальности — такова современная теория струн, и мы должны быть очень осторожны, приписывая чрезмерную достоверность такой теории лишь в силу ее модности. Однако прежде чем рассуждать о современных физических идеях, будет уместно упомянуть некоторые модные научные теории прошлого, которые сегодня всерьез уже не воспринимаются. Таких теорий много, и я уверен, что большинство читателей почти ничего не знают о них — на том достаточном основании, что если теория всерьез не воспринимается, то и изучать ее мы, скорее всего, не будем, — конечно, если не являемся маститыми специалистами по истории науки. Но среди физиков таких не много. Позвольте по меньшей мере упомянуть несколько наиболее известных из этих теорий.

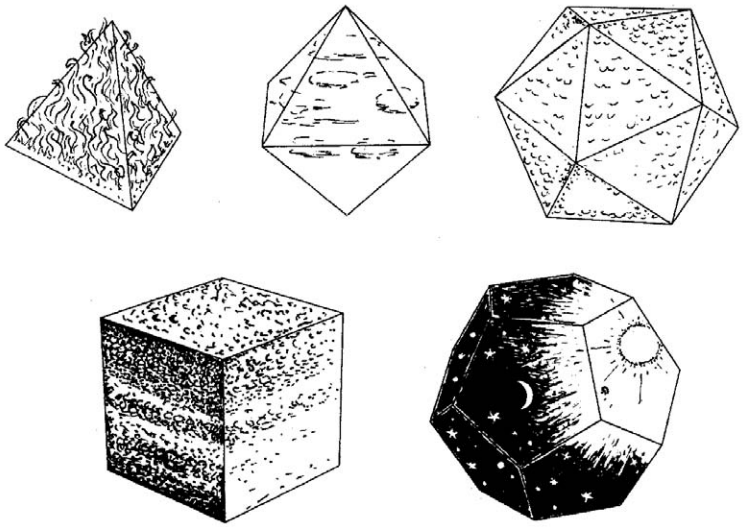


Рис. 1.1. Пять элементов, известных в Древней Греции: огонь (тетраэдр), воздух (октаэдр), вода (икосаэдр), земля (куб) и эфир (додекаэдр)

Например, в Древней Греции существовала теория, согласно которой платоновы тела должны быть связаны с так называемыми первоэлементами материи, как показано на рис. 1.1. Здесь *огонь* представлен в виде правильного тетраэдра, *воздух* — в виде октаэдра, *вода* — в виде икосаэдра, а *земля* — в виде куба. Наряду с ними впоследствии стали выделять небесный *эфир* (также именовавшийся *небесный свод* или *квинтэссенция*); предполагалось, что из него состоят небесные тела, и эфир изображался в виде правильного додекаэдра. Поскольку древние греки сформулировали такое мировоззрение, можно предположить, что в свое время это считалось модной теорией.

Изначально греки выделяли всего четыре элемента — огонь, воздух, воду и землю, — и такое число первоэлементов хорошо сочеталось с четырьмя известными на тот момент правильными многогранниками. Однако когда впоследствии был открыт додекаэдр, теорию потребовалось дополнить, чтобы она объясняла существование такого нового многогранника! Соответственно, в систему многогранников была привнесена небесная субстанция, из которой состояли такие якобы совершенные тела, как Солнце, Луна и планеты, а также хрустальные сферы, к которым эти тела предположительно прикреплялись. Древним ученым казалось, что эта субстанция подчинялась совершенно иным законам, чем те, что действуют на Земле. Предполагалось, что эфир находится в вечном движении и этим отличается от обычных субстанций, тела из которых

постепенно теряют скорость и останавливаются. Возможно, из этого можно извлечь урок о том, как даже современные затейливые теории, изначально представляемые как нечто вроде бы определенное, впоследствии могут существенно видоизменяться, а их исходные положения распространяются до немыслимых ранее пределов, подстраиваясь под новые теоретические или эмпирические факты. Насколько я понимаю, древние греки считали, что законы, управляющие движением звезд, планет, Луны и Солнца, действительно очень отличались от законов, которым подчиняются земные тела. Должен был появиться Галилей с его представлениями об относительности движения, а затем Ньютон с универсальной теорией тяготения, на которого также весьма повлияли представления Кеплера об орбитальном движении планет, чтобы мы могли осознать, что небесные и земные тела на самом деле подчиняются одним и тем же законам.

Впервые познакомившись с этими древнегреческими идеями, я воспринимал их как чистейшую романтическую фантазию, лишенную какого-либо математического обоснования (не говоря уже о физическом). Но недавно я узнал, что эти идеи опирались на более серьезный теоретический базис, чем мне когда-то казалось. Некоторые из упомянутых многогранников можно разрезать на кусочки, из которых затем легко сложить другие правильные многогранники (например, два куба можно разрезать таким образом, чтобы из них получилось два тетраэдра и октаэдр). Вероятно, эта закономерность соотносилась с физическими свойствами и использовалась как геометрическая модель для описания возможных превращений одних элементов в другие. Как минимум это была смелая и образная догадка о природе вещественных субстанций, которая в свое время не казалась беспочвенной, ведь так мало было известно об истинной природе и свойствах физической материи. Это была одна из первых попыток описать реальные вещества на основе красивой математической структуры — вполне в духе тех работ, которыми физики-теоретики по-прежнему пытаются заниматься сегодня, — где теоретические выводы из модели можно было проверить на материале реальных физических явлений. Здесь явно действуют и эстетические критерии, и эти идеи казались Платону привлекательными. Тем не менее излишне говорить, что детализация этих идей не смогла достойно выдержать проверку временем, иначе мы, конечно же, не отвергли бы идею, столь красивую с математической точки зрения!

Рассмотрим еще некоторые теории. Птолемеева модель движения планет, в которой Земля считалась неподвижной и располагалась в центре мироздания, была исключительно успешной и на протяжении многих веков не подвергалась сомнению. Вращение Солнца, Луны, планет было принято понимать в контексте *эпициклов*, согласно которым поведение планет можно объяснить в терминах

взаимного наложения равномерных круговых движений. Система, хотя и получилась достаточно сложной, хорошо согласовывалась с наблюдениями, кроме того, она была не лишена некоторой математической красоты и представляла собой теорию, которая довольно точно прогнозировала движение планет. Необходимо отметить, что эпициклы действительно можно убедительно обосновать, если рассматривать движения планет вокруг неподвижной Земли. Те движения, которые мы фактически наблюдаем, будучи на Земле, слагаются из вращения Земли (из-за которого кажется, что небеса циклически движутся вокруг земной оси, проходящей через полюса) и общих видимых движений Солнца, Луны и планет, причем эти движения в целом локализуются в плоскости эклиптики. Движения небесных тел кажутся нам почти идеальными круговыми вращениями вокруг различных осей. Геометры были давно знакомы с типом движений, когда одно вращение происходит в суперпозиции с другим вращением, поэтому было вполне логично предположить, что эта идея может проявляться и в более общем виде, то есть при движении планет.

Более того, эпициклы как таковые обладают интересной геометрией, а сам Птолемей был отличным геометром. В своей астрономической работе он опирался на красивую и мощную математическую теорему, которую, возможно, сам и доказал, поскольку сегодня она носит его имя. (Согласно этой теореме, если на плоскости есть четыре точки (A, B, C и D) и все эти точки — в таком порядке — лежат на одной окружности, то расстояния между ними удовлетворяют тождеству $AB \cdot CD + BC \cdot DA = AC \cdot BD$.) Такая теория движения планет была общепринятой на протяжении примерно четырнадцати веков, до тех пор пока ее не заменили (а в конечном итоге полностью опровергли) великолепные работы Коперника, Галилея, Кеплера и Ньютона. Сегодня теория Птолемея считается абсолютно неверной. Однако ее определенно можно было назвать модной, и при этом она была исключительно успешной в течение четырнадцати столетий (с середины II века до середины XVI века) и довольно точно описывала все наблюдаемые движения планет (с учетом того что время от времени в нее вносились те или иные корректировки), пока ближе к концу XVI века Тихо Браге не провел более точные измерения.

Другая модная теория, которую мы отвергли, была очень популярной на протяжении ста лет — с 1667 года, когда ее впервые сформулировал Иоганн Бехер, до 1778 года, когда ее фактически опроверг Антуан Лавуазье. Это флогистонная теория горения, согласно которой в любом горючем веществе содержится элемент под названием *флогистон*, и в процессе горения это вещество испускает свой флогистон в атмосферу. Флогистонная теория хорошо описывала большинство фактов о горении, известных на тот момент. Например, известно, что когда горение происходит в небольшой герметичной емкости, оно обычно

прекращается раньше, чем выгорит все топливо. Считалось, что при этом воздух в емкости настолько насыщен флогистоном, что больше уже не в состоянии принять. По иронии судьбы именно Лавуазье был автором другой модной, но ложной теории. Он считал теплоту материальной субстанцией и называл ее *теплород*. Эту теорию опроверг в 1798 году сэр Бенджамин Томпсон, граф Румфордский.

Успешность приведенных в этих двух примерах теорий можно понять, поскольку каждая из них весьма походила на более адекватную систему вещей, которая пришла ей на смену. В случае с птолемеевской динамикой переход к более удовлетворительной коперниковской гелиоцентрической системе мира производится путем геометрического преобразования. Мы просто берем за точку отсчета движения Солнце, а не Землю. Сначала, когда все описывалось в контексте эпициклов, разница была невелика — разве что гелиоцентрическая система казалась более упорядоченной: в ней те планеты, что располагались ближе к Солнцу, двигались быстрее [Gingerich, 2004; Sobel, 2011]. Но в принципе, обе системы на том этапе были эквивалентны. Однако когда Кеплер открыл три закона планетарного движения по *эллиптическим* орбитам, ситуация полностью изменилась, так как геоцентрическое описание такого движения не имело никакого геометрического смысла. Законы Кеплера стали ключом, позволившим открыть путь к исключительно точной и всеобъемлющей ньютоновской картине *всемирного тяготения*. Тем не менее сегодня можно не считать геоцентрическую картину столь возмутительной, какой она казалась в XIX веке, так как мы можем рассматривать ее в контексте *принципа общей ковариантности* , присущего общей теории относительности Эйнштейна (см. разделы 1.7, A.5 и 2.13) и позволяющего принять крайне неуклюжие координатные описания (например, геоцентрическую систему, где положение Земли со временем не изменяется) как допустимые. Аналогично, теорию флогистона можно переформулировать так, чтобы она близко соответствовала современным представлениям о горении; обычно считается, что при горении вещество соединяется с атмосферным кислородом, и флогистон можно было бы попросту считать «отрицательным кислородом». Так мы получим довольно последовательный переход от флогистоновой картины к общепринятой современной. Когда Лавуазье тщательно измерил массы веществ и показал, что масса флогистона должна была бы быть отрицательной, начался постепенный отказ от этой картины. Тем не менее «отрицательный кислород» — не такая уж абсурдная концепция с точки зрения современной физики частиц, где предполагается, что у каждого типа частиц (в том числе у сложных) должна быть античастица, следовательно, «атом антикислорода» ничуть не противоречит современной теории. Однако отрицательной массы у него не будет!

Иногда теории на некоторое время выходят из моды, но позже могут вновь стать модными в результате более поздних исследований. Здесь в качестве примера я приведу идею, выдвинутую лордом Кельвином (Уильямом Томпсоном) в 1867 году, согласно которой атомы (тогда считавшиеся элементарными частицами) могут состоять из миниатюрных узелковых структур. В то время идея привлекла значительное внимание, и математик Дж. Г. Тейт принялся систематически изучать эти узелки. Однако эта теория абсолютно не соответствовала истинным физическим свойствам атомов, поэтому о ней почти забыли. Тем не менее в последнее время идеи такого рода вновь стали вызывать интерес отчасти потому, что были созвучны представлениям теории струн. Математическая теория узлов также «воскресла» примерно в 1984 году, и виной тому работа Вона Джонса, чьи эпохальные идеи коренятся в теоретических размышлениях, связанных с квантовой теорией поля [Jones, 1985; Skyrme, 1961]. Впоследствии методы теории струн использовал Эдвард Виттен ([Witten, 1989], стремившийся сформулировать своеобразную квантовую теорию поля (так называемую *топологическую квантовую теорию поля*), которая в определенном смысле вписывает эти новые разработки в математическую теорию узлов.

В качестве возродившейся гораздо более древней идеи о крупномасштабной Вселенной я мог бы упомянуть — правда, не вполне серьезно — любопытное совпадение, случившееся примерно в то самое время, когда я читал в Принстоне лекцию, на которой основана эта самая глава (17 октября 2003 года). В той лекции я упоминал древнегреческую идею, согласно которой эфир должен быть связан с правильным додекаэдром. Я даже не знал, что в это самое время газеты писали о гипотезе Люмине и др. [Luminet et al., 2003], согласно которой трехмерная пространственная геометрия космоса на самом деле может обладать довольно сложной топологией, напоминающей (с оговоркой) противостоящие грани *правильного додекаэдра*. Соответственно, можно сказать, что платоновская идея о додекаэдрическом космосе также возродилась в наше время!

Амбициозная идея «теории всего», призванная описать все физические процессы, в том числе все элементарные частицы и их физические взаимодействия, широко обсуждается в последние годы, особенно в связи с теорией струн. Предполагается, что это будет полная теория всех физических явлений, основанная на некоторых представлениях о простейших частицах и/или полях, действующих в соответствии с некоторыми силами или другими динамическими принципами, управляющими движениями всех составных элементов. Здесь также можно говорить о возрождении древней идеи, в чем мы сейчас убедимся.

В то самое время, когда Эйнштейн готовил окончательную формулировку своей общей теории относительности, ближе к концу 1915 года математик Дэвид Гиль-

берт предложил собственный метод вывода уравнений поля теории Эйнштейна[1], воспользовавшись так называемым *вариационным принципом*. (В этой очень общей процедуре используется уравнение Эйлера — Лагранжа, содержащее лагранжиан (я уже упоминал и то и другое в разделе 1.1); см., например, книгу «Путь к реальности» [Penrose, 2004], глава 20. Далее эта книга будет обозначаться аббревиатурой ПкР — «Путь к реальности»). Эйнштейн, предпочитавший более прямолинейный подход, сформулировал свои уравнения именно в том виде, который позволял продемонстрировать, какие свойства будет иметь гравитационное поле (описываемое в терминах искривления пространства-времени), испытывающее влияние своего «источника», а именно общего распределения масс/энергий всех частиц или материальных полей и т. д., собранных в виде тензора энергии-импульса **T** (о нем шла речь в разделе 1.1).

Эйнштейн не оставил никаких конкретных указаний по поводу уравнений, описывающих механизм этого влияния; предполагалось, что такие уравнения будут взяты из какой-то иной теории, описывающей рассматриваемые в ней специфические материальные поля. В частности, одним из таких полей считалось электромагнитное поле, его полное описание выводится из уравнений великого шотландского физика и математика Джеймса Клерка Максвелла, полученных им в 1864 году. Эти уравнения позволили полностью унифицировать электрические и магнитные поля и тем самым объяснить природу света, а также многих других сил природы, управляющих внутренней организацией обычных веществ. В данном контексте это поле должно было считаться материей и играть соответствующую роль в тензоре энергии-импульса **T**. Кроме того, в **T** могут быть вовлечены и другие типы полей, а также всевозможные иные частицы. Они могли подчиняться любым уравнениям, которые оказались бы адекватны; при этом такие поля и частицы также считались бы материей. Такие подробности не были важны для теории Эйнштейна, поэтому отдельно не рассматривались.

В то же время Гильберт старался сформулировать свои предположения в более всеобъемлющем виде. Поэтому он описал систему, которую сегодня можно было бы назвать «*теорией всего*». Гравитационное поле должно было описываться так же, как в теории Эйнштейна, но Гильберт не оставлял исходный член **T** неуказанным, как это делал Эйнштейн, а предполагал, что этот член должен быть взят из очень специфической теории, модной в то время, — так называемой *теории Ми* [Mie, 1908, 1912a, b, 1913]. Она была связана с нелинейной модификацией электромагнитной теории Максвелла и предложена Густавом Ми в качестве схемы, призванной описать *все* аспекты материи. Соответственно, всеобъемлющая

[1] Комментарии к спорному вопросу о том, кто из ученых был первым, см. [Corry et al., 1997].

гипотеза Гильберта позиционировалась как полная теория материи (включая электромагнетизм) и тяготения. В то время сильные и слабые взаимодействия еще не были известны, но предложение Гильберта действительно могло трактоваться как «теория всего», если говорить современным языком. Однако полагаю, что лишь немногие современные физики хоть что-то слышали о некогда модной теории Ми, тем более о том, что она входила в состав гильбертовской общей теории относительности, претендовавшей на роль теории всего. Эта теория не играет никакой роли в современных представлениях о материи. Надеюсь, это послужит уроком нынешним физикам-теоретикам, собирающимся формулировать собственные «теории всего».

1.3. Физика частиц как основание для теории струн

Одним из таких теоретических построений является теория струн, и многие нынешние физики-теоретики в самом деле до сих пор считают, что это построение открывает верный путь к теории всего. Теория струн сформировалась из нескольких идей, о которых я впервые услышал около 1970 года (от Леонарда Сасскинда); тогда они показались мне поразительно интересными и довольно убедительными. Однако прежде чем описать эти идеи, я должен изложить их в подходящем контексте. Мы должны постараться понять, почему замена одного феномена другим — точечной частицы на небольшую петлю или искривление пространства, а такая замена и была сделана в исходной версии теории струн, может претендовать на роль основания для физической картины реальности.

На самом деле эта идея казалась привлекательной по многим причинам. По иронии судьбы, одна из наиболее конкретных причин, связанная с наблюдаемыми свойствами физических взаимодействий между адронами, по-видимому, совершенно потеряла актуальность под напором более современных исследований в теории струн, и я не уверен, что она до сих пор сохраняет какое-либо значение в этой сфере, кроме сугубо исторического. Тем не менее следует обсудить эту причину (этим мы подробнее займемся в разделе 1.6), а также поговорить о некоторых других проблемах, лежащих в основаниях современной физики частиц и породивших базовые принципы теории струн.

Во-первых, расскажу, что такое адрон. Как известно, обычный атом состоит из положительно заряженного ядра и отрицательно заряженных электронов, вращающихся вокруг него. Ядро состоит из протонов и нейтронов — вместе они именуются *нуклонами* (N), причем протон имеет положительный заряд, равный единице (единица положительного заряда подбиралась так, чтобы по модулю она была равна отрицательному заряду электрона), а заряд нейтрона нулевой.

Электрическая сила притяжения между положительными и отрицательными зарядами удерживает отрицательно заряженные электроны на их орбитах вокруг положительно заряженного ядра. Однако если бы на них действовали только электрические силы, то само ядро (кроме ядра атома водорода, состоящего из единственного протона) развалилось бы на отдельные составляющие, так как все протоны имеют один и тот же положительный заряд и они отталкиваются друг от друга. Это означает, что должна быть и другая, более мощная сила, обеспечивающая целостность ядра, и она называется *сильным* (ядерным) взаимодействием. Кроме того, существует и так называемое *слабое* взаимодействие, особенно важное в контексте ядерного распада, но это не основной компонент силы, действующей между нуклонами. Мы поговорим о слабом взаимодействии позже.

Сильному взаимодействию подвержены не все частицы — например, в нем не участвуют электроны. Однако те частицы, которые подчиняются этой силе, сами по себе сравнительно массивны и называются *адронами* (от греческого слова ἁδρός, означающего «крупный»). Итак, протоны и нейтроны являются адронами, но сегодня известны и многие другие разновидности адронов. Среди них есть родственники протонов и нейтронов, именуемые *барионами* (от греческого слова βαρύς, означающего «*тяжелый*»), в число которых, кроме самих нейтронов и протонов, входят лямбда (Λ), сигма (Σ), кси (Ξ), дельта (Δ) и омега (Ω). Большинство адронов бывают нескольких сортов с различными значениями электрического заряда и существуют в виде последовательности возбужденных (более быстро вращающихся) вариантов. Все эти частицы массивнее протона и нейтрона. Они не встречаются в обычных атомах, поскольку очень нестабильны и быстро распадаются, в конечном итоге порождая протоны и нейтроны и избавляясь от избыточной массы в виде энергии (в соответствии со знаменитым эйнштейновским $E = mc^2$). Протон, в свою очередь, весит примерно столько же, сколько 1836 электронов, а нейтрон в 1839 раз тяжелее электрона. Промежуточное положение между барионами и электронами занимают *мезоны* — это еще один класс адронов, наиболее известными из них являются пион (π) и каон (K). Оба они существуют в трех вариантах: с положительным и отрицательным зарядом (π^+ и π^-, каждый массой примерно 273 электрона; K^+ и $\overline{K}^-$, каждый массой примерно 966 электронов), а также без заряда (масса π^0 составляет примерно 264 электрона, масса K^0 и $\overline{K}^0$ составляет примерно 974 массы электрона). Если над символом частицы ставится черточка, то он обозначает античастицу; однако отметим, что антипион эквивалентен пиону, а каон отличается от антикаона. Опять же, у этих частиц могут быть возбужденные аналоги, вращающиеся быстрее.

Возможно, все это покажется вам очень сложным — далекий отголосок тех головокружительных времен начала XX века, когда казалось, что протон, нейтрон, электрон (и еще пара безмассовых частиц, например фотон — частица

света) — вот, в общем, и все элементарные частицы. Шли годы, ситуация все более усложнялась, пока наконец не оформилась общая картина, названная *стандартной моделью физики частиц* [Zee, 2010; Thomson, 2013], это произошло между 1970 и 1973 годами. Согласно этой модели, все адроны состоят из *кварков* и/или их античастиц, именуемых *антикварками*. Сегодня считается, что любой барион состоит из трех кварков, а любой (обычный) мезон — из кварка и антикварка. Кварки могут иметь один из шести различных *ароматов*, которые называются (довольно странно и нелогично) *верхний, нижний, очарованный, странный, истинный и красивый*. Ароматы имеют соответственно следующие электрические заряды: $\frac{2}{3}, -\frac{1}{3}, \frac{2}{3}, -\frac{1}{3}, \frac{2}{3}, -\frac{1}{3}$.

На первый взгляд дробные значения заряда кажутся особенно странными, но общий электрический заряд наблюдаемых свободных частиц (например, барионов и мезонов) всегда выражается целым числом.

Стандартная модель не только систематизирует кажущуюся неразбериху среди элементарных частиц, но и хорошо описывает важнейшие силы, действующие на них. Как сильное, так и слабое взаимодействие трактуется в контексте красивой математической процедуры — так называемой *калибровочной теории*, важнейшую роль в которой играет *расслоение* (краткое описание расслоения приводится в разделе A.7, и к нему я буду возвращаться, в частности, в разделе 1.8). *Базовое пространство* $\mathcal{M}$, на котором строится расслоение (об этом см. раздел A.7), — это пространство-время, а в случае сильного взаимодействия (этот пример более прозрачен с математической точки зрения) слой $\mathcal{F}$ описывается в терминах так называемого *цвета*, присваиваемого отдельным кваркам (конкретный кварк может иметь разные цвета). Соответственно, теория физики сильных взаимодействий называется *квантовой хромодинамикой* (КХД). Я не хочу вдаваться в подробное обсуждение квантовой хромодинамики, поскольку для этого потребовалось бы больше математики, чем я готов здесь приводить (см. [Tsou and Chan, 1993; Zee, 2003]). Более того, эта теория «немодна» в том смысле, который я использую в этой книге, поскольку при всей кажущейся экзотичности и странности ее идей она работает очень хорошо и не только образует согласованный и крепко связанный математический аппарат, но и замечательно подтверждается экспериментами. Квантовая хромодинамика изучается на любом физическом факультете, где серьезно рассматривается теория сильных взаимодействий, но она просто «немодна» в принятом здесь смысле, а широко изучается потому, что для этого есть серьезные научные основания!

Однако, несмотря на все достоинства стандартной модели, существуют серьезные научные причины стремиться выйти за ее пределы. Одна из таких при-

чин заключается в том, что стандартная модель включает около 30 числовых параметров, которых эта теория вообще не объясняет. Таковы, в частности, массы кварков и лептонов, величины, именуемые *параметрами смешивания фермионов* (например, угол Кабиббо, угол Вайнберга, тета-угол, калибровочные связности и параметры, относящиеся к механизму Хиггса). С этой проблемой связан и другой серьезный недостаток, проявлявшийся и в других теориях, существовавших еще до появления стандартной модели и разрешенный в них лишь отчасти. Это неприятная проблема *бесконечностей* (бесконечности — это бессмысленные ответы, получаемые из расходящихся выражений, например показанные в разделе A.10), возникающих в квантовой теории поля (КТП). КТП — это разновидность квантовой механики, играющая центральную роль не только в КХД и других приложениях стандартной модели, но и во всех современных подходах к физике частиц, а также во многих других аспектах фундаментальной физики.

Мне все еще требуется немало рассказать о квантовой механике, но я займусь этим в главе 2. А пока давайте обратим внимание только на одну специфическую, но фундаментальную черту квантовой механики. Эту черту можно рассматривать как корень проблемы бесконечностей в КТП. Мы также увидим, как традиционный метод обращения с этими бесконечностями не позволяет в сколько-нибудь полной мере решить проблему вывода тридцати с чем-то необъяснимых чисел в составе стандартной модели. Теория струн во многом выросла из хитроумной попытки как-нибудь дистанцироваться от бесконечностей КТП, о чем мы поговорим в разделе 1.6. Следовательно, она дает некоторую надежду на то, что мы найдем путь к разгадке тайны этих необъяснимых чисел.

1.4. Принцип суперпозиции в КТП

Краеугольным камнем квантовой механики является принцип суперпозиции, фигурирующий во *всех* вариантах квантовой теории, а не только в КТП. В частности, он имеет ключевое значение в тех критических обсуждениях, которые изложены в главе 2. В данной главе, чтобы дать вам представление об источнике проблемы с бесконечностями в КТП, я вкратце познакомлю вас с этим принципом, однако речь о квантовой механике в основном пойдет в главе 2 (см., в частности, разделы 2.5 и 2.7).

Для того чтобы подчеркнуть роль принципа суперпозиции в КТП, рассмотрим ситуации следующего плана. Допустим, есть физический процесс, приводящий к конкретному наблюдаемому результату. Предположим, этот результат возникает благодаря некоторому промежуточному действию Ψ, но возможно

и другое промежуточное действие **Ф**, которое может привести, в сущности, к такому же наблюдаемому результату. Затем в соответствии с принципом суперпозиции мы должны признать, что, в известном смысле, *оба действия*, и **Ψ**, и **Ф**, вполне могли произойти на промежуточном этапе *параллельно*! Разумеется, это не слишком логично, поскольку в обычных макроскопических масштабах мы не сталкиваемся со случаями, в которых две явно разные возможности реализовывались бы одновременно. Однако в случае с субмикроскопическими событиями, когда мы не можем наблюдать, какое именно промежуточное действие имело место — первое или второе, мы должны допустить, что оба действия могли произойти одновременно и находились при этом в так называемой *квантовой суперпозиции*.

Архетипический пример такого явления — знаменитый эксперимент с двумя щелями, который часто описывают с целью дать представление о квантовой механике. Рассматривается ситуация, в которой пучок квантовых частиц (скажем, электронов или фотонов) направляется на экран и при этом должен пройти через две расположенные одна рядом с другой параллельные щели (рис. 1.2 *а*). В рассматриваемой ситуации каждая частица, достигнув экрана, оставляет четкий темный след в этой точке, что указывает на *корпускулярную* природу частицы. Но после того как на экран попадет множество частиц, выстраивается интерференционная картина, состоящая из светлых и темных полос. Темные полосы соответствуют месту, где на экран попало сравнительно много частиц, а светлые — месту, куда их попало относительно мало (рис. 1.2 *г*). Стандартный анализ[1] ситуации позволяет заключить, что каждая отдельная квантовая частица должна в некотором смысле пройти сразу через *обе* щели; при этом два альтернативных возможных маршрута странным образом оказываются в суперпозиции.

Вот каковы рассуждения, приводящие нас к столь странному выводу: если любая из щелей закрыта, но вторая при этом открыта (рис. 1.2 *б, в*), то не возникает никаких полос, просто достаточно однородное распределение, причем центральная часть экрана кажется наиболее бедной в распределении точек. Однако когда обе щели открыты, на экране видны более темные области, расположенные между

[1] Здесь имеется в виду традиционный анализ ситуации. Логично, что из-за такого странного на первый взгляд вывода предлагаются и другие варианты интерпретации событий, происходящих на этом промежуточном этапе существования частицы. Наиболее примечательная альтернативная трактовка излагается в так называемой теории де Бройля — Бома, согласно которой сама частица всегда проходит либо через одну щель, либо через другую, но ей сопутствует «волна-пилот», которая ведет частицу и должна *сама* «прозондировать» оба варианта, один из которых могла бы выбрать частица (см. [Bohm and Hiley, 1993]). Я вернусь к этой точке зрения в разделе. 2.12.

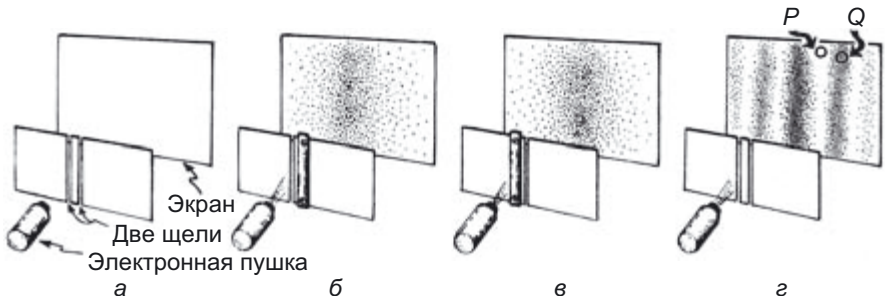


Рис. 1.2. Эксперимент с двумя щелями. Электроны направляются на экран через две щели, расположенные рядом (*а*). Если открыта всего одна щель (*б*, *в*), то на экране возникает множество точек, хаотично разбросанных на той площади, куда направлена траектория пучка. Однако если открыты обе щели (*г*), то получается «чересполосица». На экране появляются такие места (например, *P*), куда частица попасть не может, но попала бы, если бы была открыта всего одна щель. Более того, в другие места (например, *Q*) попадает вчетверо больше частиц, чем когда открыта одна щель

светлыми полосами. Эти темные области возникают на участках, которые были бы совершенно светлыми, если бы была открыта всего одна щель. Каким-то образом, когда частице открыты оба пути, попадание частиц в более темные области *блокируется*, а в более светлые — *активизируется*. Если бы каждая частица просто могла пойти по одному из маршрутов, доступных, если открыта либо одна щель, либо другая, то эффекты от прохождения этих маршрутов просто суммировались бы, и у нас бы не было этих странных интерференционных полос. Происходящее объясняется лишь тем, что частице открыты оба маршрута, она «прощупывает» оба этих маршрута, и в результате мы имеем итоговый эффект. В определенном смысле оба эти маршрута *сосуществуют* для частицы, когда она находится между источником и экраном.

Разумеется, это противоречит поведению макроскопических тел, к которому мы привыкли. Например, если из одной комнаты в другую ведут две разные двери и мы заметили, что кошка пошла через одну комнату и оказалась в другой, то, конечно же, мы решим, что она прошла либо через одну дверь, либо через другую, но не через обе двери одновременно. То есть если объект сравним по размеру с кошкой, то мы могли бы постоянно отслеживать его местоположение, не вмешиваясь в его действия так или иначе, и, таким образом, получили бы точные данные о том, через какую именно из двух дверей он прошел. Если бы нам удалось осуществить подобное хотя бы с одной квантовой частицей, участвующей в вышеописанном эксперименте с двумя щелями, то мы бы так сильно повлияли на ее поведение, что весь интерференционный узор на экране исчез. Волновое поведение отдельно взятой квантовой частицы, в результате

которого на экране возникают светлые и темные интерференционные полосы, предполагает, что мы *не сможем* удостовериться, через которую из щелей она прошла. Именно поэтому может возникать такое загадочное промежуточное состояние частицы — суперпозиция.

В эксперименте с двумя щелями исключительно странное поведение одиночных квантовых частиц особенно хорошо заметно, если смотреть в точку *P*, расположенную на экране прямо посередине между двумя темными полосами. Мы заметим, что частица просто не может попасть в точку *P*, когда обе щели открыты, но если открыта всего одна щель, то частица довольно легко попадает через нее в *P*. Когда открыты обе щели и частица может достичь точки *P* по одному из двух маршрутов, то эти возможности каким-то образом «обнуляют» друг друга. Однако в другой точке экрана, скажем *Q*, где интерференционный узор особенно темный, оба доступных маршрута словно усиливают друг друга, когда открыты обе щели (см. рис. 1.2 *г*). Вероятность того, что частица попадет в точку *Q*, вчетверо выше при двух открытых щелях, чем если бы была открыта всего одна щель, — не вдвое, как это было бы не с квантовой частицей, а с обычным объектом, подчиняющимся законам классической физики, а вчетверо! Эти странные свойства являются следствием так называемого *правила Борна*, соотносящего интенсивность с фактической вероятностью случая, — об этом мы вскоре поговорим.

Кстати, слово «*классический*» в контексте физических теорий, моделей или ситуаций означает просто «*неквантовый*». В частности, общая теория относительности Эйнштейна является классической, несмотря на то что именно в рамках этой теории были сформулированы многие ключевые идеи квантовой теории (например, атом Бора). Другими словами, классические системы *не* подвергаются странным суперпозициям альтернативных возможностей, которые мы рассмотрели выше и которые действительно характерны для квантового поведения, но об этом чуть ниже.

Я вынужден отложить до главы 2 подробную дискуссию об основах нашего современного понимания квантовой физики (см., в частности, раздел 2.3 и далее). А пока рекомендую вам просто взять на заметку странное математическое правило, согласно которому современная квантовая механика описывает промежуточные состояния. Но что это за странное правило? Квантовый формализм постулирует, что подобное промежуточное состояние в суперпозиции, где имеются всего две альтернативные возможности, $\boldsymbol{\Psi}$ и $\boldsymbol{\Phi}$, должно математически выражаться в виде своеобразной суммы двух возможностей $\boldsymbol{\Psi} + \boldsymbol{\Phi}$ или, в более общем виде, линейной комбинации (см. разделы A4 и A5).

$$w\boldsymbol{\Psi} + z\boldsymbol{\Phi},$$

где w и z — *комплексные числа* (числа, содержащие $i = \sqrt{-1}$, как описано в разделе А.9), причем оба они не равны нулю! Более того, далее мы будем вынуждены принять, что такие квантовые суперпозиции должны иметь возможность *сохраняться* в квантовой системе вплоть до момента, когда состоится акт наблюдения за этой системой, и в этот момент суперпозиция альтернатив должна будет смениться смесью вероятностей этих альтернатив. Это действительно странно, но в разделах 2.5–2.7 и 2.9 мы увидим, как использовать эти комплексные числа (иногда именуемые амплитудами) — и как они удивительным образом связываются с вероятностями, а также с эволюцией физических систем на квантовом уровне с течением времени (уравнение Шрёдингера); кроме того, на фундаментальном уровне они связаны с тонкостями поведения спина квантовой частицы и даже с трехмерностью обычного физического пространства! Хотя связи между этими амплитудами и вероятностями (правило Борна) не будут подробно рассматриваться в этой главе (так как для этого потребовалось бы ввести понятия *ортогональность* и *нормирование* для $\boldsymbol{\Psi}$ и $\boldsymbol{\Phi}$, а этот материал лучше отложить до раздела 2.8), суть правила Борна можно выразить следующим образом.

> Измерение, призванное определить, находится система в состоянии $\boldsymbol{\Psi}$ или в состоянии $\boldsymbol{\Phi}$, при наличии суперпозиции $w\boldsymbol{\Psi} + z\boldsymbol{\Phi}$ дает:
>
> отношение вероятности $\boldsymbol{\Psi}$ к вероятности $\boldsymbol{\Phi}$ = отношение $|w|^2$ к $|z|^2$.

Отметим (см. разделы А.9 и А.10), что квадрат модуля $|z|^2$ комплексного числа z равен сумме квадратов его вещественной и мнимой частей и, следовательно, является квадратом расстояния от начала координат в комплексной плоскости до z (см. рис. А.42 в разделе А.10). Также можно отметить, что факт возникновения вероятностей на основе *квадратов* модулей этих амплитуд объясняет упомянутое выше *четырехкратное* увеличение интенсивности, поэтому вклад обеих щелей в общую картину при эксперименте взаимно усиливается (см. также конец раздела 2.6).

Здесь нужно быть внимательным, чтобы понять, что знак «*плюс*» в этих суперпозициях сильно отличается от обычного понятия «*и*» (несмотря на то что в обычных рассуждениях «*плюс*» и «*и*» обычно понимаются как одно и то же) и даже от «*или*». Здесь имеется в виду, что в некотором смысле обе возможности словно *складываются* друг с другом абстрактным математическим образом. Таким образом, если говорить об эксперименте с двумя щелями, где $\boldsymbol{\Psi}$ и $\boldsymbol{\Phi}$ соответствуют двум промежуточным местоположениям одной и той же частицы, выражение $\boldsymbol{\Psi} + \boldsymbol{\Phi}$ не описывает *две* частицы, по одной в каждой точке (то есть речь не идет о ситуации «одна частица в точке $\boldsymbol{\Psi}$ и одна частица в точке $\boldsymbol{\Phi}$ — всего две частицы»). Также не следует считать две эти возможности обычными

альтернативами, из которых произойдет одна *или* другая, но мы не знаем, какая именно. Мы действительно должны полагать, что это одна частица каким-то образом находится в двух местах сразу, то есть существует в суперпозиции благодаря странной операции квантово-механического «сложения». Разумеется, это кажется очень странным, и физики начала XX века не стали бы рассматривать такую возможность, если бы у них не было на то очень серьезных причин. В главе 2 мы исследуем некоторые из этих причин, а пока просто прошу читателя поверить, что такой формализм действительно работает.

Важно понять, что, согласно стандартной квантовой механике, такая процедура суперпозиции считается *универсальной*, а значит, интерпретируется так, как будто у промежуточного состояния существует не две альтернативы, а больше. Например, если бы у нас были три альтернативные возможности: $\boldsymbol{\Psi}$, $\boldsymbol{\Phi}$ и $\boldsymbol{\Gamma}$, то мы бы рассматривали тройные суперпозиции вида $w\boldsymbol{\Psi} + z\boldsymbol{\Phi} + u\boldsymbol{\Gamma}$ (где w, z и u — комплексные числа, среди которых есть хотя бы одно ненулевое). Соответственно, если бы существовало четыре альтернативных состояния, то мы должны были бы рассмотреть четверные суперпозиции, и т. д. Квантовая механика требует этого, и такое проявление квантовых механизмов на субмикроскопическом уровне отлично подтверждается экспериментально. Да, это действительно странно, но с математической точки зрения здесь все хорошо и непротиворечиво. До сих пор речь шла всего лишь о математике *векторного пространства* с комплексными скалярами, рассматриваемой в разделах A.3, A.4, A.9 и A.10, а в разделе 2.3 и далее мы увидим, что квантовые суперпозиции — явление более универсальное. Однако в квантовой теории поля все значительно усложняется, поскольку нам часто приходится рассматривать ситуации, в которых присутствует бесконечное множество промежуточных возможностей. Соответственно, мы начинаем рассматривать бесконечные суммы альтернатив, и проблема приобретает угрожающие очертания: речь идет о том, что такая бесконечная сумма действительно может дать нам последовательность членов, сумма которых может *расходиться* до бесконечности (об этом рассказано в разделах A.10 и A.11).

1.5. Сила фейнмановских диаграмм

Давайте попробуем разобраться, как возникают такие расходимости. В физике частиц приходится рассматривать ситуации, в которых несколько частиц сталкиваются и образуют другие частицы, причем некоторые из них могут распасться и породить третьи частицы, а пары этих частиц могут снова слиться, и т. д. и т. п., так что подобные процессы могут сильно усложняться. Как правило, специалисты по физике частиц имеют дело с такими ситуациями: существует некоторое множество сталкивающихся частиц — часто это происходит на око-

лосветовых скоростях, — и такая комбинация столкновений и распадов дает на выходе какое-то новое множество частиц. Весь процесс будет связан с обширной квантовой суперпозицией всевозможных промежуточных процессов, которые могут произойти и не противоречат исходным и конечным условиям. Пример подобного сложного процесса показан на *фейнмановской диаграмме* на рис. 1.3.

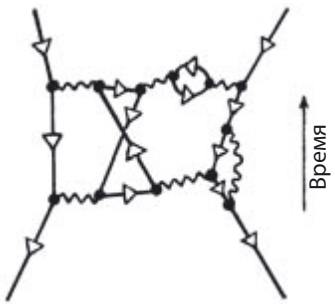


Рис. 1.3. Диаграмма Фейнмана (стрелка, указывающая вверх, обозначает направление течения времени) — это схематическое представление пространства-времени (с четкой математической интерпретацией), описывающее процесс на уровне частиц, зачастую сопряженный с возникновением частиц, аннигиляцией и обменом промежуточными частицами. Волнистой линией обозначены фотоны. Треугольная стрелка обозначает электрический заряд (положительный — если стрелка указывает вверх, и отрицательный — если вниз)

Мы не сильно погрешим против истины, если скажем, что фейнмановская диаграмма — это пространственно-временная схема конкретных процессов, происходящих с определенным множеством частиц. Мне нравится обозначать время стрелкой, направленной вверх, так как я работаю в области теории относительности, но не являюсь профессионалом в физике частиц или КТП; настоящие профессионалы обычно изображают ход времени слева направо. Фейнмановские диаграммы (или графы) названы в честь выдающегося американского физика Ричарда Фейнмана. Простейшие диаграммы такого рода показаны на рис. 1.4. На рис. 1.4 *а* продемонстрирован распад частицы на две, а на рис. 1.4 *б* — слияние двух частиц с образованием третьей.

На рис. 1.4 *в* показан обмен частицей (например, фотоном, квантом электромагнитного поля, или света, обозначенным волнистой линией) между двумя другими частицами. В этом процессе используется термин *обмен*, и хотя он широко известен в физике частиц, здесь он несет немного необычное значение, так как всего *один* фотон переходит от одной частицы к другой, хотя переходит таким образом, что (и это неслучайно) нельзя сказать, какая частица является

донором, а какая — акцептором. Фотон, фигурирующий в таком обмене, обычно называется *виртуальным*, а его скорость не ограничена требованиями теории относительности. Термин «*обмен*» в общеупотребительном значении применим скорее в ситуации, показанной на рис.1.5 *б*, где происходит обмен двумя фотонами.

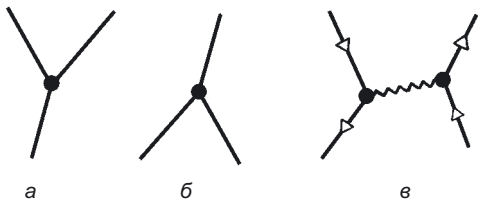


Рис. 1.4. Простейшие фейнмановские диаграммы: *а*) частица распадается на две; *б*) две частицы сливаются, образуя третью; в) три частицы с противоположными зарядами (например, электрон и позитрон) «обмениваются» фотоном

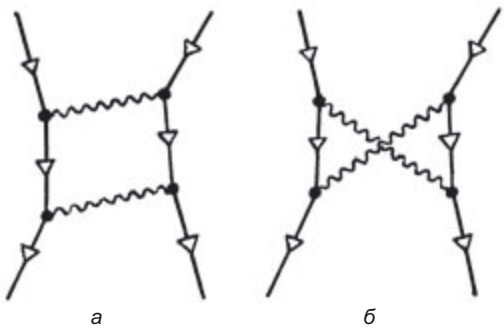


Рис. 1.5. Обмен двумя фотонами

Можно сказать, что целостная фейнмановская диаграмма состоит из множества таких универсальных элементов, которые сочетаются во всевозможных комбинациях. Однако принцип суперпозиции не позволяет считать, что столкновение двух частиц — это акт, который можно запечатлеть всего на одной фейнмановской диаграмме, поскольку одновременно существует множество альтернатив, и истинный физический процесс представляет собой сложную линейную суперпозицию множества таких фейнмановских диаграмм. По каждой конкретной фейнмановской диаграмме требуется рассчитать, в какой степени общая суперпозиция — в сущности, комплексное число вроде w и z, с которыми мы сталкивались в разделе 1.4, — отражается на результате. Эти числа называются *комплексными амплитудами* (см. разделы 1.4 и 2.5).

Однако необходимо помнить, что одно лишь расположение связей на диаграмме не дает исчерпывающей информации. Нам также требуется знать значения энергий и импульсов всех частиц, вовлеченных в процесс. В случае с внешними частицами (как входящими, так и исходящими) эти значения уже бывают присвоены и их можно просто взять, но энергии и импульсы промежуточных частиц (также называемых внутренними) обычно могут принимать самые разные значения согласно следующему ограничению: энергия и импульс суммируются на каждой вершине, причем импульс обычной частицы вычисляется как произведение скорости на массу (см. разделы. A.4 и A.6). (Импульс обладает важным свойством — *консервативностью*, или более простым языком: он сохраняется в замкнутой системе. Таким образом, при любом столкновении между частицами общий импульс входящих частиц должен быть равен общему импульсу исходящих.) Следовательно, какими бы сложными ни казались суперпозиции, хотя бы на примере все более сложных последовательных диаграмм, возникающих в процессе, на самом деле все *гораздо* сложнее, поскольку обычно в каждой диаграмме внутренние частицы могут принимать *бесконечное* множество значений энергии и импульса (не противоречащих заданным значениям энергии и импульса, которыми обладают внешние частицы).

Итак, даже если мы имеем дело всего с одной фейнмановской диаграммой и знаем исходные и конечные значения, нам, вполне возможно, придется просуммировать бесконечное число таких процессов. Технически такое сложение принимает вид непрерывного интеграла, а не дискретной суммы (см. разделы A.7, A.11 и рис. A.44), но в данный момент такие различия для нас несущественны. Подобное происходит с фейнмановскими диаграммами, содержащими *замкнутую петлю*, такими, например, какие представлены на рис. 1.5. На *древовидной диаграмме* (см., например, рис. 1.4 и 1.6) замкнутых петель нет, и значения

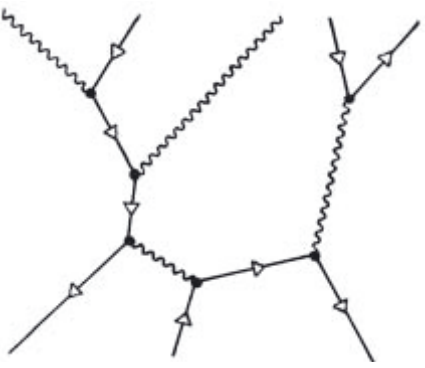


Рис. 1.6. Древовидная диаграмма, которая не содержит замкнутых петель

внутренних энергий и импульсов жестко определяются входящими и выходящими параметрами. Однако такие древовидные диаграммы не позволяют постичь подлинно квантовой природы взаимодействий между частицами; для этого на схему нужно добавить замкнутые петли. При этом вся сложность с замкнутыми петлями заключается в том, что по такой петле может циркулировать фактически неограниченная энергия с неограниченным импульсом, а при сложении всего этого мы получаем *расходимость*.

Рассмотрим этот случай подробнее. Одна из простейших ситуаций, в которой возникает замкнутая петля, показана на рис. 1.5 *а*, где изображен обмен двумя частицами. Проблема возникает из-за того, что, хотя на каждой вершине диаграммы значения энергии и трех компонентов импульса должны суммироваться точно (то есть сумма исходных значений равна сумме конечных), здесь нет достаточного числа уравнений, чтобы задать внутренние значения этих величин. (Для каждого из четырех компонентов энергии-импульса в отдельности существуют три независимых уравнения, поскольку на каждой из четырех вершин имеем уравнение сохранения, но одно из них избыточно и просто еще раз выражает сохранение энергии и импульса для всего процесса. Однако на каждый компонент приходится по четыре неизвестных, одно от каждой внутренней линии. Поэтому уравнений оказывается недостаточно, чтобы определить все неизвестные.) Мы всегда можем приплюсовать (или вычесть) одну и ту же величину энергии-импульса, пока идет обход петли. Приходится складывать это бесконечное множество возможностей, вовлекая в вычисления потенциально все более высокие значения энергии-импульса; это и приводит к потенциальной расходимости.

Итак, мы видим, что при непосредственном применении квантовых законов мы, вероятно, получим расходимость. Однако это еще совсем не означает, что «верный» ответ таких вычислений в рамках квантовой теории поля действительно равен бесконечности. Было бы полезно напомнить о расходящихся рядах, рассматриваемых в разделе А.10, где в некоторых случаях сумме ряда можно присвоить конечное значение, несмотря на то что простое сложение его членов дает в результате бесконечность. Хотя в КТП ситуация несколько иная, между двумя примерами есть некоторое явное сходство. За годы исследований специалисты по КТП разработали множество вычислительных методов, позволяющих обойти такие бесконечные ответы. Если подойти к делу с умом, как это сделано в примерах из раздела А.10, можно откопать «истинный» конечный ответ, который мы не получили бы простым сложением членов. Соответственно, мастерам КТП зачастую удается выжать конечные ответы из безумно расходящихся выражений, с которыми приходится иметь дело, хотя многие применяемые при этом процедуры отнюдь не столь прямолинейны, как простое аналитическое продолжение, — об этом методе мы поговорим в разделе А.10 (см. также раз-

дел 3.10 с описанием некоторых любопытных подводных камней, на которые можно наткнуться, придерживаясь даже «прямолинейных» процедур).

Здесь следует упомянуть один ключевой момент, связанный с коренной причиной многих таких расходимостей, именуемых *ультрафиолетовыми*. В принципе, проблема состоит в том, что в замкнутой петле ничто не ограничивает масштабов энергии и импульса, передающихся по этой петле, и расходимость возникает из-за того, что требуется суммировать все более и более высокие значения энергии (и импульса). В то же время, согласно квантовой механике, очень высокие энергии могут существовать только в течение очень коротких промежутков времени. Этот принцип исходно описывается знаменитой формулой Макса Планка $E = h\nu$, где E — энергия, ν — частота, а h — постоянная Планка. Итак, высокие значения энергии соответствуют высоким частотам, и, соответственно, интервалы времени между импульсами получаются крошечными. Аналогично, очень сильные мпульсы действуют лишь на минимальных расстояниях. Если предположить, что на минимальных расстояниях и в минимальные промежутки времени с пространством-временем творится что-то странное (действительно, большинство физиков согласились бы, что без «если» не обходятся никакие рассуждения о квантовой гравитации), то в верхнем конце шкалы значений энергии и имплься может происходить некое фактическое обрезание допустимых значений этих величин. Соответственно, в будущем может появиться какая-нибудь новая теория о структуре пространства-времени, и если она будет описывать резкие изменения, происходящие в кратчайшие промежутки времени и на минимальных расстояниях, то эта теория, возможно, действительно поможет получить конечные результаты для вычислений КТП, которые сейчас считаются расходящимися и возникают из-за замкнутых петель в фейнмановских диаграммах. Такие времена и расстояния должны быть гораздо короче, чем те, с которыми мы имеем дело на уровне обычных процессов физики частиц, причем зачастую считается, что для теории квантовой гравитации будут релевантны величины именно такого порядка, например *планковское время* (около 10^{-43} с) или планковская длина (около 10^{-35} м, о ней мы упоминали в разделе 1.1). Итак, эти значения будут примерно в 10^{20} меньше тех величин, которыми сегодня оперирует физика элементарных частиц.

Здесь также следует отметить, что в КТП выделяются и так называемые *инфракрасные расходимости*. Они возникают на противоположном конце спектра при экстремально малых значениях энергии и импульса, проявляющихся на колоссальных расстояниях и в течение очень длительных периодов времени. Здесь проблемы возникают не с замкнутыми петлями, а с фейнмановскими диаграммами, что показаны на рис. 1.7. В данном случае при рассматриваемом процессе может излучаться неограниченное количество *мягких фотонов* (то есть

фотонов с крайне малыми энергиями), и при суммировании всех этих значений вновь возникает бесконечность. Обычно специалисты по КТП считают инфракрасные расходимости не столь серьезными, как ультрафиолетовые, и всячески стараются от них отмахнуться (как минимум временно). Однако в последнее время появляются данные, заставляющие внимательнее отнестись к инфракрасным расходимостям. Здесь, обсуждая интересующие меня темы, я не буду заострять внимание на вопросах инфракрасных расходимостей и подробнее остановлюсь на том, как в стандартной КТП решается проблема ультрафиолетовых расходимостей, возникающих в замкнутых петлях на фейнмановских диаграммах, и каким образом идеи теории струн позволяют надеяться, что мы сможем разобраться с этой головоломкой.

Рис. 1.7. Инфракрасные расходимости возникают при излучении бесконечного множества «мягких» фотонов

В данном контексте следует отдельно рассказать о стандартной процедуре перенормировки, принятой в КТП. Давайте вкратце рассмотрим, как она происходит. Согласно различным прямым расчетам КТП, мы получаем бесконечный масштабный фактор между так называемым *голым зарядом* частицы (например, электрона) и *одетым зарядом*, причем именно последний измеряется при проведении экспериментов. Так происходит из-за постепенного добавления новых величин в ходе таких процессов, как показан на фейнмановской диаграмме на рис. 1.8, и эти величины снижают измеряемую величину заряда. Проблема в том, что на рис. 1.8 (и на многих других подобных диаграммах) приплюсовывается бесконечность (на этой диаграмме есть замкнутые петли). Соответственно, оказывается, что голый заряд должен быть бесконечным, чтобы можно было найти конечное значение наблюдаемого (одетого) заряда. Базовая философия процедуры перенормировки заключается в необходимости признать следующее: КТП может быть не вполне точна на сверхмалых расстояниях, где как раз и возникают расходимости, и если внести в теорию пока неизвестные уточнения, то, возможно, они помогут обеспечить необходимое обрезание, которое приведет нас к конечным результатам. Следовательно, такая процедура требует, чтобы мы отказались от попыток вычислить истинные природные значения этих масштабных факторов (для заряда и других величин, например массы и т. д.), а вместо этого собрали все подобные

бесконечные масштабные факторы, которыми нас обременяет КТП, фактически упаковали эти бесконечные величины в красивые маленькие конвертики и так и продолжали их игнорировать, подбирая значения голого заряда (а также массы и т. д.) так, чтобы получить значения, получаемые в эксперименте. Довольно примечательно, что в соответствующих КТП, именуемых *перенормируемыми*, это можно делать систематически, получая таким образом конечные ответы и для многих других вычислений КТП. Значения таких величин, как одетый заряд (масса и т. д.), узнаются из наблюдений, а не рассчитываются в рамках соответствующей КТП, и дают нам около 30 параметров, которые вносятся в стандартную модель на основе экспериментов, как было указано выше.

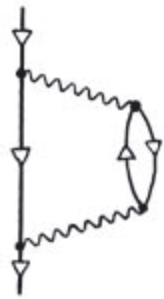


Рис. 1.8. Для работы с такими расходящимися диаграммами применяется процедура перенормировки заряда

Прибегая к подобным процедурам, иногда удается получить и исключительно точные значения, не относящиеся к КТП. Так, *магнитный момент* электрона определяют путем КТП-вычисления, которое уже стало стандартным. Большинство частиц ведут себя как маленькие магниты (помимо того, что некоторые частицы еще и несут электрический заряд), и магнитный момент подобной частицы характеризует, насколько силен этот магнит. Дирак впервые предвычислил магнитный момент электрона, получив его непосредственно из своего фундаментального уравнения для электрона (о нем мы вкратце упоминали в разделе 1.1), и предугаданное им значение практически идеально подтвердилось в ходе экспериментов. Однако оказывается, что к этому значению можно добавить поправки, обусловленные косвенными КТП-процессами, и эти поправки должны учитываться при расчете магнитного момента одиночного электрона. Поправка, даваемая квантовой теорией поля, составляет 1,001159652... от «чистого» исходного дираковского значения. Экспериментально получаем значение поправки 1,00115965218073... [Hanneke et al., 2011]. Соответствие поистине невероятное! По замечанию Ричарда Фейнмана, это все равно что определить

расстояние от Нью-Йорка до Лос-Анджелеса с точностью до толщины человеческого волоса [Feynman, 1985]. Этот результат замечательно подтверждает *перенормируемую* КТП-теорию электронов и фотонов (так называемую *квантовую электродинамику*, или КЭД), где электроны описываются теорией Дирака, а фотоны — электромагнитными уравнениями Максвелла (см. раздел 1.2) и где их взаимодействие согласуется со стандартным уравнением Лоренца, описывающим реакцию заряженной частицы на воздействие электромагнитного поля. Последнее в квантовом контексте следует из *калибровочных процедур* Германа Вейля (см. раздел 1.7). Итак, мы видим, что теория и эксперимент действительно согласуются в чрезвычайно высокой степени, а это означает, что в данной теории содержится некая глубокая истина, пусть с математической точки зрения она и не является строго непротиворечивой.

Перенормировку можно считать временным решением и продолжать надеяться, что однажды будет открыта некая более точная версия квантовой теории поля, где такие бесконечности вообще не будут возникать и для этих масштабных факторов удастся рассчитать не просто конечные значения, но и фактические «голые» значения — а следовательно, и экспериментально наблюдаемые значения заряда, массы и т. п. для различных элементарных частиц. Несомненно, активная разработка теории струн во многом объясняется как раз надеждой на то, что именно эта теория могла бы привести нас к такой «улучшенной КТП». Но существует и более скромный метод, до сих пор, несомненно, более успешный, чем теория струн, — просто проверять, применима ли процедура перенормировки в рамках данной теории, и по этому критерию выбирать наиболее многообещающие варианты из общего множества различных КТП. Оказывается, лишь *некоторые* варианты КТП поддаются процедуре перенормировки (выше упоминалось, что они называются перенормируемыми), а другие не поддаются. Итак, перенормировка считается мощным признаком для отбора наиболее многообещающих КТП. На самом деле удалось открыть (здесь следует особо отметить работы Герарда 't Хоофта начиная с 1971 года и более поздние ['t Hooft, 1971; 't Hooft and Veltman, 1972]), что использование такой симметрии, какую требуют калибровочные теории, упомянутые в разделе 1.3, помогает формулировать перенормируемые КТП. Этот факт стал мощным стимулом для формулировки стандартной модели.

1.6. Исходные ключевые идеи теории струн

Давайте рассмотрим, как во всю эту картину вписываются исходные идеи теории струн. Как упоминалось выше, проблемы ультрафиолетовых расходимостей обусловлены квантовыми процессами, происходящими на очень малых расстояниях и за очень короткие промежутки времени. Можно считать, что корень

проблемы таков: мы полагаем, что материальные объекты состоят из *частиц* и каждая такая элементарная частица представляет собой *точку* в пространстве. Разумеется, можно понимать точечную природу голой частицы как нереалистичное приближение. Однако если, напротив, считать такой примитивный объект чем-то «размазанным», то мы сталкиваемся с противоположной проблемой: как описать нечто столь разреженное, не беря в расчет, что оно состоит из более мелких компонентов. Кроме того, в подобных моделях всегда существуют и более тонкие проблемы, касающиеся возможных противоречий с теорией относительности (согласно которой скорость распространения информации конечна), возникающих, если некая «размазанная» сущность должна проявлять свойства единого целого.

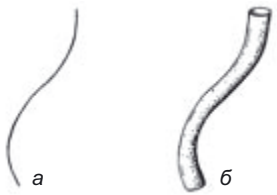


Рис. 1.9. Мировая линия обычной (точечной) частицы — это кривая в пространстве-времени (*а*); в теории струн она превращается в двустороннюю мировую трубку (струнный мировой лист) (*б*)

Теория струн подсказывает свои ответы для решения этой головоломки. Она предполагает, что базовый компонент материи не нульмерный в пространственном отношении, как точечная частица, не трехмерный, как некая распределенная сущность, но одномерный, подобный кривой линии. Хотя эта идея и может показаться странной, следует учитывать, что с точки зрения четырехмерного пространства-времени даже точечная частица в классическом смысле не описывается как обычная точка, поскольку это пространственная точка, сохраняющаяся во времени. Поэтому такое пространственно-временное описание фактически характеризует *одномерное* многообразие (см. раздел А.5), именуемое мировой линией частицы (рис. 1.9 *а*). Соответственно, в теории струн кривая линия должна восприниматься как двумерное многообразие, или *поверхность*, расположенная в пространстве-времени (рис. 1.9 *б*), именуемая струнным *мировым листом*.

По-моему, одна из самых привлекательных черт теории струн (как минимум — ее исходного варианта) заключалась в том, что эти двумерные струнные истории — они же струнные мировые листы — можно было в некотором смысле считать *римановыми поверхностями* (см. раздел 1.9, особенно рис. 1.30, где показана их

связь с поворотом Вика). Риманова поверхность, подробнее описанная в разделе А.10, — это *комплексное* одномерное пространство (с учетом того, что одно измерение, выраженное комплексным числом, соответствует двум измерениям, выраженным вещественными числами). Поэтому в таком комплексном пространстве может проявляться магия комплексных чисел. Римановым поверхностям действительно присущи многие «магические» аспекты. Причем сам факт, что эти поверхности (то есть комплексные кривые) играют роль на уровне, где царят комплексно-линейные законы квантовой механики, открывает возможность для тонкой взаимосвязи, а может, и для гармоничного единства между двумя разными аспектами физики микромира.

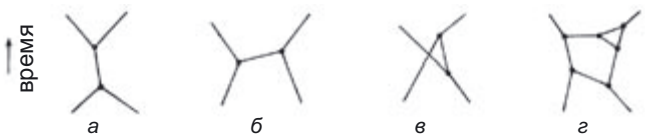


Рис. 1.10. Три разных древовидных графа, где мы имеем две (неуказанные) частицы на входе и две на выходе (*а*, б, *в*), и пример с замкнутыми петлями (*г*)

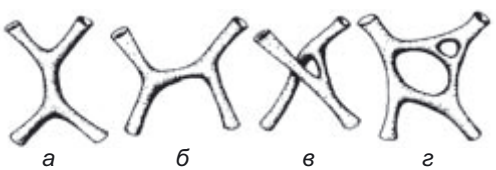


Рис. 1.11. Струнные версии соответствующих процессов, показанных на рис. 1.10

Для того чтобы точнее описать роль, которую должна играть эта фундаментальная струнная идея, вернемся к фейнмановским диаграммам, рассмотренным в разделе 1.5. Если считать, что линии на этих диаграммах соответствуют истинным мировым линиям элементарных частиц, причем эти частицы с топологической точки зрения принципиально напоминают *точки*, то вершины диаграмм соответствуют столкновениям между частицами, и мы можем полагать, что ультрафиолетовые расходимости обусловлены точечной природой этих столкновений. Если, напротив, считать элементарные частицы крошечными петлями, то история такой петли будет похожа на узкую трубку в пространстве-времени. Тогда у нас будут не точечные вершины, как на диаграммах Фейнмана, а трубочки, которые можно аккуратно «приваривать» друг к другу, — задача для хорошего водопроводчика. На рис. 1.10 *а*, б, *в* я изобразил несколько фейнмановских диаграмм (для неуказанных частиц), а на рис.1.10 *г* — более типичную

диаграмму, в которой *присутствуют* замкнутые петли. На рис. 1.11 показано, как могли бы выглядеть их струнные аналоги. Теперь представим, что поверхности с рис. 1.11, каждая из которых представляет собой струнную историю, вместе с их перемычками — это *римановы* поверхности, и мы можем изучать лежащие в их основе физические процессы, опираясь на эту красивую математическую теорию. В частности отметим, что замкнутые петли стандартной фейнмановской теории (из-за которых возникают ультрафиолетовые расходимости) просто обеспечивают многосвязность в топологии римановых поверхностей. Каждая замкнутая петля на диаграмме Фейнмана просто дает нам еще одну «ручку» для топологии римановой поверхности (технически здесь происходит возрастание рода, где *род* римановой поверхности равен числу ее ручек) (см. рис. 1.44 *а* в разделе 1.16 и рис. А.11 в разделе А.5, где показаны примеры топологических ручек).

Также отметим, что исходные и конечные состояния в фейнмановской теории соответствуют дыркам, или проколам, в наших римановых поверхностях, и именно в этих точках можно вводить информацию о таких феноменах, как импульс или энергия. В некоторых научно-популярных описаниях топологии поверхностей термином «*дырка*», или «*отверстие*», именуется как раз то, что я называю здесь «*ручкой*». Но придерживаясь используемой здесь терминологии, можно сказать, что некомпактные римановы поверхности, присутствующие в теории струн, также имеют своеобразное отверстие (или прокол), поэтому нужно выражаться осторожно и не путать два этих очень разных понятия. В разделе 1.16 мы увидим, что отверстия/проколы в римановых поверхностях могут играть и другие роли.

Здесь я должен описать одно раннее обоснование теории струн, о котором я вскользь упоминал в начале раздела 1.3, говоря о некоторых наблюдаемых аспектах адронной физики, озадачивавших ученых на тот момент. На рис. 1.10 *а, б, в* я изобразил три фейнмановские диаграммы, на каждой из которых представлен низкоуровневый процесс с участием двух частиц (скажем, адронов), где мы имеем две частицы на входе и две частицы на выходе. На рис. 1.10 *а* два адрона сливаются в третий, который практически сразу же распадается на два других адрона; на рис. 1.10 *б* исходная пара адронов обменивается одним адроном и на выходе получается другая пара адронов. Рисунок 1.10 *в* похож на 1.10 *б* с той оговоркой, что адроны на выходе поменялись местами. Итак, зная исходные и конечные условия, мы легко можем обнаружить, что в любой из трех ситуаций существует множество вариантов того, каким получится внутренний адрон, и, чтобы получить верный ответ, нам придется просуммировать все эти варианты. Все обстоит именно так, но может показаться, что для получения верного ответа на таком уровне вычислений нам потребуется сложить все три суммы (то есть суммы, вычисленные отдельно для каждой из трех возможностей, представленных на рис. 1.10). Однако создается впечатление, что результат

вычисления каждой из трех сумм получается одним и тем же, и нам незачем складывать все три суммы, ведь каждая из них приводит к искомому ответу!

С учетом того, что было сказано выше об использовании фейнмановских диаграмм, это может показаться очень странным. Логично предположить, что все возможности следует суммировать, а природа нам словно подсказывает, что достаточно будет любого из процессов, изображенных на любой из трех разных с виду диаграмм на рис. 1.10 *а*, *б*, *в*, и что если учесть все эти процессы, то получится серьезный «перебор». Это проще понять при помощи полной формулировки КХД, если попробовать описать все эти адронные процессы на уровне взаимодействий простейших кварков, а не адронов, ведь адроны считаются составными объектами, а подсчет независимых состояний следует делать в контексте элементарных кварков. Но на тот момент, когда создавалась теория струн, полной формулировки КХД еще не существовало и казалось, что очень уместно прорабатывать другие пути для решения этой задачи (а также определенных связанных с ней задач). Способ такого решения в рамках теории струн показан на рис. 1.11 *а*, *б*, *в*, где я изобразил струнные версии всех трех вариантов, показанных на рис. 1.10 *а*, *б*, *в*. Мы отмечали, что струнные представления всех трех процессов топологически *идентичны*. Это означает, что с точки зрения «струнников» можно сделать вывод: три процесса, показанные на рис. 1.10 *а, б, в*, действительно не следует считать раздельно — это всего лишь три разных взгляда на один и тот же глубинный процесс.

Однако не все струнные диаграммы одинаковы. Рассмотрим струнную версию рис. 1.10 *г*, а именно рис. 1.11 *г*. *Петли*, возникающие на этой фейнмановской диаграмме, более высокого порядка и представлены в виде *топологических ручек* в струнных историях (см. рис. 1.*44 а* и *б* в разделе 1.11 и рис. А.11 в разделе А.5). Однако, опять же, пользуясь методами теории струн, мы получаем ключевое преимущество. Теория струн избавляет нас от расходящихся выражений такого плана, к каким нас приводили традиционные фейнмановские диаграммы, содержавшие замкнутые петли, взамен дает очень красивый способ работы с петлями, а именно позволяет рассматривать их в контексте двумерной топологии, в которой математики успели хорошо освоиться в рамках исключительно плодотворной теории римановых поверхностей (см. раздел А.10).

Подобные рассуждения подсказали отличный и интуитивно понятный мотив, чтобы всерьез отнестись к идее теории струн. Существовал и более «технический» повод, из-за которого ряд физиков взялись за исследования в этом интересном направлении. В 1970 году Йоитиро Намбу, получивший в 2008 году Нобелевскую премию за исследования в иной области — за открытие механизма спонтанного нарушения симметрии в субатомной физике, предложил объяснить

при помощи струнной идеи замечательную формулу, описывающую столкновения адронов, предложенную Габриэле Венециано двумя годами ранее. Можно отметить, что струны Намбу напоминали скорее резиновые ленты, поскольку сила их воздействия возрастала пропорционально растяжению струны (однако отличались от обычных резиновых лент тем, что сила становилась нулевой лишь в той точке, где обнулялась сама длина струны). Итак, теперь понятно, что струнная идея изначально должна была привести к теории *сильных взаимодействий*, и в данном контексте она предлагала вариант, который для своего времени был новаторским и, безусловно, привлекательным, особенно с учетом того, что КХД еще не успела оформиться в удобную теорию. Ключевую составляющую КХД, так называемую *асимптотическую свободу*, удалось описать лишь позже, в 1973 году. Это сделали Дэвид Гросс и Фрэнк Вильчек, а независимо от них — Дэвид Политцер, за что все трое в 2004 году были удостоены Нобелевской премии по физике. Струнная гипотеза давала картину, казавшуюся мне, как и многим другим, весьма достойной проработки. Но еще раз отметим: к разработке базовых исходных идей теории струн ученых подтолкнул интерес к природе *адронных* (сильных) взаимодействий.

Однако пытаясь развить полноценную квантовую теорию таких струн, ученые столкнулись с так называемой *аномалией*, которая привела их на очень странную почву. *Аномалией* называется ситуация, в которой некая теория, описанная в классических терминах — в данном случае речь идет о динамической теории струноподобных элементарных единиц, трактуемой в контексте обычной классической (то есть ньютоновской) физики, — утрачивает некое ключевое свойство, когда к ней применяются законы квантовой механики. Обычно нарушается та или иная симметрия. В случае теории струн симметрия заключалась в фактической инвариантности при изменении координатных параметров, описывающих струну. Без такой инвариантности параметров математическое описание струны не имело смысла именно из-за аномального нарушения инвариантности. Однако около 1970 года был сделан очень примечательный вывод: если число пространственно-временных измерений увеличить с четырех до 26 (то есть чтобы получилось 25 пространственных измерений и одно временно́е) — признаться, очень странная идея, — то элементы теории, порождающие такую аномалию, чудесным образом исчезают (Goddard and Thorn, 1972; см. также Greene, 1999, п. 12) и квантовая версия теории наконец-то начинает работать!

По-видимому, многие находят нечто романтичное и притягательное в идее, что может существовать мир высших измерений, недоступный для непосредственного восприятия, и более того, что такая высшая размерность может быть неотъемлемой составляющей реального обитаемого мира! Но я отреагировал на нее совсем не так. Узнав эту новость, я подумал, что, несмотря на математи-

ческую занимательность этой гипотезы, я не могу всерьез воспринимать ее как заслуживающую внимания физическую модель известной нам Вселенной. Итак, поскольку мне не показали, что может существовать какая-то другая (радикально иная) трактовка всего этого, улетучился весь изначальный интерес и азарт, который струнные идеи разожгли во мне как в физике. Я полагаю, что среди физиков-теоретиков не я один отреагировал подобным образом, но на то были и особые причины, по которым меня решительно не устраивало предложение так резко увеличить число пространственных измерений. Я подробно расскажу об этих причинах в разделах 1.9–1.11, 2.9, 2.11, 4.1 и особенно подробно в разделе 4.4, но в данном контексте мне потребуется вкратце описать принятый среди теоретиков-струнников взгляд на вещи, благодаря которому их *не смущает* кажущееся противоречие между очевидной трехмерностью физического пространства (с одномерным физическим временем) и этими постулируемыми 25 пространственными измерениями (плюс одним временным), существования которых теперь очевидно требовала теория струн.

Речь идет о так называемых *бозонных струнах*, которые предположительно должны соответствовать частицам-*бозонам*. В разделе 1.14 я расскажу, что все квантовые частицы делятся на два класса: к первому относятся бозоны, а ко второму — *фермионы*. Характерное различие бозонов и фермионов заключается в их статистических свойствах, а также в том, что спин бозонов всегда равен *целому числу* (в абсолютных единицах, см. раздел 2.1), а спин фермиона — *полуцелому*. Эти проблемы будут рассмотрены в разделе 1.14, где мы обсудим их в связи с гипотезой суперсимметрии — эта гипотеза была выдвинута с целью попытаться объединить бозоны и фермионы в общую картину. Мы увидим, что эта гипотеза играет ключевую роль в большей части современной теории струн. Действительно, Майкл Грин и Джон Шварц (1984; см. также [Green, 1999]) обнаружили, что с учетом суперсимметрии число измерений, при котором теория струн становится работоспособной, снижается с 26 до 10 (то есть девять пространственных измерений и одно временное). Струны в такой теории, именуемые *фермионными*, описывают фермионы; связь между ними и бозонами существует на уровне суперсимметрии.

Теоретики-струнники, чтобы не было так мучительно больно за это колоссальное и очевидно абсурдное несоответствие между теорией и наблюдаемыми фактами, обращаются к более ранней гипотезе, которую в 1921 году выдвинул немецкий математик Теодор Калуца[1], а впоследствии развил шведский физик Оскар Клейн. Сегодня их теория известна под названием *теория Калуцы — Клейна*. В соответ-

[1] В некоторых источниках Калуцу называют поляком. Дело в том, что его родной город Ополе (*нем.* Оппельн) ныне находится на территории Польши.

ствии с ней гравитация и электромагнетизм одновременно описываются в пятимерном пространстве-времени. Как же Калуца и Клейн объясняли, почему пятое измерение пространства-времени, описываемое их теорией, недоступно обитателям нашей Вселенной для непосредственного наблюдения? В исходной картине Калуцы пятимерное пространство-время должно было иметь в точности такие параметры, какие предполагает чистая эйнштейновская теория гравитации, но в этом пространстве должна быть строгая симметрия вдоль особого векторного поля **k** в пятимерном пространстве (см. раздел А.6, рис. А.17), причем геометрия в направлении **k** абсолютно не меняется. В терминологии дифференциальной геометрии **k** — это так называемый *вектор Киллинга*, то есть векторное поле, порождающее такую непрерывную симметрию (см. раздел А.7, рис. А.29). Более того, любой физический объект, описываемый в пространстве-времени, также будет иметь постоянное описание вдоль **k**. Поскольку любому объекту будет присуща такая симметрия, ничто в пространстве-времени «не выдаст» существования этого направления, и *фактическое* пространство-время относительно своего содержимого будет четырехмерным. Тем не менее структура четырехмерного фактического пространства-времени, обусловленная существованием пяти измерений, будет интерпретироваться в рамках этого четырехмерного пространства как электромагнитное поле, удовлетворяющее уравнениям Максвелла и присутствующее в эйнштейновском тензоре энергии **T** именно таким образом, каким и должно[1]. Действительно, это была исключительно хитроумная идея. Фактически пятимерное пространство Калуцы является *расслоением* $\mathcal{B}$ в том смысле, как этот феномен понимается в разделе А.7, с одномерным слоем. Базисное пространство — это наше четырехмерное пространство-время $\mathcal{M}$, но не *погруженное* естественным образом в пятимерное пространство $\mathcal{B}$. Это связано с тем, что элементы в четырехмерной плоскости «повернуты» перпендикулярно направлениям **k**, и этот поворот описывает электромагнитное поле (рис. 1.12).

Затем в 1926 году Клейн предложил иную трактовку пятимерного пространства Калуцы, изложив такую идею: дополнительное измерение в направлении **k** будет «маленьким» в том смысле, что оно свернуто в крошечную петлю (S^1). Чтобы представить себе такую картину, обычно приводят аналогию со шлангом (рис. 1.13). В этой аналогии четыре макроскопические измерения обычного пространства-времени приравниваются к единственному направлению (длине)

---

[1] Чтобы вы могли оценить «скрученную» дифференциальную геометрию пятимерного пространства Калуцы, описанную на техническом геометрическом языке, отметим, что для того чтобы **k** мог быть вектором Киллинга, ковариантная производная **k**, выраженная в виде *ко*вектора, должна быть антисимметричной, и в таком случае мы убедимся, что эта двумерная форма фактически представляет собой поле Максвелла в четырехмерном пространстве.

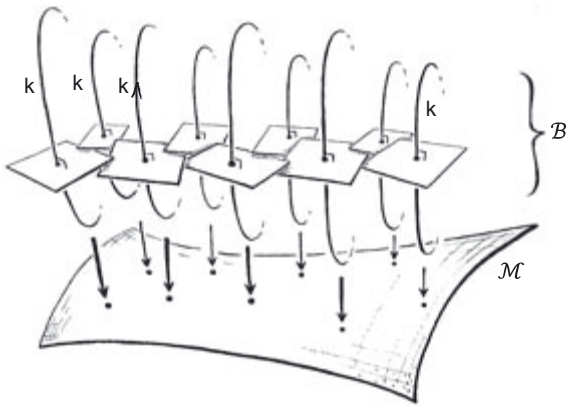


Рис. 1.12. Поскольку существует симметрия вдоль **k**-направлений вектора Киллинга, пятимерное пространство Калуцы — Клейна является расслоением $\mathcal{B}$, наложенным на знакомое нам четырехмерное пространство-время $\mathcal{M}$, где **k** направлен вдоль слоев S^1 (вертикальные кривые). Поле Максвелла зашифровано в «повороте» слоев, не дающем ортогональным четырехмерным пространствам спутаться друг с другом. Получаются согласованные участки четырехмерного пространства-времени, которые в противном случае были бы отображениями пространства-времени $\mathcal{M}$

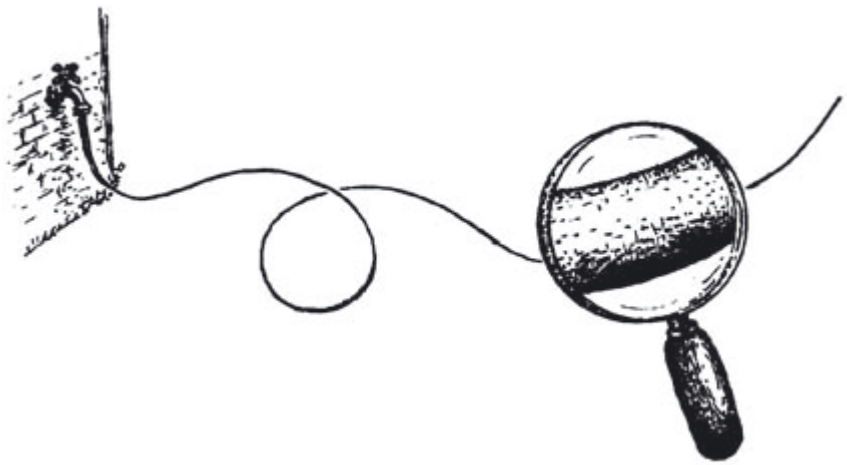


Рис. 1.13. Метафора со шлангом помогает понять предположение Клейна о том, что дополнительные измерения должны быть крошечными — возможно, иметь размеры порядка планковской длины. Если рассматривать шланг издалека, то он кажется одномерным, и это свойство можно сравнить с мнимой четырехмерностью пространства-времени. Под лупой дополнительное измерение шланга становится заметным, и в этом оно напоминает гипотетические субмикроскопические пространственные измерения

шланга, а дополнительное «малое» пространственно-временное измерение из теории Калуцы — Клейна соответствует крошечной петле, охватывающей шланг, — возможно, эта петля имеет планковские масштабы порядка ~10^{-35} м (см. раздел 1.5). Если рассматривать шланг издалека, то он кажется одномерным, и дополнительное измерение, благодаря которому поверхность шланга на практике является двумерной, напрямую не наблюдается. Соответственно, в этой картине Калуцы — Клейна пятое измерение аналогично крошечному контуру по окружности шланга; считается, что на уровне повседневного опыта заметить такое измерение невозможно.

Аналогично теоретики-струнники предположили, что дополнительные 22 измерения из теории струн будут совсем «крошечными», подобно пятому, дополнительному измерению Калуцы — Клейна, и, следовательно, их невозможно заметить на макроскопическом уровне. Соответственно, считали они, те 22 пространственные измерения, которые, по-видимому, требуются для устранения аномалий в теории струн, могут оставаться для нас незаметными. Действительно, как упоминалось в начале этого раздела, струнное обоснование адронной физики, по-видимому, предполагает, что масштаб адронов (около 10^{-15} м) вполне сравним по «величине» с этими дополнительными пространственными измерениями, которые в самом деле получаются слишком крошечными для обыденного восприятия, но критически важны для частиц такого размера, как адроны. Как мы увидим в разделе 1.9, в более современных версиях теории струн предполагается, что дополнительные измерения заметны в гораздо более мелких масштабах — возможно, в пределах 10^{-33}–10^{-35} м.

Есть ли смысл в такой гипотезе? Мне кажется, при ее принятии возникает глубокая проблема, связанная с *функциональной свободой*. Эта проблема упоминалась в предисловии и подробно обсуждается в разделе A.2 и разделе A.8, к которым отсылаю неискушенного читателя (см. также [Cartan, 1945] и [Bryant et al., 1991]). Если говорить о классических полях, подчиняющихся обычным уравнениям, описывающим, как эти поля распространяются во времени, то число пространственных измерений, вовлеченных в этот процесс, приобретает огромное значение. Ведь при увеличении числа пространственных измерений, в которых мы допускаем существование поля (и при наличии все того же единственного временного измерения), функциональная свобода радикально возрастает. С обозначениями, используемыми в разделе A.2, функциональная свобода c-компонентного поля, свободно описываемого в пространстве с d пространственными измерениями, описывается формулой:

$$\infty^{c\infty^{d}}.$$

Разница между такой степенью свободы и степенью свободы для C-компонентного поля в пространстве с иным числом пространственных измерений (D) выражается в виде:

$$\infty^{C\infty^D} \gg \infty^{C\infty^d}, \text{ если } D > d,$$

независимо от относительного числа компонентов на точку, то есть от C и c. Удвоенный знак неравенства >> призван показать крайнюю, недосягаемую колоссальность, причем функциональная свобода, выраженная в левой части неравенства, превышает аналогичное значение свободы в правой части при условии увеличения числа пространственных измерений, причем неважно, чему именно равны числовые значения компонентов C и c (см. разделы A.2 и A.8). Основной момент заключается в том, что для обычных классических полей с конечным числом компонентов на точку — предполагается, что там будут действовать обычные уравнения полей, описывающие детерминированное развитие во времени исходя из (фактически) свободно задаваемых данных в d-мерном начальном пространстве, — число d становится критически важным. Такая теория не может быть эквивалентна иной подобной теории, где начальное пространство содержит иное число измерений (D). Если D больше d, то свобода в теории для D-пространства всегда неизмеримо выше, чем в теории для d-пространства!

Хотя эта ситуация, на мой взгляд, совершенно ясна в контексте классических теорий поля, в случае квантовых теорий (поля) это вовсе не так очевидно. Тем не менее квантовые теории обычно моделируются на базе классических, поэтому отклонения квантовой теории от классической, на которой она основана, в первую очередь должны представлять собой обычные корректировки классической теории на уровне квантовой. В случае квантовых теорий такого рода требуются очень веские причины, чтобы полагать, что между квантовыми теориями, предусматривающими разное число пространственных измерений, может сохраняться какая-либо эквивалентность.

Закономерно возникают вопросы о физической релевантности квантовых теорий, например теорий струн с дополнительными измерениями, где число пространственных измерений не ограничивается теми тремя, которые мы непосредственно воспринимаем. Что происходит с той степенью свободы, которая потенциально доступна в системе за счет дополнительных пространственных измерений? Возможно ли, что эти безграничные степени свободы окажутся скрыты и не будут определять физику в таких картинах мира?

В определенном смысле это *возможно* даже в классической теории, но при условии, что этих дополнительных степеней свободы *просто не существует*. Именно такова была ситуация с исходной теорией Калуцы о пятимерном пространстве-

времени, где требовалось наличие *точной* непрерывной симметрии в одном дополнительном измерении. Симметрия была обусловлена тем, что в исходной картине Калуцы **k** по своей природе являлось векторным полем Киллинга, поэтому функциональная свобода уменьшалась до уровня, свойственного обычной теории с тремя пространственными измерениями.

Следовательно, чтобы проверить правдоподобие струнных идей, предполагающих существование высших измерений, необходимо сначала попытаться понять, что именно стремились сделать Калуца и Клейн. Они хотели построить геометрическую картину электромагнетизма в духе общей теории относительности Эйнштейна, представив эту силу как проявление самой природы пространства-времени. Из раздела 1.1 мы помним, что общая теория относительности, впервые целиком опубликованная в 1916 году, позволила Эйнштейну вписать природу гравитационного поля в структуру искривленного четырехмерного пространства-времени. На тот момент из фундаментальных сил природы были известны лишь гравитационное и электромагнитное поля, и было логично полагать, что с определенной точки зрения полное описание электромагнетизма, учитывающее его взаимосвязь с гравитацией, должно также быть выполнено в контексте некоей пространственно-временной геометрии. Примечательно, что Калуце действительно удалось достичь этой цели, но ценой привнесения дополнительного измерения в пространственно-временной континуум.

1.7. Время в общей теории относительности Эйнштейна

Прежде чем подробнее рассмотреть пятимерное пространство-время из теории Калуцы — Клейна, познакомимся с методом описания электромагнитных взаимодействий, который в итоге вошел в состав стандартной модели. Здесь нас в первую очередь интересует, каким образом описываются электромагнитные взаимодействия квантовых частиц (речь идет о лоренцовской расширенной версии теории Максвелла, которая, как упоминалось в разделе 1.5, демонстрирует реакцию заряженных частиц на электромагнитное поле), а также обобщения этих законов для сильного и слабого взаимодействий в стандартной модели. Эту картину впервые составил в 1918 году великий немецкий математик и физик-теоретик Герман Вейль. (В 1933—1955 годах Вейль стал одним из столпов Принстонского института перспективных исследований — в то же самое время там работал и Эйнштейн. Однако, как и Эйнштейн, Вейль внес свой основной вклад в физику ранее, когда жил в Германии и Швейцарии.) Исходная идея Вейля заключалась в следующем: расширить общую теорию относительности

Эйнштейна таким образом, чтобы максвелловский электромагнетизм (великая теория, вкратце упомянутая в разделах 1.2 и 1.6) естественным образом вписывался в геометрическую структуру пространства-времени. Для этого Вейль ввел концепцию так называемой *калибровочной связности*. Наконец, после того как в теорию были внесены некоторые ювелирные изменения, идея Вейля стала играть центральную роль при трактовке взаимодействий в стандартной модели физики частиц. В терминах математики (в основном под влиянием Анджея Траутмана (Trautman, [1970])) такая идея калибровочной связности сегодня трактуется в контексте *расслоения* (см. раздел A.7), проиллюстрированного на рис. 1.12 (о чем уже упоминалось в разделе 1.3). Важно понимать, в чем исходная идея Вейля о калибровочной связности отличалась от высказанной чуть позже идеи Калуцы — Клейна и чем на нее походила.

В разделе 1.8 я немного подробнее опишу, как Вейль представил свое геометрическое расширение эйнштейновской общей теории относительности, в которое собирался включить теорию Максвелла. Мы убедимся, что теория Вейля не предполагает никакого увеличения размерности пространства-времени, однако снижает роль *метрики*, лежащей в основе теории Эйнштейна. Итак, для затравки я хочу поговорить о той фактической роли, которую в эйнштейновской системе играет *метрический тензор* **g**. Действительно, это базовая величина, определяющая псевдориманову структуру пространства-времени. Как правило, физики пользуются нотацией вроде g_{ab} (или g_{ij}, или $g_{\mu\nu}$, или тому подобной) для обозначения *набора компонентов* тензорной величины **g**, но я не собираюсь вдаваться в детали этих проблем и даже не планирую объяснять, что именно означает термин «тензор» с математической точки зрения. В данном случае нас интересует самая непосредственная физическая интерпретация, которая может присваиваться **g**.

Предположим, у нас есть кривая $\mathcal{C}$, соединяющая две точки — или два *события* — P и Q в пространственно-временном многообразии $\mathcal{M}$, где $\mathcal{C}$ представляет историю некоторой массивной частицы, развивающуюся от события P до более позднего события Q. (Термин «*событие*» часто означает точку в пространстве-времени.) Кривая $\mathcal{C}$ именуется *мировой линией* этой частицы. Итак, в теории Эйнштейна величина **g** нужна для определения «*длины*» кривой $\mathcal{C}$, причем физически эта длина интерпретируется как *временной* интервал (а не расстояние) между точками P и Q, который был бы измерен *идеальными часами*, установленными на частице (рис. 1.14 *a*).

Необходимо помнить, что, согласно эйнштейновской теории относительности, «ход времени» — это не строго абсолютный процесс, синхронно протекающий во всей Вселенной. Напротив, следует мыслить в терминах неразрывного *пространства-времени*. Пространство-время никоим образом не «шинкуется» на

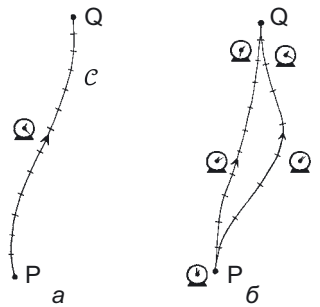


Рис. 1.14. Пространственно-временная метрика **g** присваивает «длину» любому отрезку мировой линии частицы C и, таким образом, представляет собой временной интервал, который измерили бы идеальные часы, движущиеся по мировой линии (*а*); если две такие мировые линии соединяют конкретные события P и Q, то результаты измерений для двух различных мировых линий могут различаться (*б*)

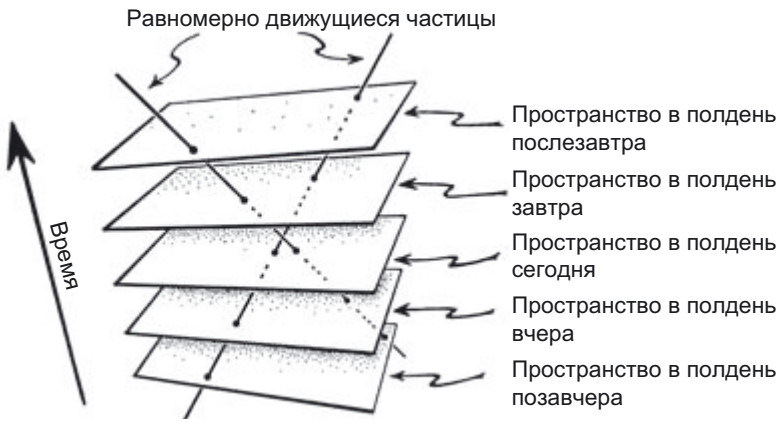


Рис. 1.15. Ньютоновское представление об идеальном времени (где можно представить себе идеальные часы, ежедневно отбивающие полдень одновременно во всем пространстве). Теория относительности отвергает такую точку зрения, но мы можем условно трактовать пространство-время именно таким образом в качестве отличной аппроксимации для объектов, скорости которых много меньше скорости света

трехмерные участки пространства, каждый их которых содержит совокупность событий, происходящих «одновременно». У нас *нет* никаких абсолютных «универсальных часов», которые просто тикают, а каждому удару этих часов соответствует отдельное трехмерное пространство с одновременными событиями, а следующий удар часов — это уже новая совокупность одновременных событий, и т. д., причем все эти трехмерные пространства слагаются в пространство-время (рис. 1.15: можно представить себе, что наши универсальные часы отбивают

каждый полдень). Вполне допустимо условно воспринимать пространство-время именно таким образом, это позволит соотнести эту четырехмерную картину с повседневным опытом, согласно которому существует трехмерное пространство, явления которого «развиваются во времени». Однако обязательно нужно понимать, что в подобном разбиении пространства-времени нет ничего особенного или «ниспосланного Богом» по сравнению с другими вариантами. *Целостное* пространство-время — это абсолютный феномен, но мы не можем считать приоритетным какой-либо вариант *разбиения* пространства-времени и считать, что именно он соответствует универсальному понятию, которое можно было бы охарактеризовать как «время как оно есть». (Все это — элементы принципа общей ковариантности, подробнее описанного в разделе A.5, согласно которому конкретные координаты отсчета, в частности координата времени, не обязаны иметь непосредственный физический смысл.) Напротив, для мировой линии каждой конкретной частицы характерен собственный ход времени, зависящий как от самой мировой линии, так и от метрики **g**, как это описано выше. Правда, различия между временем одной частицы и временем другой очень невелики, если только относительные скорости этих частиц не становятся достаточно высокими и сравнимыми со скоростью света (или если мы не оказываемся в каком-либо месте, где возникает колоссальное искривление пространства-времени, обусловленное гравитацией), причем такая незначительность есть необходимое условие, благодаря которому мы не замечаем подобной рассогласованности темпа времени на уровне повседневного опыта.

В теории относительности Эйнштейна если у нас есть две мировые линии, соединяющие два конкретных события P и Q (см. рис. 1.14 *б*), то их «длина» (то есть измеренное истекшее время) может действительно различаться в первом и во втором случае (этот эффект многократно удавалось непосредственно измерить, например, при помощи сверхточных часов на очень быстро летящих самолетах или на самолетах, летящих на сильно различающихся высотах [Will, 1993]. В этом неочевидном факте проявляется так называемый парадокс близнецов из специальной теории относительности, согласно которому космонавт, слетавший с огромной скоростью от Земли к далекой звезде и обратно, может провести в полете значительно меньше времени, чем его брат-близнец проведет на Земле, пока первый брат будет в пути. У двух близнецов разные мировые линии, однако они соединяют одни и те же события, а именно P (когда они вместе и космонавт только готовится к старту) и Q (когда космонавт возвращается на Землю).

На рис. 1.16 ситуация показана в рамках специальной теории относительности (где движения в основном равномерны), и на этом рисунке отмечено третье событие (R) — прибытие космонавта к далекой звезде. Аналогично на рис. 1.17 показано, как субъективный ход времени зависит от метрики, и эта картина

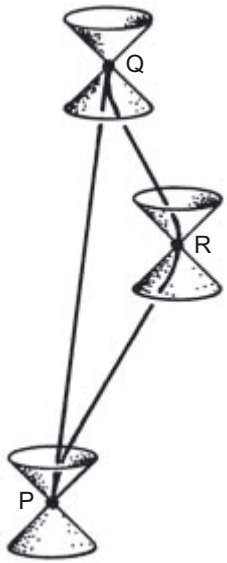


Рис. 1.16. Так называемый парадокс близнецов специальной теории относительности. Близнец, оставшийся на Земле и движущийся вдоль мировой линии PQ, проживает больший промежуток времени, чем близнец-космонавт с мировой линией PRQ (здесь возникает любопытная ситуация, обратная известному неравенству треугольников из евклидовой геометрии: PR + RQ > PQ). Световые конусы рассмотрены на рис. 1.18

также применима в более универсальном контексте *общей* теории относительности, где для мировой линии массивной частицы «длина» отрезка этой линии определяется метрикой **g** и дает собственный интервал времени за этот период. На обоих рисунках изображены световые конусы — важное физическое проявление эйнштейновской метрики **g**, дающие пространственно-временное описание *распространения света* в пространстве-времени. Мы видим, что для любого события проходящая через него мировая линия частицы или космонавта должна находиться внутри двойного светового конуса, в вершине которого находится это событие. Так иллюстрируется важное ограничение: скорость света нельзя (локально) превысить.

На рис. 1.18 показана физическая интерпретация той части двойного светового конуса, которая расположена в будущем, то есть непосредственная история (гипотетической) вспышки света при событии X. На рис. 1.18 *а* представлена трехмерная пространственная картина, а на рис. 1.18 *б* — соответствующая пространственно-временная картина, на которой исключено одно пространственное измерение. Часть двойного конуса, обращенная в прошлое, соответствует гипотетической вспышке света, сходящейся в X. На рис. 1.18 *в* видно, что при каждом

Рис. 1.17. В искривленном пространстве-времени общей теории относительности метрический тензор **g** позволяет вычислить собственное время. Так обобщается плоское представление пространства-времени из специальной теории относительности, показанное на рис. 1.16

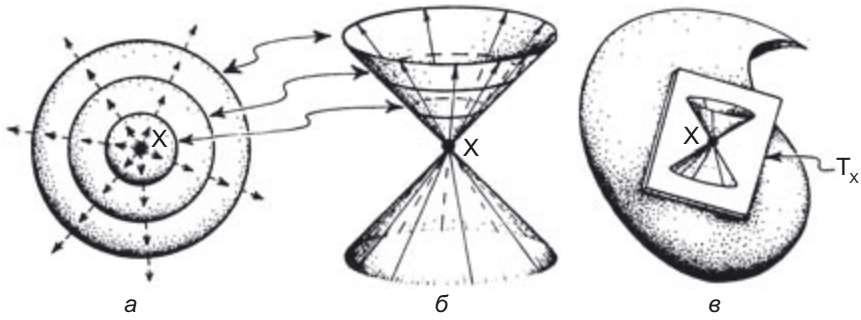


Рис. 1.18. В каждой точке X пространства-времени задается двойной световой конус, определяемый метрикой **g**, состоящий из светового конуса прошлого и светового конуса будущего. Собственное время вдоль поверхности конуса всегда равно нулю. Световой конус будущего локально интерпретируется как история гипотетической вспышки света, излученной в X. *а*) пространственная картина; *б*) пространственно-временная картина (в которой исключено одно пространственное измерение), где световой конус прошлого соответствует истории гипотетической вспышки света, сходящейся в X; *в*) технически световой конус представляет собой бесконечно малую структуру в непосредственной близости от события X, то есть он расположен в касательном пространстве T_X

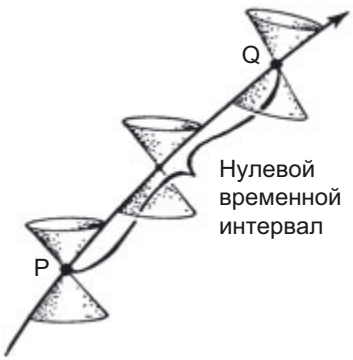


Рис. 1.19. Вдоль луча света (или любой нулевой кривой) время между любыми двумя событиями P и Q всегда равно нулю

событии X нулевой конус на самом деле является *бесконечно малой* структурой и локализуется фактически лишь в *касательном пространстве* при X (см. раздел A.5 и рис.A.10).

Эти конусы ориентированы в направлениях, вдоль которых время *исчезает*. Это происходит потому, что геометрия пространства-времени в строгом смысле является *псевдо*римановой, а не римановой (как отмечалось в разделе 1.1). Зачастую псевдориманова геометрия данного конкретного типа именуется термином *лоренциан*, и в ее пространственно-временной структуре присутствует всего одно временное измерение и (n – 1) пространственных измерений, а в каждой точке пространственно-временного многообразия присутствует такой двойной световой конус. Световые конусы представляют наиболее важную особенность структуры пространства-времени, поскольку демонстрируют границы распространения информации.

Как собственное время, вычисляемое через **g**, непосредственно соотносится с этими световыми конусами? До сих пор мы рассматривали мировые линии, представлявшие собой истории обычных частиц, обладающих массой, и считалось, что скорость этих частиц ниже световой, поэтому их мировые линии должны находиться *внутри* световых конусов. Однако также нужно рассмотреть *безмассовые* частицы вроде фотонов (частиц света), и подобные частицы будут перемещаться со скоростью света. Согласно теории относительности, если бы часы летели со скоростью света, то вообще бы не фиксировали ход времени! Следовательно, длина «мировой линии» (измеряемая по кривой) между двумя событиями P и Q для безмассовой частицы всегда *нулевая* (рис. 1.19), независимо от того, насколько эти события удалены друг от друга. Такая мировая линия

именуется *нулевой* кривой. Некоторые нулевые кривые являются *геодезическими* (см. ниже), а мировая линия свободного фотона считается геодезической линией нулевой длины.

Совокупность всех геодезических линий, проходящих через конкретную точку P в пространстве-времени, заполняет световой конус P (рис. 1.20), *нулевой конус* в точке P описывает лишь бесконечно малую структуру на *вершине* светового конуса P (см. рис. 1.18). Нулевой конус задает пространственно-временные *направления* в точке P, определяющие скорость света, то есть эта структура в *касательном пространстве* в точке P указывает, какие направления имеют нулевую длину согласно метрике **g**. В литературе термин *световой конус* часто используется в том смысле, который я здесь вкладываю в термин *нулевой конус*. Я в дальнейшем буду использовать термин *световой конус* для обозначения объемного тела, а *нулевой конус* — для обозначения его поверхности. Световой конус (как и нулевой конус, описанный выше) состоит из двух частей, первая из которых определяет *нулевые направления в будущем*, а вторая — *нулевые направления в прошлом*. Теория относительности накладывает на частицы, обладающие массой, ограничение, а именно: их скорость не может превысить локальную скорость света. Это требование выражается явно, как и тот факт, что все *касательные направления* к мировым линиям частиц, обладающих массой, лежат внутри нулевых конусов соответствующих событий (рис. 1.21). Подобные гладкие кривые, у которых все касательные направления лежат строго в пределах нулевых конусов, называются *времениподобными*. Следовательно, мировые линии частиц, обладающих массой, всегда являются времениподобными.

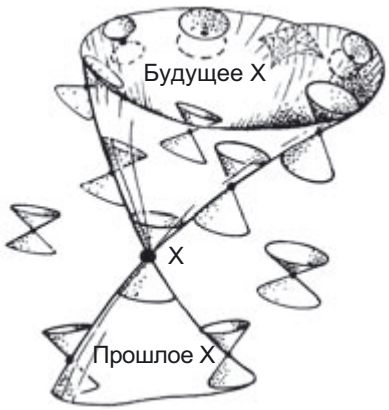


Рис. 1.20. Световой конус события X — это локус в пространстве-времени, описываемый нулевыми геодезическими линиями, проходящими через X. Касательная структура на его вершине X — это нулевой конус при X

Понятие времениподобной кривой дополняется *пространственноподобной* трехмерной поверхностью — она же пространственноподобная $(n-1)$-мерная поверхность, или пространственноподобная *гиперповерхность*, если имеется в виду n-мерное пространство-время. Все касательные направления к такой гиперповерхности лежат *вне* нулевых конусов (рис. 1.21). В общей теории относительности используется понятие *обобщенного момента времени* — единого момента времени t, заданного для некоей фиксированной гиперповерхности. Очевидно, что выбор такой гиперповерхности во многом произволен, но эта сущность нам необходима, если мы собираемся рассуждать о проблемах вроде *детерминизма* динамических явлений; в таких случаях можно потребовать указать на такой гиперповерхности начальные условия, предполагая, что эти условия должны (локально) определять эволюцию системы в прошлом или будущем путем решения соответствующих уравнений (как правило, речь идет о дифференциальных уравнениях, см. раздел А.11).

Можно обратить внимание и на другую особенность теории относительности: если «длина» (в данном случае — истекшее измеренное время) мировой линии $\mathcal{C}$, соединяющей P и Q, превышает длину любой иной мировой линии между P

Рис. 1.21. Нулевые касательные векторы при X задают нулевые конусы, как на рис. 1.18, но существуют еще и времениподобные кривые; если они направлены в будущее, то описывают касательные векторы (4-скорости) к мировым линиям обладающих массой частиц. Кроме того, существуют пространственноподобные кривые, направленные за пределы конуса и являющиеся касательными к пространственноподобным поверхностям, проходящим через X

и Q, то линия 𝒞 должна быть *геодезической*[1], которая в искривленном пространстве-времени является аналогом «прямой линии» (рис. 1.22). Любопытно, что такая максимизация «длин» в пространстве-времени *противоположна* ситуации в обычной евклидовой геометрии, где прямая линия между двумя точками P и Q соответствует кратчайшему из путей между P и Q. Согласно теории Эйнштейна, мировая линия частицы, свободно движущейся под действием гравитации, всегда является геодезической. Однако в истории с путешествием космонавта, показанной на рис. 1.16, речь идет о движении с ускорением, а значит, траектория не геодезическая.

Плоское пространство-время специальной теории относительности, в котором отсутствует гравитационное поле, называется пространством Минковского (которое я предпочитаю обозначать символом $\mathbb{M}$) в честь русско-немецкого математика Германа Минковского, впервые сформулировавшего в 1907 году идею

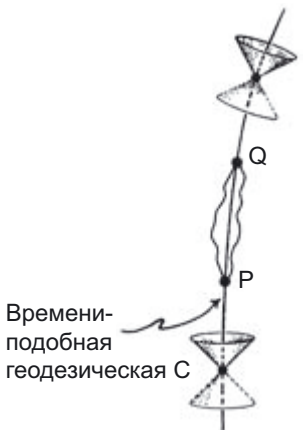


Рис. 1.22. Времениподобная кривая, соответствующая максимальному промежутку времени между двумя разнесенными на ней событиями P и Q, обязательно является геодезической

[1] И наоборот, *каждая* мировая линия 𝒞, оказывающаяся геодезической, обладает таким характерным свойством в *локальном* смысле, то есть если для любой точки P на кривой 𝒞 найдется достаточно небольшой открытый участок $\mathcal{N}$ многообразия $\mathcal{M}$, содержащий P, то для любой пары точек на 𝒞 внутри $\mathcal{N}$ мировая линия максимальной длины окажется совпадающей с 𝒞. С другой стороны, если две точки слишком далеко разнесены вдоль геодезической 𝒞, то можно обнаружить, что 𝒞 не всегда обеспечивает максимальное расстояние. Это связано с тем, что между двумя рассматриваемыми точками на 𝒞 имеются пары сопряженных точек [Penrose, 1972; Hawking and Ellis, 1973].

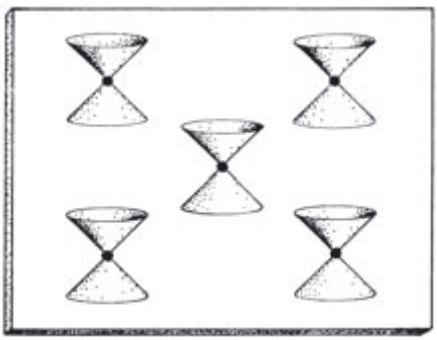


Рис. 1.23. Пространство Минковского — это плоское пространство-время, описываемое специальной теорией относительности. Нулевые конусы в нем распределены совершенно равномерно

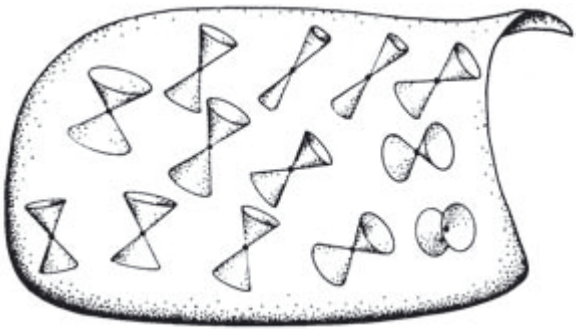


Рис. 1.24. Согласно общей теории относительности, нулевые конусы могут распределяться неоднородно

пространства-времени. Здесь нулевые конусы однородно распределены в пространстве (рис. 1.23). В теории относительности Эйнштейна также принимается эта идея, но считается, что нулевые конусы могут располагаться и неоднородно, что объясняется наличием гравитационного поля (рис. 1.24). Метрика **g** (десять компонентов на точку) определяет это распределение нулевых конусов, но это распределение зависит не только от метрики. Иногда такое распределение нулевых конусов именуют *конформной* структурой пространства-времени (девять компонентов на точку); см., в частности, раздел 3.5. Кроме этой лоренцевой конформной структуры, **g** определяет *масштаб* (один компонент на точку) и тем самым фиксирует, в каком темпе идеальные часы должны отсчитывать время в теории Эйнштейна (рис. 1.25). Подробнее о взаимосвязи часов и теории относительности рассказано, например, в работах [Rindler, 2001] и [Hartle, 2003].

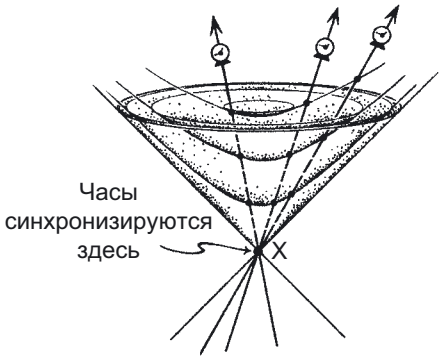


Рис. 1.25. Метрический масштаб в момент события X будет зависеть от того, с какой скоростью идеальные часы проходят через X. Здесь через X проходит несколько идентичных часов, и каждые задают свой метрический масштаб, при котором «удары» различных часов будут синхронизовываться друг с другом при помощи показанных на рисунке воронкообразных поверхностей (на самом деле это гиперболические трехмерные поверхности)

1.8. Вейль и его калибровочная теория электромагнетизма

Исходная идея Вейля, высказанная в 1918 году и призванная инкорпорировать электромагнетизм в общую теорию относительности, предполагала, в частности, редуцировать метрическую структуру пространства-времени до *конформной структуры*, как это описано выше, чтобы исключить абсолютную меру для темпа времени, хотя нулевые конусы в ней все-таки определялись [Weyl, 1918]. Кроме того, в теории Вейля присутствует и концепция идеальных часов, поэтому для времениподобной кривой можно определить меру «длины» относительно любых *конкретных* подобных часов, хотя *темп*, в котором часы отсчитывают время, будет зависеть от движения часов. Однако в теории Вейля нет *абсолютной* шкалы времени, так как ни одни из идеальных часов не являются более предпочтительными, нежели другие. Более того, мы могли бы иметь двое таких часов, тикающих абсолютно синхронно, когда при каком-то событии P они находятся в состоянии покоя друг относительно друга, но если при другом событии Q они станут двигаться по разным маршрутам в пространстве-времени, то может оказаться, что *темп* первых и вторых часов не совпадает, и эти часы не будут тикать синхронно, когда окажутся в состоянии покоя друг относительно друга при событии Q (рис. 1.26 *а*). Важно отметить, что эта ситуация отлична от «парадокса близнецов» из эйнштейновской теории относительности и более

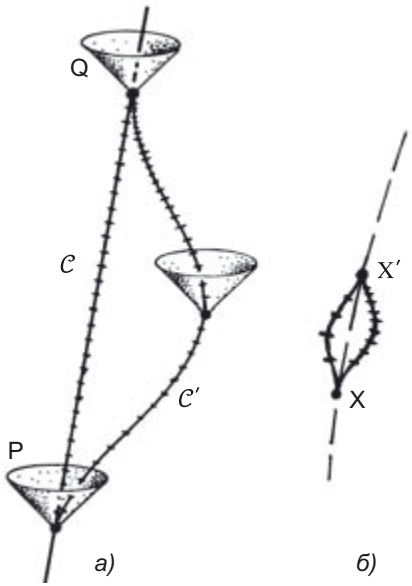


Рис. 1.26. Идея Вейля о калибровочной связности предполагает, что метрическая шкала не данность и может быть перенесена из точки P в точку Q по соединяющей их кривой C, причем перенос из P в Q по иной кривой C′ может дать иной результат (*а*); Вейлева *калибровочная кривизна* возникает из *бесконечно малой* вариации этого различия, и Вейль исходно предполагал, что эта кривизна и будет тензором электромагнитного поля Максвелла (*б*)

экстремальна. В данном случае, хотя *показания* часов и могут зависеть от их истории, *темп* часов от истории не зависит. Более общая вейлева геометрия подводит нас к любопытному «искривлению» пространства-времени на уровне темпа часов, причем такое рассогласование скорости хода часов происходит в бесконечно малых масштабах (рис. 1.26 *б*). Аналогичная ситуация прослеживается, когда различие кривизны поверхности можно выразить в угловой мере, в чем мы вскоре убедимся (рис. 1.27). Вейлю удалось показать, что величина **F**, описывающая предложенный им тип кривизны, в точности удовлетворяет тем же уравнениям, что и величина, описывающая свободное электромагнитное поле в теории Максвелла! Поэтому Вейль предположил, что его величина **F** физически тождественна электромагнитному полю Максвелла.

Временные и пространственные меры, в сущности, эквивалентны друг другу в окрестности любого события P, если у нас есть концепция нулевого конуса в P, поскольку он фиксирует скорость света в P. Говоря простым языком, скорость света позволяет преобразовывать пространственные меры во временные

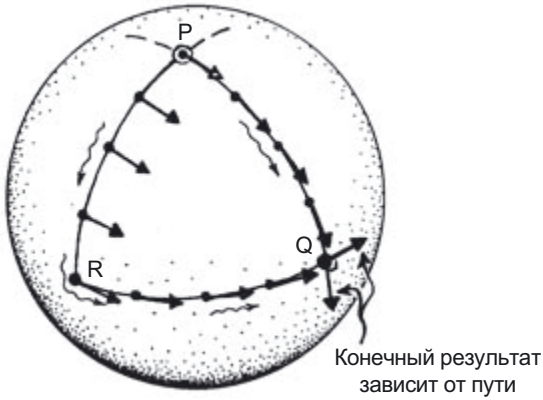


Рис. 1.27. Аффинная связность выражает идею параллельного переноса касательных векторов вдоль кривых, где различие между переносами по разным кривым позволяет узнать меру кривизны. Это можно проследить на сфере, где перенос касательного вектора по ортодромии из P в Q дает совершенно не тот результат, нежели при переносе по дуге большого круга от P к R и далее по подобной дуге от R к Q

и обратно. Следовательно, интервал времени длительностью в год соответствует расстоянию в один световой год, одна секунда — одной световой секунде и т. д. На самом деле в современных экспериментах временные интервалы можно непосредственно определять с гораздо большей точностью, чем пространственные. Так, сегодня считается, что метр равен *точно* 1/299792458 световой секунды, а точная скорость света, измеряемая в метрах в секунду, выражается целым числом 299792458! Соответственно, кажется крайне уместным рассуждать о структуре пространства-времени в терминах *хронометрии*, а не геометрии, как предлагал выдающийся специалист по теории относительности Дж. Л. Синг [J. L. Synge 1921, 1956].

Я описал идею Вейля в контексте измерения времени, но, вероятно, Вейль имел в виду в первую очередь изменения положения в пространстве, и его система именовалась *калибровочной теорией*, причем под «калибровкой» понималась шкала, по которой отмерялись физические значения длины. Суть замечательной идеи Вейля заключалась в том, что шкалу не требуется определять глобально, раз и навсегда для всего пространства-времени, но если калибровка проведена при одном событии P, а вдобавок известно, что кривая $\mathcal{C}$ соединяет P с другим событием Q, то калибр можно однозначно перенести по $\mathcal{C}$ от P к Q. Однако если существует другая геодезическая $\mathcal{C}'$, соединяющая P и Q, то перенос шкалы к Q вдоль $\mathcal{C}'$ может дать иной результат. Математическая величина, характеризующая такую процедуру «переноса калибровки», называется *калибровочной*

связностью, а различия, возникающие при выборе разных путей, позволяют оценить меру *калибровочной кривизны*. Следует отметить, что блестящая идея Вейля о калибровочной связности, вероятно, возникла у него благодаря знакомству со связностью другого типа, которой по определению обладает любое (псевдо-)риманово многообразие — так называемой *аффинной связностью*, сопряженной с параллельным переносом касательных векторов вдоль кривых, и такая связность также зависит от избранного пути, что можно убедительно продемонстрировать на сфере (см. рис. 1.27).

Узнав об оригинальной идее Вейля, Эйнштейн был очень заинтригован, но отметил, что с физической точки зрения предложенная Вейлем картина имеет серьезный недостаток, связанный со следующим физическим свойством: *масса* частицы жестко определяет ее собственное время вдоль мировой линии. Это получается (рис. 1.28) путем комбинации квантового соотношения Планка

$$E = h\nu$$

с собственным уравнением Эйнштейна

$$E = mc^2.$$

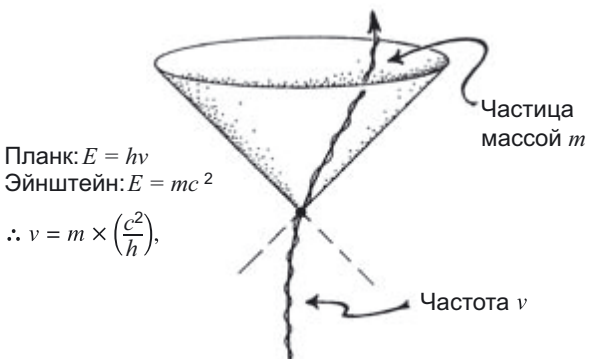


Рис. 1.28. Любая стабильная частица, обладающая массой m, — это точные квантово-механические часы, тикающие с частотой $\nu = mc^2/h$

Здесь E — энергия покоя частицы (в ее собственной покоящейся системе координат), m — масса частицы, а ν — частота (то есть «темп тиканья» частицы), приобретаемая частицей в соответствии с базовыми законами квантовой механики (см. раздел 2.2); h и c — это константы: постоянная Планка и скорость света. Итак, объединив эти тождества согласно уравнению $h\nu\ (= E) = mc^2$, убеждаемся,

что у каждой отдельной частицы всегда есть точная частота, прямо пропорциональная ее массе:

$$\nu = m \times \frac{c^2}{h},$$

причем c^2/h — универсальная константа. Следовательно, масса любой стабильной частицы очень точно определяет ход ее «часов», задаваемый этой частотой.

Однако в гипотезе Вейля любой такой темп часов по определению являлся не фиксированной величиной, а свойством, зависящим от истории частицы. Соответственно, *масса* частицы должна была зависеть от ее истории. В частности, возвращаясь к вышеописанной ситуации, если считать два электрона *идентичными* частицами (чего действительно требует квантовая теория) в момент события P, то, вероятно, при втором событии Q эти электроны уже имели бы разные массы, если бы достигли Q по разным маршрутам. В таком случае, прибыв к Q, они бы уже *не были* идентичны! А это противоречит надежно установленным принципам квантовой теории, требующим, чтобы законы, применимые к идентичным частицам, коренным образом отличались от тех, что применяются к неидентичным.

Таким образом, создавалось впечатление, что идея Вейля противоречит некоторым из основополагающих квантово-механических принципов. Однако в результате невероятного поворота событий сама квантовая теория выручила идею Вейля уже после того, как эту теорию удалось полностью сформулировать (это сделали, прежде всего, Дирак [Dirac, 1930], фон Нейман [von Neumann, 1932] и сам Вейль [Weyl, 1927]). Как мы увидим в главе 2 (см. разделы 2.5 и 2.6), квантовое описание частиц дается в терминах *комплексных чисел* (см. раздел A.9). В разделе 1.4 мы уже познакомились с принципиальной ролью комплексных чисел, выступающих в качестве коэффициентов (величины w и z) в контексте квантово-механического принципа суперпозиции. Далее (см. раздел 2.5) мы узнаем, что если умножить все эти коэффициенты на одно и то же комплексное число u, модуль которого равен единице, то есть $|u| = \sqrt{u\bar{u}} = 1$, что эквивалентно тому, что u лежит на *единичной окружности* в комплексной плоскости (см. раздел A.10, рис. A.13), то с физической точки зрения ситуация останется без изменений. Отметим, что по формуле Котса — де Муавра — Эйлера (см. раздел A.10) такое *унимодулярное* комплексное число u всегда можно записать в виде

$$u = e^{i\theta} = \cos\theta + i\sin\theta,$$

где θ — угол (измеряемый в радианах против часовой стрелки), который линия, соединяющая u с началом координат, образует с положительной действительной осью (см. рис. A.13 в разделе A.10).

В контексте квантовой механики унимодулярный комплексный коэффициент часто именуется *фазой* (или фазовым углом) и с точки зрения квантового формализма трактуется как параметр, недоступный для непосредственного наблюдения (см. раздел 2.5). Неочевидное изменение, превращающее своеобразную, но экстраординарную идею Вейля в ключевую составляющую современной физики, заключается в следующем: заменить вейлевский вещественный положительный масштабный фактор — *калибровку* — на *комплексную фазу* квантовой механики. Именно по этим историческим причинам и прописался в физике термин «*калибровочная*», хотя, пожалуй, видоизмененную таким образом теорию Вейля было бы правильнее именовать *фазовой теорией*, а *калибровочную связность — фазовой связностью*. Однако если сейчас изменить терминологию таким образом, это скорее многих запутает, чем кому-то поможет.

Если быть точным, то та фаза, которая фигурирует в теории Вейля, не вполне тождественна (универсальной) фазе из квантово-механического формализма: две эти фазы различаются коэффициентом, который соответствует *электрическому заряду* рассматриваемой частицы. Важнейшая основополагающая черта теории Вейля заключается в существовании так называемой *непрерывной группы симметрий* (см. раздел A.7, последний абзац), применимой к любому событию P в пространстве-времени. В исходной теории Вейля группа симметрий состоит из всех положительных вещественных коэффициентов, каждый из которых может увеличивать или уменьшать шкалу. Эти возможные коэффициенты суть просто различные *положительные вещественные числа*, пространство которых математики именуют $\mathbb{R}^+$, поэтому интересующая нас в данном случае группа симметрий иногда называется мультипликативной группой $\mathbb{R}^+$. В более поздней версии вейлевской теории электромагнетизма, которая, кроме того, оказалась точнее с физической точки зрения, элементы группы представляют собой вращения в комплексной плоскости (без отражения) под названием SO(2) или иногда U(1), а элементы этой группы выражаются комплексными числами $e^{i\theta}$ с единичным модулем, причем эти элементы дают разные углы вращения единичной окружности в комплексной плоскости, и я буду обозначать такую окружность с радиусом, равным единице, просто S^1.

Следует отметить (см. также последний абзац в разделе A.7, где идет речь об этой нотации, а также о понятии «*группа*»), что O в SO(2) означает «ортогональный» — фактически мы имеем дело с *группой* вращений, в которой сохраняется квадратичная форма от координат. В данном случае вращение происходит в двух измерениях, о чем свидетельствует число 2 в нотации SO(2). S означает «специальный», то есть в данной картине не учитываются *отражения*. Что касается U(1), U здесь означает «унитарный» (сохраняющий присущую комплексным векторам единичную норму) и относится к своеобразному вращению в простран-

стве комплексных чисел, о чем мы поговорим в разделах 2.5–2.8. Независимо от терминологии, нас интересует только вращение обычной окружности S^1, без отражений.

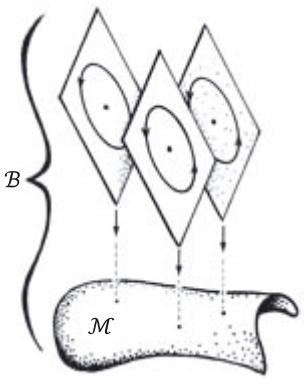


Рис. 1.29. В геометрии Вейля электромагнетизм выражается как связность на расслоении $\mathcal{B}$, наложенном на пространство-время $\mathcal{M}$. (Круглые) слои S^1 наиболее удобно представлять в виде единичных окружностей на копиях комплексной плоскости

Отметим, что связность Вейля — не такой феномен, который применяется к пространственно-временному многообразию $\mathcal{M}$, поскольку окружность S^1, по сути вообще не находится в пространстве-времени. Напротив, S^1 относится к абстрактному пространству, непосредственно связанному с квантовой механикой. Мы по-прежнему можем считать, что S^1 играет и некую геометрическую роль, а именно выступает в качестве *слоя* в расслоении $\mathcal{B}$, базовое пространство которого — это пространственно-временное многообразие $\mathcal{M}$. Такая геометрия показана на рис. 1.29. Слои — это круги S^1, но из рисунка следует, что их лучше всего представлять как единичные окружности, расположенные в копиях комплексной плоскости (Весселя) (см. раздел A.10). (Отсылаю читателя к разделу A.7, где рассматривается феномен расслоения.) Вейлевская концепция калибровочной связности в самом деле является геометрической, но не такой, которая напрямую придает структуру пространству-времени; вместо этого сообщаемая ею структура присваивается расслоению $\mathcal{B}$, которое есть не что иное, как 5-многообразие, тесно связанное с 4-многообразием пространства-времени.

В стандартной модели (см. раздел 1.3) также применяются расширенные формулировки идей Вейля, описывающие сильные и слабые взаимодействия в физике частиц; эти формулировки выражены в терминах калибровочной связности — здесь опять же уместна отсылка к описанию расслоения (см. раздел A.7).

В каждом случае базовым пространством является, как и ранее, четырехмерное пространство-время, но слой должен быть пространством $\mathcal{F}$, обладающим более высокой размерностью, чем одномерная окружность S^1, которое, как было указано выше, может использоваться для описания электромагнетизма. Эти расширения калибровочного подхода Вейля к теории Максвелла именуются *теориями Янга — Миллса* [Chan and Tsou, 1998]. Если говорить о сильных взаимодействиях, то $\mathcal{F}$ будет пространством с такой же симметрией, как и пространство доступных цветов для кварка в соответствии с описанием, приведенным в разделе 1.3. В данном случае применяется группа симметрий, известная под названием SU(2) (или U(2)), но в теории слабых взаимодействий ситуация немного осложняется. Все дело в том, что симметрия считается нарушенной, а процесс, в результате которого такое нарушение произошло, имел место на самых ранних этапах расширения Вселенной. На самом деле с типичным описанием этого процесса сопряжены некоторые проблемы, которые лично меня несколько беспокоят, ведь, строго говоря, сама идея калибровочной симметрии не работает, если не соблюдается абсолютно *точная* симметрия (см. раздел A.7 и раздел 28.3 в книге «ПкР»). По-моему, это большая удача, что можно переформулировать обычную процедуру так, чтобы объяснить возникновение слабого взаимодействия при помощи механизма, имеющего несколько иную физическую интерпретацию. В таких формулировках постулируется существование «цветных» кваркоподобных частиц в составе лептонов (аналогичных обычным кваркам, входящим в состав адронов), и в таком случае симметрия слабого взаимодействия всегда остается ненарушенной ['t Hooft, 1980b; Chan and Tsou, 1980].

1.9. Функциональная свобода в модели Калуцы — Клейна и струнной модели

Итак, существуют *два* альтернативных пятимерных пространства, в каждом из которых возможна геометрическая процедура, позволяющая вписать электромагнетизм Максвелла в геометрию искривленного пространства-времени. Как 5-многообразие $\mathcal{B}$, представленное в виде S^1-расслоения в процедуре Вейля, описанной в разделе 1.8, соотносится с пятимерной картиной электромагнитных взаимодействий в пространстве-времени, построенной Калуцей и Клейном? На самом деле они очень близки друг к другу, и их вполне можно считать *идентичными*! Пятимерное пространство-время Калуцы после модификации, предложенной Клейном, имеет в качестве дополнительного измерения крошечную окружность S^1, и это измерение топологически идентично расслоению $\mathcal{B}$, которое мы получаем в процедуре Клейна. Обычно оба они просто образуют произведение пространств $\mathcal{M} \times S^1$, в состав которого входят обычное

четырехмерное пространство-время $\mathcal{M}$ и окружность S^1 (см. рис. A.25 в разделе A.7 и рис. 1.29). Более того, в пространстве Калуцы — Клейна автоматически присутствует структура, похожая на S^1-расслоение. Чтобы идентифицировать в нем слои S^1, просто ищем *закрытые* геодезические (относящиеся при этом к подходящему топологическому семейству). Однако пятимерные пространства Калуцы — Клейна и Вейля немного различаются по типу структуры, которая присваивается пространству в каждом из случаев. Процедура Вейля требует, чтобы мы присвоили $\mathcal{B}$ *калибровочную связность* (см. раздел 1.8), причем $\mathcal{B}$ понимается как расслоение над четырехмерным пространством-временем $\mathcal{M}$, тогда как в теории Калуцы — Клейна все пятимерное многообразие понимается как пространство-время и, соответственно, *метрика* $\mathbf{g}$ присваивается всей структуре. Однако получается, что калибровочная связность Вейля *уже* неявно присутствует в конструкции Калуцы, поскольку оказывается, что она определяется при помощи обычной *аффинной связности*, рассмотренной в разделе 1.8 (это соблюдается для любого риманового пространства и, следовательно, для пятимерного пространства Калуцы) и применяемой к направлениям, которые ортогональны слоям S^1. Следовательно, пятимерное пространство Калуцы — Клейна уже содержит калибровочную связность Вейля и действительно может быть *идентифицировано* как вейлевское расслоение $\mathcal{B}$.

Однако на самом деле пространство Калуцы — Клейна дает нам нечто *большее*, поскольку обладает метрикой со следующим свойством: если оно удовлетворяет соответствующим эйнштейновским вакуумным уравнениям поля ${}^5\mathbf{G} = \mathbf{0}$ (согласно которым тензор энергии ${}^5\mathbf{T}$ пятимерного пространства считается равным нулю), то мы получаем не только вейлевскую связность, но и, что примечательно, следующий факт: электромагнитное поле Максвелла $\mathbf{F}$, возникающее из вейлевской связности, действует (посредством присущей ему плотности массы/энергии) как источник гравитационного поля. Уравнения, описывающие подобную ситуацию, принято называть *уравнениями Эйнштейна — Максвелла*. Этот поразительный факт *не* выводится напрямую из вейлевского подхода.

Для того чтобы более строго описать структуру пятимерного пространства Калуцы — Клейна, отмечу, что вышеприведенное утверждение верно с одной оговоркой, а именно: та версия теории Калуцы — Клейна, которой я здесь придерживаюсь, требует, чтобы длина у всех петель S^1 в пятимерном пространстве была одинаковой. (В некоторых версиях теории допускается, что эта длина может варьироваться, и таким образом появляется возможность для введения дополнительного скалярного поля.) Я также выдвигаю еще одно требование: эта постоянная длина подбирается такой, чтобы обеспечить правильное значение константы $8\pi\gamma$ в эйнштейновских уравнениях (см. раздел 1.1). Что еще важнее, так это то, что, говоря о теории Калуцы — Клейна, я имею в виду именно исход-

ную версию этой теории, где все пятимерное пространство обязательно обладает точной симметрией и поэтому должно иметь полную *симметрию вращения* относительно направления S^1 (см. по существу аналогичный рис. 1.29). Иными словами, вектор **k** фактически является вектором Киллинга, так что пятимерное пространство может скользить вдоль линий S^1, что не влияет при этом на его метрическую структуру.

Теперь давайте поговорим о проблеме *функциональной свободы* в теории Калуцы — Клейна. Если принять эту теорию в такой форме, какую я только что описал, то дополнительное измерение не приводит к возникновению избыточной функциональной свободы. Поскольку симметрия вращения обязательно должна соблюдаться вдоль кривых S^1, свобода получится такой же, как и в обычном четырехмерном пространстве-времени, для которого характерна стандартная детерминированная эволюция на основе данных из исходного *трехмерного пространства* — фактически такая же, как для уравнений Эйнштейна — Максвелла:

$$\infty^{8\infty^3},$$

причем так и должно быть в соответствии с классической физической теорией, описывающей нашу Вселенную.

Здесь я хочу подчеркнуть, что речь идет о ключевом свойстве калибровочных теорий (а теории этого класса с ошеломительным успехом объясняют основополагающие силы природы), которое заключается в том, что должна существовать (конечномерная) *симметрия*, присущая слою расслоения $\mathcal{F}$, к которому применяется калибровочная теория. Как прямо сказано в разделе A.7, именно наличие (непрерывной) симметрии у нашего слоя $\mathcal{F}$, в принципе, позволяет калибровочной теории работать. В случае вейлевской теории электромагнитных взаимодействий эта симметрия соответствует группе поворотов окружности U(1) (или, что эквивалентно, SO(2)), которая должна точно применяться к слоям $\mathcal{F}$ (значение этих символов объясняется в конце раздела A.7). В подходе Вейля эта симметрия также глобально распространяется на все пятимерное многообразие $\mathcal{B}$; она же была обозначена в исходной процедуре Калуцы — Клейна. Чтобы сохранить эту тесную связь между пространственно-временным подходом, предполагающим наличие высших измерений (предложенным Калуцей) и калибровочным подходом Вейля, кажется необходимым сохранить симметрию слоев и действительно (радикально) повысить функциональную свободу, рассматривая пространства слоев $\mathcal{F}$ как полноценные элементы пространства-времени с внутренними степенями свободы.

А как насчет теории струн? Здесь складывается принципиально иная ситуация, поскольку эта теория явно требует, чтобы дополнительные пространственные

измерения непосредственно участвовали в формировании динамической свободы. Считается, что такие дополнительные пространственные измерения должны играть роль *подлинных* пространственных измерений. Это важнейшая часть философии, лежащей в основе теории струн и способствующей ее развитию, поскольку таким образом предполагается наличие неких «колебаний», обусловленных этими дополнительными измерениями и способных объяснить все сложные силы и параметры, необходимые, чтобы примирить теорию струн со всеми неотъемлемыми свойствами частиц. На мой взгляд, это глубоко ошибочная философия, поскольку, допуская свободное вовлечение *дополнительных* пространственных измерений в общую динамику, мы открываем ящик Пандоры с нежелательными степенями свободы и слабой надеждой на то, что нам удастся удержать их все под контролем.

Тем не менее сторонники теории струн, не обращая внимания на подобные сложности, которые непременно возникали бы из-за избыточной функциональной свободы в дополнительных пространственных измерениях, пошли совершенно иным путем по сравнению с предложенным в исходной картине Калуцы — Клейна. Пытаясь устранить аномалии, возникающие в силу требований инвариантности параметров в квантовой теории струн, теоретики уже примерно с 1970 года пытались приспособить к бозонным струнам полностью динамическое 26-мерное пространство-время, где было бы 25 пространственных измерений и одно временное. Затем вслед за крайне влиятельными теоретическими разработками Майкла Грина и Джона Шварца, выполненными в 1984 году, теоретики-струнники смогли понизить число измерений до девяти (для фермионных струн) при помощи так называемой *суперсимметрии* (см. раздел 1.14; она уже упоминалась в разделе 1.6), но такое сокращение числа дополнительных пространственных измерений практически ничего не значит для решения тех вопросов, которые я собираюсь освещать, поскольку оно не снижает пространственную размерность до трех непосредственно воспринимаемых измерений.

Пытаясь разобраться с различными наработками в теории струн, я усматривал в ней еще один запутанный аспект, когда пробовал понять проблемы функциональной свободы. Именно при изучении этих проблем зачастую пересматриваются представления о том, чем же является размерность пространства-времени. Полагаю, многие из тех, кто не в теме, также сталкивались с подобными трудностями, пытаясь понять математическую структуру теории струн. Идея о наличии окружающего пространства, обладающего конкретной размерностью, по-видимому, играет в теории струн не столь важную роль, как в традиционной физике, причем настолько не важную, что лично меня это не устраивает. Особенно сложно оценивать функциональную свободу, принимаемую в физической теории, если не иметь четкого представления, о какой именно размерности пространства-времени идет речь.

Для того чтобы пояснить эту проблему, позвольте мне вернуться к одному из наиболее привлекательных аспектов ранних струнных идей, которые были описаны в разделе 1.6. Тогда мы говорили, что струнные истории могут рассматриваться как *римановы поверхности*, то есть как комплексные кривые (см. раздел А.10), которые с математической точки зрения довольно красивы. Иногда такая струнная история обозначается термином *мировой лист* (аналогично идее о мировой линии частицы в традиционной теории относительности; см. раздел 1.7). На заре теории струн эта тема иногда рассматривалась с точки зрения двумерной *конформной теории поля* [Francesco et al., 1997; Kaku, 2000; Polchinski, 1994, глава 1, 2001, глава 2], причем, грубо говоря, аналог пространства-времени в такой теории как раз и *представлял бы собой* двумерный мировой лист! (Вспомните из раздела 1.7, что означает термин *конформный* в контексте пространства-времени.) Так мы получили картину, в которой функциональная свобода выражается следующим образом:

$$\infty^{a\infty^1},$$

где a — некоторое положительное число. Как увязать это с гораздо большей функциональной свободой $\infty^{b\infty^3}$, требуемой в традиционной физике?

По-видимому, ответ таков: мировой лист будет в некотором смысле «прощупывать» окружающее пространство и его физику в терминах некоторого разложения в степенной ряд, причем необходимая информация (фактические коэффициенты степенного ряда) будет предоставляться в терминах бесконечного числа параметров (по сути, это будут голоморфные величины на мировом листе; см. раздел А.10). В упрощенном виде такое бесконечное число параметров можно передать, подставив в вышеприведенное выражение $a = \infty$, но такой вариант будет довольно бесполезен (почему — я расскажу ближе к концу раздела А.11). В данном случае я говорю отнюдь не о том, что функциональная свобода может быть в каком-то смысле плохо определена или нерелевантна. Суть, однако, заключается в следующем: если теория сформулирована в терминах коэффициентов степенных рядов или анализа мод, то довольно непросто достоверно узнать, какова же именно функциональная свобода (см. раздел А.11). К сожалению, формулировки такого рода часто используются в различных подходах к теории струн.

Возможно, физики-теоретики считают, что не так уж и важно четко понимать, что же именно представляет собой размерность пространства-времени. В некотором смысле эту размерность можно было бы считать энергетическим эффектом, поэтому допустимо было бы предположить, что, если энергия системы возрастает, в ней открывается доступ к дополнительным пространственным измерениям. Соответственно, можно полагать, что существуют скрытые измерения и некоторые из них могут обнаружиться, когда энергия возрастет. Неоднозначность этой кар-

тины меня несколько беспокоит, особенно в связи с вопросом о функциональной свободе, который для данной теории является неотъемлемым.

Этот вопрос встает ребром в случае с *гетеротической* теорией струн. Она существует в двух версиях — *теория типа HO* и *теория типа HE1* (заглавные буквы — латинские). Разница между двумя этими теориями сейчас непринципиальна, но чуть ниже я кое-что об этом расскажу. Странная черта гетеротической теории струн заключается в том, что она одновременно проявляет свойства и 26-мерной, и 10-мерной теории струн (второй вариант сопровождается суперсимметрией) в зависимости от того, идет ли речь о левоориентированных или правоориентированных возбуждениях струны. Эта разница (зависящая от ориентации, которую требуется придать струне) также требует объяснения, к которому я вскоре перейду. По-видимому, такой конфликт размерностей чреват проблемами, если мы попытаемся определить функциональную свободу для данного случая (чтобы это сделать, каждую из двух схем нужно трактовать как классическую теорию).

Возникает явная путаница, для устранения которой официально предлагается считать пространство-время десятимерным в *обоих* случаях (одно временное измерение и девять пространственных). Если говорить о левосторонних возбуждениях, то необходимо взять сразу все 28 измерений вместе и считать, что именно они образуют пространство, в котором может колебаться струна. Однако в случае с правосторонними возбуждениями различные направления в рамках 26 измерений интерпретируются по-разному. Считается, что десять из них дают направления, в которых может колебаться струна, а остальные 16 являются направлениями *слоев*. Поэтому, рассуждая о вибрации струны в таких правоориентированных модах, мы имеем в виду вибрацию расслоения (см. раздел A.7) с 10-мерным базовым пространством и 16-мерным слоем.

Что касается расслоений вообще, то со слоем должна быть ассоциирована *группа симметрий*, и в теории типа HO таковой считается SO(32) (группа вращения сферы без отражений в 32 измерениях; см. в конце раздела A.7). В гетеротической теории типа HE такую роль играет группа $E_8 \times E_8$, где E_8 — группа симметрий, тип которой особенно интересен с математической точки зрения; она именуется «*наибольшей особой простой группой*». Несомненно, что чисто математический интерес, присущий этой исключительно простой группе — E_8 является самой крупной и наиболее интересной из подобных групп, — отчасти объясняется и эстетической привлекательностью (см. раздел 1.1). Однако с функциональной свободой связана важная проблема, которая заключается в следующем: какая бы группа симме-

[1] Современная классификация содержит шесть типов теорий струн: 1) бозонная; 2) тип I; 3) тип IIA; 4) тип IIB; 5) тип HO и 6) тип HE. — *Примеч. ред.*

трий ни применялась для описания слоя, эта свобода будет описываться выражением вида $\infty^{a\infty^9}$, приемлемым для фермионных (правоориентированных) мод колебаний струн, но будет казаться, что она описывается выражением $\infty^{b\infty^{25}}$, которое применяется для описания бозонных (левоориентированных) мод. Эта проблема тесно связана с той, которую мы рассматривали выше, говоря о различиях функциональной свободы в исходной теории Калуцы — Клейна (или в вейлевской теории круговых расслоений) с функциональной свободой $\infty^{8\infty^3}$ и в теории о полноценном пятимерном пространстве-времени, где наблюдалась бы гораздо большая функциональная свобода, описываемая выражением $\infty^{b\infty^4}$. Здесь необходимо четко различать размерность $d + r$ в общем пространстве $\mathcal{B}$ расслоения (с r-мерным слоем $\mathcal{F}$) и размерность d-мерного базового пространства $\mathcal{M}$. Эти вопросы подробнее описаны в разделе A.7.

Вышеизложенная проблема касается функциональной свободы, присущей пространству-времени в целом, и эта свобода совершенно не зависит от того, какой струнный мировой лист мог бы присутствовать в этом пространстве. Однако в данном контексте нас интересует как раз та свобода, которой могут обладать струнные мировые листы (см. раздел 1.6), расположенные в данном пространстве-времени. Как такое возможно, что при некоторых (фермионных) вариантах смещения пространство-время кажется 10-мерным, тогда как при других вариантах (бозонных) оно представляется 26-мерным? В случае с бозонными модами картина относительно проста. Струна может колебаться в окружающем пространстве-времени с функциональной свободой $\infty^{24\infty^1}$ (значение 1 присутствует здесь потому, что, хотя струнный лист и является двумерной плоскостью, мы рассматриваем лишь одномерное пространство правоориентированных возбуждений). Но когда мы начинаем рассматривать фермионные моды, мы считаем, что струна располагается в 10-мерном пространстве-времени, а не в расположенном над ним 26-мерном расслоении. Это означает, что сама струна должна *подчиняться* движениям слоя из того расслоения, что лежит над ней. Этот феномен серьезно отличается от бозонных мод, поскольку теперь сама струна превращается в 18-мерное *подрасслоение* 26-мерного общего пространства в рамках того пространства-времени, где она (струна) располагается. (Этот факт обычно упускается теоретиками. Эффективное пространство-время является 10-мерным *факторным* пространством (см. рис. 1.32 в разделе 1.10 и раздел A.7) 26-мерного расслоения, поэтому и струнный мировой лист также должен быть факторным пространством, на сей раз относящимся к 18-мерному подрасслоению). Функциональная свобода этих мод также выражается в форме $\infty^{a\infty^1}$ (где a зависит от группы расслоения), но *геометрическая картина* теперь совершенно отличается от той, что наблюдалась с бозонными модами, поскольку при бозонных модах струна считается двумерной мировой трубкой (как на рис. 1.11),

но в случае фермионных мод струна технически должна представлять собой подрасслоение с общей размерностью 18 (= 2 + 16). Мне очень сложно построить непротиворечивую картину всех этих процессов, и я пока не встречал нормального обсуждения этих геометрических проблем.

К тому же здесь следует более предметно рассказать о геометрической природе левоориентированных и правоориентированных мод, безотносительно того, как должно трактоваться окружающее пространство-время, поскольку здесь перед нами встает еще одна проблема, которой я пока не касался. Как мы уже говорили, струнные мировые листы можно считать *римановыми поверхностями*. Однако это не вполне верно для вышеприведенных описаний. Я пошел на одну уловку, которая довольно часто применяется в дискуссиях о квантовой теории поля и открыто используется здесь. Речь о том, что я прибегаю к одному методу, который называется *поворот Вика* и о котором я до сих пор в явном виде не упоминал.

Что такое поворот Вика? Это математическая процедура, исходно связанная с преобразованием различных задач квантовой теории поля в пространстве-времени Минковского $\mathbb{M}$ (это плоское пространство-время специальной теории относительности; см. конец раздела 1.7) в гораздо более удобоваримые, решаемые в обычном евклидовом четырехмерном пространстве $\mathbb{E}^4$. Данная идея связана с тем фактом, что лоренцева метрика пространства-времени $\mathbf{g}$ в теории относительности превращается в псевдоевклидову, если заменить стандартную координату времени t на it (где $i = \sqrt{-1}$, см. раздел A.9). Такой фокус иногда называют *евклидизация*, и когда задача решается в евклидовой форме, ее обратное преобразование происходит путем *аналитического продолжения* (см. раздел A.10, а также раздел 3.8); так получается решение в целевом пространстве-времени Минковского $\mathbb{M}$. Идея поворота Вика сегодня так часто используется в квантовой теории поля, что в самых различных ситуациях ее применяют просто на автомате, упоминая лишь вскользь, а правомерность ее применения едва ли когда-нибудь оспаривалась. Действительно, область применения этой процедуры широка, но она *не является* универсальной. Особенно важно, что применение ее весьма сомнительно в контексте *искривленных* конфигураций пространства-времени, возникающих в общей теории относительности. В таком случае процедуру, как правило, просто нельзя применить, так как отсутствует *естественная координата времени*. В теории струн такая проблема существует как в 10-мерном пространстве-времени (общие ситуации с искривленным пространством), так и на уровне струнного мирового листа[1].

[1] Интересный вариант поворота Вика используется в методе Хартла — Хокинга по квантованию времени [Hartle and Hawking, 1983]. Однако, строго говоря, это совершенно другая процедура со своими характерными проблемами.

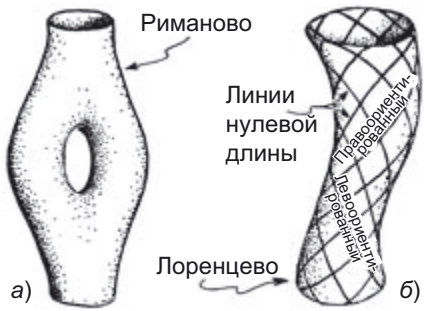


Рис. 1.30. Два разных взгляда на струнные мировые листы. На рис. *а* представлены мировые трубки, являющиеся римановыми поверхностями, которые могут ветвиться и плавно соединяться различными способами. На рис. *б* приведено изображение (времениподобное) струнной истории как лоренцева 2-многообразия, где можно показать левоориентированные и правоориентированные моды возбуждения, но ветвление не допускается. Связь между двумя изображениями прослеживается через поворот Вика, но такая операция весьма сомнительна в контексте общей теории относительности с искривленным пространством-временем

Мне кажется, что подобные сложности ставят перед нами задачи, которые никто пока не пытался решить в рамках теории струн. Однако позвольте мне проигнорировать здесь подобные моменты; вместо этого давайте рассмотрим, каким образом евклидизация отразится на струнном мировом листе. Струнную историю можно визуализировать в виде цельной петли, которая каким-то образом движется, не превышая локальной скорости света. В таком случае мировой лист будет времениподобной двумерной плоскостью, которая унаследует лоренцеву 2-метрику от лоренцевой 10-метрики окружающего пространства-времени. Такая 2-метрика будет присваивать пару нулевых направлений каждой точке на мировом листе. Если аккуратно прослеживать эти направления в одну или в другую сторону, то мы получим правоориентированную или левоориентированную спиральную кривую на цилиндрическом мировом листе. Возбуждения, постоянно возникающие вдоль одного или другого из этих семейств кривых, дадут левоориентированные или правоориентированные моды, о которых шла речь ранее (см. рис. 1.30 *б*). Однако такие цилиндрические листы никогда не смогут ветвиться и не дадут нам такой картины, которая показана на рис. 1.11, поскольку лоренцева структура нарушается в тех местах, где трубка ветвится. Такая топология может возникать лишь для евклидизированных струн, показанных на рис. 1.30 *а* и представляющих собой римановы поверхности, которые обладают метрикой риманова типа без нулевых направлений и могут интерпретироваться в качестве *комплексных кривых* (см. раздел А.10). В таком случае евклидизированные право- и левоориентированные моды приравниваются

соответственно к *голоморфным* и *антиголоморфным* функциям на римановой поверхности (см. раздел А.10).

Проблема *функциональной свободы*, которую я обсуждаю в этом разделе, — не просто физическая сложность, которой, по-видимому, никто всерьез не занимался в рамках теории струн, по крайней мере, я не нашел такого обсуждения в стандартной литературе. Более того, я не нашел и никаких материалов, которые касались бы непосредственных геометрических вопросов, возникающих в связи с якобы принципиальной, но на деле крайне неоднозначной процедурой поворота Вика, упомянутой выше. У меня сложилось впечатление, что большинство из наиболее очевидных физических и геометрических проблем, связанных с теорией струн, вообще никогда как следует не обсуждались!

Возьмем, к примеру, гетеротическую теорию струн: все струны считаются *замкнутыми*, то есть не имеющими разрывов (см. раздел 1.6 и в особенности раздел 1.16). Если мы попытаемся представить эти струны в непосредственном физическом смысле, то есть до того, как с ними будет проделан «волшебный» поворот Вика, то должны будем считать мировой лист *времениподобным*, как на рис. 1.30 *б*. Мировой лист не должен иметь отверстий, и он так и должен оставаться времениподобной трубкой, уходящей в будущее. Продолжаться в будущее он может до бесконечности, а в таком случае не может считаться по-настоящему замкнутым. Это один из многих вопросов, адекватного рассмотрения которых я не нашел в каком-либо из известных мне материалов по теории струн.

Столь любопытное отсутствие внятного геометрического представления теории струн в обычных физических терминах кажется мне очень странным, особенно учитывая, что эта теория нередко именуется *теорией всего*. Более того, такое отсутствие четкой геометрической и физической картины при подобных рассуждениях резко контрастирует с крайне изощренной геометрией и тщательными математическими выкладками, применяемыми при изучении настоящих 6-многообразий (как правило, речь идет о пространствах Калаби — Яу: см. разделы 1.13 и 1.14). Считается, что такие многообразия могут дать искомые свернутые дополнительные шесть пространственных измерений, заметных в планковских масштабах и предположительно необходимых для непротиворечивой формулировки теории струн. Мне кажется невероятно странным, что исключительно сведущая фракция в лице физического сообщества, с одной стороны, принимает столь исключительно утонченную геометрию, а с другой — не обращает внимания на явное пренебрежение общей геометрической согласованностью!

В тех выкладках о функциональной свободе, которым посвящены следующие два раздела, я исхожу из того, что пространство-время десятимерно, но мои аргументы применимы не только к данному конкретному числу измерений.

Классический аргумент из раздела 1.11 о том, что такие дополнительные измерения были бы исключительно нестабильны, актуален для любой «многомерной» теории, в которой имеется как минимум два дополнительных (микроскопических) пространственных измерения и которая подчиняется десятимерным (девять пространственных и одно временное измерение) эйнштейновским вакуумным уравнениям $^{10}\mathbf{G} = \mathbf{0}$ ($\Lambda = 10$). В стандартной литературе высказываются аргументы относительно того, что исходная 26-мерная бозонная теория струн действительно катастрофически нестабильна, но это не особенно важно для тех аргументов, что я привожу здесь (они более общие).

Аргумент из раздела 1.10 коренным образом отличается от аргумента, который я привожу в разделе 1.9 и который, в свою очередь, направлен против распространенного квантово-механического утверждения, что дополнительные пространственные измерения будет невозможно возбудить при каких-либо хотя бы отдаленно достижимых энергиях. Опять же этот аргумент актуален независимо от конкретного числа дополнительных пространственных измерений, но для определенности я обсуждаю его в контексте десятимерной теории, популярной в настоящее время. Ни в первом, ни во втором случае я не буду затрагивать проблемы суперсимметрии, поэтому геометрические построения останутся довольно однозначными. Полагаю, что наличие суперсимметрии не изменило бы эти аргументы коренным образом, поскольку они позволяют рассмотреть несуперсимметричную геометрическую «основу» (см. раздел 1.14).

Во всех этих аргументах я придерживаюсь точки зрения, которая кажется обязательной для теории струн, а именно: предполагается, что пространственные измерения полностью динамические. Следовательно, несмотря на сходство между дополнительными измерениями в теории струн и тем дополнительным измерением, которое ввели Калуца и Клейн (струнные теоретики часто подчеркивают это сходство), я должен вновь заострить внимание на огромной и существенной разнице между исходной картиной Калуцы — Клейна и той гипотезой, которую имеют в виду струнники. Во всех известных мне версиях теории струн с дополнительными измерениями, кроме, пожалуй, варианта с 16 несогласованными измерениями в гетеротических моделях, описанных в этом разделе ранее, фактически не существует ничего, аналогичного *симметрии вращения* в дополнительных измерениях, которая была принята в теории Калуцы — Клейна. Более того, такая симметрия даже открыто отрицается [Greene, 1999]. Соответственно, функциональная свобода в теории струн представляется избыточной. Если быть точным, в десятимерной теории, которая популярна сегодня, функциональная свобода соответствует $\infty^{k\infty^9}$, а не $\infty^{k\infty^3}$, как следовало бы ожидать в реалистичной физической теории. Основной момент заключается в том, что, тогда как в теории Калуцы — Клейна мы *не можем* свободно получать

произвольные структурные вариации вдоль дополнительного пространственного измерения S^1 (это связано с обязательной симметрией вращения), в теории струн такая свобода открыто допускается. В этом и заключается причина, по которой в теории струн возможна избыточная функциональная свобода.

Насколько мне известно, эта проблема, касающаяся классических (то есть неквантовых) вопросов, никогда всерьез не рассматривалась профессиональными струнниками. В то же время высказывалось мнение, что подобные моменты, в сущности, не важны для теории струн, поскольку их нужно рассматривать с точки зрения квантовой механики (или квантовой теории поля), а не классической теории поля. Действительно, когда пытаешься обсудить со струнными теоретиками вопрос об избыточной функциональной свободе в дополнительных шести «малых» измерениях, от него зачастую пытаются отмахнуться, прибегая к довольно общему квантово-механическому аргументу, который кажется мне порочным по сути. Я расскажу об этом аргументе в следующем разделе, а потом (в разделе 1.11) попробую самостоятельно показать полную его необоснованность; более того, логика присутствия тех дополнительных пространственных измерений, о которых говорят струнные теоретики, предполагает существование совершенно нестабильной Вселенной, где эти измерения могли бы динамически схлопываться с катастрофическими последствиями для привычной нам макроскопической геометрии пространства-времени.

Эти аргументы связаны в первую очередь со степенями свободы в геометрии пространства-времени как таковой. Существует отдельная проблема избыточной функциональной свободы в других полях, определяемых в пространственно-временных многообразиях с дополнительными измерениями, но эта проблема тесно связана с нашей. Я вкратце коснусь этих вопросов ближе к концу раздела 1.10, где уделю особое внимание экспериментальным ситуациям, которые порой считаются важными в данном контексте. В разделе 2.11 я упомяну о другой смежной проблеме. Возможно, мои опасения могут показаться не вполне убедительными, эта проблема меня весьма беспокоит; несмотря на то что я больше нигде ее не упоминаю, она тем не менее заслуживает дальнейшего изучения.

1.10. Квантовые препятствия для функциональной свободы

В этом (и следующем) разделе я привожу аргументы, которые, на мой взгляд, очень убедительно доказывают, что мы не можем избавиться от проблемы избыточной функциональной свободы в теориях с дополнительными пространственными измерениями — даже в квантово-механическом контексте. В сущ-

ности, именно этот аргумент я представил в январе 2002 года на кембриджской конференции, устроенной по случаю 60-летия Стивена Хокинга [Penrose, 2003; см. также ПкР, разделы 31.11 и 31.12]. Здесь я изложу свои соображения в более акцентированном виде. Для того чтобы понять насущные квантовые проблемы в том виде, как они обычно подаются, нужно более глубоко рассмотреть стандартные методы квантовой теории.

Рассмотрим простую квантовую систему, такую как покоящийся атом (например, атом водорода). В принципе, мы обнаружим, что в атоме будет некоторое количество четко разграниченных энергетических уровней (так, в атоме водорода это будет некоторое количество орбит, на которых может находиться электрон). Кроме того, будет состояние с *минимальной* энергией — так называемое *основное состояние*. Предполагается, что если атом находится в любом другом стационарном состоянии, в котором обладает большей энергией, то рано ли поздно он перейдет в основное состояние, так как будет испускать фотоны, — это произойдет при условии, что окружающая атом среда не слишком горячая (то есть не высокоэнергетическая). В некоторых ситуациях могут действовать правила отбора, препятствующие некоторым из таких переходов, но это не влияет на общий ход рассуждений. Напротив, если атому доступно достаточно много энергии извне (как правило, речь идет об электромагнитной энергии, так называемой *фотонной бане*, то есть вновь о фотонах в данном квантово-механическом контексте) и эта энергия передается атому, то атом может перейти из низкоэнергетического (базового) состояния в высокоэнергетическое. В каждом случае энергия E любого фотона, участвующего в процессе, будет связана с конкретной частотой ν согласно формуле Планка: $E = h\nu$ (см. разделы 1.5, 1.8, 2.2 и 3.4).

Теперь давайте вернемся к рассматриваемым в теории струн вариантам пространства-времени с дополнительными измерениями. По-видимому, среди теоретиков царит всеобщее самодовольное убеждение, что по какой-то причине функциональная свобода (огромная!), заключенная в дополнительных измерениях, никогда не проявляется в обычных обстоятельствах. Скорее всего, все дело в следующей точке зрения: такие степени свободы, *деформирующие* геометрию дополнительных шести малых измерений, практически невозможно перевести в возбужденное состояние, так как на возбуждение этих дополнительных степеней свободы потребовалась бы колоссальная энергия.

На самом деле бывают такие особые деформации дополнительных пространственных измерений, которые можно было бы возбудить *вообще* без всякой энергии. Это справедливо для десятимерного пространства-времени, когда шесть дополнительных измерений считаются пространствами Калаби — Яу (см. разделы 1.13 и 1.14). Такие деформации называются *нулевыми колебани-*

ями (или *нулевыми модами*), и с ними связаны серьезные проблемы, которые не вызывают энтузиазма у теоретиков-струнников. Нулевые моды не требуют избыточной функциональной свободы, которую мы здесь обсуждаем, и о них мы поговорим далее — в разделе 1.16. В этом и в следующем разделах речь пойдет о таких деформациях, которые действительно открывают доступ к полной чрезмерной функциональной свободе и для возбуждения которых требуются значительные энергии.

Чтобы оценить масштабы энергий, которые для этого требуются, обратимся к планковской формуле $E = h\nu$; при этом мы не сильно погрешим против истины, если возьмем за частоту ν величину, обратную тому времени, которое потребовалось бы свету для распространения в этих дополнительных измерениях. В таком случае «размер» этих маленьких дополнительных измерений зависит от того, какую именно версию теории струн мы рассматриваем. В исходной 26-мерной теории мы могли бы говорить о значениях порядка 10^{-15} м, и в таком случае потребовалась бы примерно такая энергия, которая генерируется на БАК (см. раздел 1.1). Напротив, если говорить о более новых десятимерных суперсимметричных теориях струн, то необходимая энергия была бы гораздо больше — далеко за пределами возможностей самого мощного ускорителя частиц на Земле или любого другого ускорителя частиц, который можно было бы сегодня построить. В теории струн такого типа — той, что всерьез пытается решать проблемы квантовой *гравитации*, речь шла бы об энергиях примерно *планковского* масштаба, ассоциированных с *планковской* длиной (эти величины кратко обсуждаются в разделах 1.1 и 1.5, а более подробно — в разделах 3.6 и 3.10). Соответственно, обычно считается, что нужен какой-то процесс, при котором *отдельные* частицы ускорялись бы до столь высоких энергий — соответствующих энергии взрыва крупного артиллерийского снаряда, — и тогда можно было бы возбудить степени свободы, заключенные в дополнительных измерениях, выведя их из основного состояния. Как минимум в тех версиях теории струн, которые оперируют столь крошечными дополнительными измерениями, считается, что такие измерения должны быть фактически неуязвимы для любых попыток перевести их в возбужденное состояние, — по меньшей мере в обозримом будущем.

Попутно можно отметить, что *существуют* и такие версии теории струн, обычно считающиеся маргинальными, в которых некоторые дополнительные измерения могут быть и довольно крупными — миллиметрового масштаба. Предполагается, что такие схемы хороши тем, что их можно было бы проверить путем наблюдения [Arkani-Hamed et al., 1998]. Однако с точки зрения функциональной свободы они связаны с совершенно конкретной проблемой: возбудить такие «большие» колебательные энергии должно быть легко даже на современных ускорителях, поэтому мне совершенно непонятно, почему это не смущает сторонников таких

теорий — ведь, согласно их выкладкам, мы уже должны были бы зафиксировать такие радикальные всплески функциональной свободы.

Лично я считаю полностью неубедительным аргумент, что возбудить функциональную свободу дополнительных пространственных измерений совершенно невозможно даже в *планковских* масштабах. Соответственно, не могу всерьез воспринимать и общее утверждение, что бездна функциональной свободы, скрытая в дополнительных измерениях, должна быть полностью неуязвима для возбуждений в «обычном» состоянии, то есть на уровне энергий, доступных в нынешней Вселенной. У меня есть причины для стооль серьезного скепсиса. Во-первых, хочу спросить: почему мы должны считать планковскую энергию в данном контексте «большой»? Полагаю, выстраивается такая картина: энергия должна поступать в систему под действием какой-то высокоэнергетической частицы, что можно было бы наблюдать в ускорителе частиц (аналогичная ситуация возникает, когда фотон попадает в атом и выводит его из основного состояния). Однако необходимо помнить, что пространство-время в картине, выстраиваемой теоретиками-струнниками — пока дополнительные измерения находятся в основном состоянии, — понимается как произведение пространств $\mathcal{M} \times \mathcal{X}$ (см. рис. А.25 в разделе А.7), где $\mathcal{M}$ напоминает наши «классические» представления о четырехмерном пространстве-времени, а $\mathcal{X}$ — пространство с дополнительными «малыми» измерениями. В десятимерной версии теории струн за $\mathcal{X}$ обычно принимается пространство Калаби — Яу— особое 6-многообразие, о котором мы немного подробнее поговорим в разделах 1.13 и 1.14. Если бы потребовалось возбуждать дополнительные измерения как таковые, то соответствующая «возбужденная мода» (см. раздел А.11) пространства-времени проявлялась бы в виде нашего пространства-времени, но с дополнительными измерениями (вида $\mathcal{M} \times \mathcal{X}'$), где $\mathcal{X}'$ — выведенная из равновесия (то есть возбужденная) система дополнительных измерений. (Разумеется, мы должны считать, что $\mathcal{X}'$ каким-то образом располагается в «квантовом» пространстве, а не в классическом, однако это не оказывает серьезного влияния на нашу дискуссию.) Обратите внимание, что, возбуждая $\mathcal{M} \times \mathcal{X}$ до состояния $\mathcal{M} \times \mathcal{X}'$, мы должны были бы проделать это со всей Вселенной (со всем пространством $\mathcal{M}$ в каждой точке $\mathcal{X}$). Поэтому, полагая, что энергия для такой пертурбации «велика», мы должны сравнить эту энергию с масштабами всей Вселенной. Мне кажется, было бы совершенно неоправданно требовать, чтобы такой квант энергии обязательно сообщался какой-то строго локальной высокоэнергетической частицей.

Более уместно было бы рассмотреть некий вариант предположительно *нелинейной* (ср. разделы А.11 и 2.4) нестабильности, влияющей на динамику Вселенной (с дополнительными измерениями) в целом. Здесь я должен четко сказать, что *не считаю* динамику «внутренних» степеней свободы, управляющую свойствами

шести дополнительных пространственных измерений, *независимой* от динамики «внешних», задающих поведение всем известного четырехмерного пространства-времени. Для того чтобы и те и другие можно было обоснованно считать компонентами реального общего «пространства-времени», должна существовать динамика, управляющая обеими совокупностями степеней свободы в рамках общей схемы (а не такой, где одна подобная совокупность считается «расслоением» поверх другой; см. разделы A.7 и 1.9). Некоторые версии эйнштейновских уравнений описывают эволюцию сразу обеих совокупностей степеней функциональной свободы. Думаю, как раз такую картину и представляют себе теоретики-струнники — как минимум на «классическом» уровне, считая, что эволюция всего десятимерного пространства-времени хорошо аппроксимируется десятимерными эйнштейновскими вакуумными уравнениями $^{10}\mathbf{G} = \mathbf{0}$ (см. следующий раздел — 1.11).

Я затрону проблему *классических* нестабильностей в разделе 1.11; сейчас же речь пойдет о *квантовых* проблемах, сухой остаток которых таков: действительно, нужно изучить картину классической физики, чтобы всерьез разобраться с проблемой стабильности. В контексте динамики всей Вселенной планковская энергия отнюдь не велика — она исчезающе мала. Например, кинетическая энергия вращения Земли вокруг Солнца примерно в миллион миллионов миллионов миллионов (то есть в 10^{24}) раз больше. Я не вижу причин, по которым какая-то крошечная часть этой энергии — возможно, намного выше планковской — не спровоцировала бы в пространстве $\mathcal{X}$ каких-нибудь минимальных изменений на участке пространства $\mathcal{M}'$ в масштабах Земли — или, например, в немного более обширном пространстве, куда поместилась бы вся система Солнце — Земля. *Плотность* этой энергии, распределенной во всем пространстве $\mathcal{M}'$, должна быть крайне мала (рис. 1.31). Соответственно, фактическая геометрия этих дополнительных пространственных измерений ($\mathcal{X}$) едва ли вообще изменилась бы во всем пространстве $\mathcal{M}'$ при возмущении, сопоставимом с планковской энергией, и я не вижу никаких причин, по которым наша сугубо локальная пространственная геометрия $\mathcal{M}' \times \mathcal{X}$ должна не исказиться до условного $\mathcal{M} \times \mathcal{X}'$, а мягко примкнуть к остальному пространству $\mathcal{M} \times \mathcal{X}$ за пределами области $\mathcal{M}'$, где разница между *геометриями* $\mathcal{X}$ и $\mathcal{X}'$ может быть исчезающе малой — намного *меньше* планковских масштабов.

Уравнения, описывающие все десятимерное пространство целиком, динамически сочетали бы уравнения для $\mathcal{M}$ и уравнения для $\mathcal{X}$, поэтому крошечное локальное изменение геометрии $\mathcal{X}$ ожидаемо могло бы произойти при сугубо локальном возмущении геометрии макроскопического пространства-времени $\mathcal{M}$ (в районе $\mathcal{M}'$). Более того, такая связь была бы взаимной. Соответственно, из-за высвобождения бесчисленных степеней функциональной свободы, потенци-

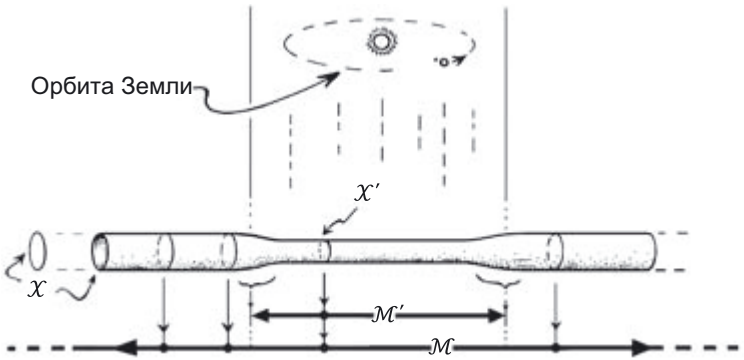


Рис. 1.31. Энергия порядка планковского масштаба потребовалась бы для возбуждения крошечного шестимерного компактифицированного дополнительного пространства $\mathcal{X}$, но даже при вращении Земли вокруг Солнца речь идет о гораздо более высокой энергии. Здесь $\mathcal{M}$ означает обычное четырехмерное пространство-время, воспринимаемое нами, а $\mathcal{M}'$ — сравнительно небольшую часть пространства, ограниченную орбитой Земли. Даже крошечная часть того возмущения пространства-времени, которое могла бы спровоцировать энергия вращения Земли, была бы достаточной, чтобы исказить $\mathcal{X}$ и превратить его в немного иное пространство $\mathcal{X}'$, занимающее область $\mathcal{M}'$

ально существующих благодаря свободе, заложенной в геометрии планковских масштабов — что, кстати, сопровождалось бы колоссальными искривлениями пространства-времени, — макроскопическая динамика подверглась бы разрушительным изменениям.

Хотя суперсимметрия и предполагает, что геометрия пространства $\mathcal{X}$ в основном состоянии должна быть строго ограничена (поэтому и возникает необходимость в существовании так называемого *шестимерного пространства Калаби — Яу*, см. разделы 1.13 и 1.14), это *не должно* сказываться на возможности пространства менять такую геометрию в динамике. Например, хотя уравнения Эйнштейна $^{10}\mathbf{G} = \mathbf{0}$ применительно к геометриям, ограниченным произведениями пространств $\mathcal{M} \times \mathcal{X}$, вполне могут быть постулированы строгие условия для геометрии пространства $\mathcal{X}$ как таковой (а также для геометрии $\mathcal{M}$), это очень специфическое произведение пространств не должно устойчиво сохраняться в динамической ситуации, и *почти всю* функциональную свободу можно было бы выразить в виде решений, не являющихся такими произведениями (см. раздел A.11). Соответственно, какие бы критерии ни использовались для ограничения дополнительных измерений и для подгонки их под конкретную геометрическую структуру (например, Калаби — Яу) в базовом состоянии, мы не можем ожидать, что подобная структура поддерживалась бы в полностью динамических ситуациях.

На данном этапе следует дать некоторые пояснения относительно сравнений с атомными квантовыми переходами, которые я приводил ранее, поскольку в начале этого раздела я увильнул от обсуждения технических сложностей, так как рассматривал атом в состоянии покоя. Чтобы атом действительно находился в покое, его состояние (волновая функция) должно быть равномерно распределено во всей Вселенной (поскольку состояние абсолютного покоя подразумевает нулевой импульс, из чего и следует однородность; см. разделы 2.13 и 4.2) примерно по такому же принципу, как $\mathcal{X}$ (или $\mathcal{X}'$) равномерно распределяется в произведении пространств $\mathcal{M} \times \mathcal{X}$. Опровергает ли это каким-либо образом вышеизложенные рассуждения? Не вижу на то причин. Как бы то ни было, процессы, происходящие с отдельным атомом, должны трактоваться как *локализованные* события, где изменение состояния атома будет выражаться в виде некоего локального процесса — например, контакта с другой условно локальной сущностью, скажем, с фотоном. Тот факт, что стационарное состояние (или не зависящая от времени волновая функция) атома технически должно считаться распределенным по всей Вселенной, совершенно не важен для самих вычислений, поскольку естественно рассматривать все пространственные соотношения относительно *центра масс* системы — в таком случае вышеупомянутая сложность исчезает.

Однако ситуация с возмущениями на уровне планковских масштабов в пространстве $\mathcal{X}$ совершенно иная, так как в данном случае основное состояние $\mathcal{X}$ по своей природе *не* локализовано в какой-либо точке обычного пространства-времени $\mathcal{M}$, а считается вездесущим, пронизывающим структуру пространства-времени во всей Вселенной. Предполагается, что геометрическое квантовое состояние $\mathcal{X}$ влияет на физические детали в какой-нибудь далекой галактике в такой же степени, как и на физику здесь, на Земле. Аргумент теоретиков-струнников, что энергия планковских масштабов была бы слишком велика по сравнению с имеющейся, чтобы можно было возбудить $\mathcal{X}$, представляется несостоятельным по разным причинам. Такие энергии не просто легкодоступны на уровне нелокализованных процессов (например, вращение Земли). Более того, если бы мы предположили, что $\mathcal{X}$ должно переходить в возбужденное состояние $\mathcal{X}'$ именно в результате такого одночастичного перехода (возможно, в будущем появятся какие-то более совершенные технологии, которые позволят сконструировать ускоритель, разгоняющий частицы до планковских энергий), приводящего к возникновению нового состояния Вселенной $\mathcal{M} \times \mathcal{X}'$, это было бы совершенно абсурдно, поскольку мы не могли бы ожидать, что физика в Туманности Андромеды немедленно изменится в результате такого события здесь, на Земле! Мы могли бы предположить гораздо менее значительное событие поблизости от Земли, эффект которого распространялся бы со скоростью света.

Такие явления можно было бы гораздо более правдоподобно описать при помощи нелинейных классических уравнений, а не резких квантовых переходов.

В свете таких соображений хочу вернуться к вопросу, которого касался ранее, и попытаться рассмотреть, каким образом планковский квант энергии, распределенный по некоторой довольно обширной области $\mathcal{M}'$ пространства $\mathcal{M}$, может повлиять на геометрию пространства $\mathcal{X}$ в этой области. Как было указано выше, $\mathcal{X}$ при этом почти не будет затронуто, если мы трактуем это изменение как результат рассредоточения планковской энергии. Соответственно, если мы собираемся рассмотреть по-настоящему значительные изменения формы или размера $\mathcal{X}$, то есть заменить $\mathcal{X}$ на $\mathcal{X}^*$, существенно отличающееся от $\mathcal{X}$, нам придется оперировать энергиями, значительно превышающими планковские (такие энергии в изобилии встречаются в известной нам физической Вселенной, например при вращении Земли вокруг Солнца). Такие энергии возникают не из-за единственного «минимального» кванта энергии (в планковских масштабах), но в результате того, что в $\mathcal{X}$ «впрыскивается» огромная последовательность квантов, каждый из которых привносит очередное изменение. Чтобы обеспечить трансформацию пространства $\mathcal{X}$ в существенно иное пространство $\mathcal{X}^*$ в пределах какой-либо достаточно обширной области, нужно задействовать колоссальное количество таких квантов (может быть, планковского порядка, а может быть, и больше). Сегодня при рассмотрении эффектов, требующих такого огромного количества квантов, принято считать, что подобные эффекты лучше всего описывать чисто классически (то есть без привлечения квантовой механики).

На самом деле, как будет показано в главе 2, проблема объяснения возникновения внешне классических явлений в результате множества квантовых событий вскрывает ряд глубоких вопросов, связанных с тем, как именно классический мир соотносится с квантовым. Это интересный (и неоднозначный) вопрос: возникает «классичность» просто из-за вовлеченности большого множества квантов или в силу какого-то иного критерия? Это я собираюсь подробно рассмотреть в разделе 2.13. Однако в рамках обсуждения текущей проблемы такие тонкости кажутся для нас не слишком важными, и я просто отмечу, что следует считать оправданным использование *классических* аргументов при обсуждении проблемы таких пертурбаций в пространстве-времени $\mathcal{M} \times \mathcal{X}$, которые действительно значительно видоизменяют пространство $\mathcal{X}$. При этом возникают свои серьезные сложности с дополнительными пространственными измерениями — о них мы поговорим в следующем разделе.

Однако прежде чем переходить к этим сложностям, касающимся сугубо конкретной формы пространственной надразмерности, возникающей в рамках теории струн, полагаю, уместно было бы провести сравнение с экспериментальной

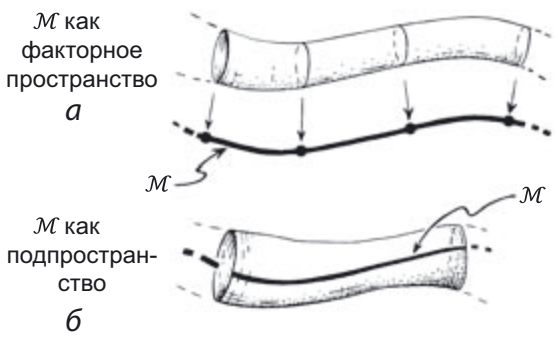


Рис. 1.32. Разница между $\mathcal{M}$ как факторным пространством (*а*) и $\mathcal{M}$ как подпространством (*б*)

ситуацией, которая зачастую преподносится как аналогичная рассматриваемой здесь. Данная ситуация иллюстрируется на примере так называемого *квантового эффекта Холла* [von Klitzing et al., 1980; von Klitzing, 1983]. Эффект связан с четко зафиксированным двумерным квантовым феноменом, который возникает в физике трехмерного пространства. Из-за высокого энергетического барьера интересующая нас система реализуется лишь в двумерной плоскости, а квантовая физика такого пространства с небольшим числом измерений совершенно «не замечает» третьего пространственного измерения, поскольку явления, происходящие в такой двумерной системе, не обладают достаточной энергией, чтобы преодолеть этот энергетический барьер. Поэтому иногда утверждается, что подобный пример служит аналогией процесса, который может происходить в многомерном пространстве теории струн, где наша привычная трехмерная физика схожим образом «не замечает» девятимерного окружающего мира, и эта физика заключена в рамках трех измерений, поскольку, предположительно, не может преодолеть некоторый энергетический барьер.

Однако это совершенно ложная аналогия. Дело в том, что вышеприведенный пример точнее соответствует представлению о мире как о совокупности бран (см. раздел 1.16), где пространство с меньшим числом измерений является *подпространством* того пространства, которое обладает бóльшим числом измерений, а не *факторным пространством*, актуальным для рассматриваемых здесь стандартных представлений из теории струн (см. выше пример с выражением $\mathcal{M} \times \mathcal{X}$, где $\mathcal{M}$ — фактор; см. раздел A.7 и рис. 1.32). Роль $\mathcal{M}$ как факторного пространства, а не подпространства, опять же связана с аргументацией из предыдущего раздела. Однако картина с подпространством важна для совершенно иной точки зрения (мир из бран), которую я подробнее опишу в разделе 1.16.

1.11. Классическая нестабильность теории струн для высших измерений

Давайте вернемся к проблеме, затронутой в разделе 1.10, — стабильности классического пространства-времени вида $\mathcal{S} = \mathcal{M} \times \mathcal{X}$, где $\mathcal{X}$ — микроскопическое компактное пространство. Не останавливаясь в своей аргументации на подробном описании природы пространства $\mathcal{X}$, я буду давать формулировки в терминах тех версий теории струн, где $\mathcal{X}$ относится к типу компактного шестимерного пространства, известного под названием *многообразие Калаби — Яу*, которое мы подробнее рассмотрим в разделах 1.13 и 1.14. В таком случае $\mathcal{S}$ будет десятимерным пространством-временем. В данном случае мы будем иметь дело с некоторыми элементами суперсимметрии (см. раздел 1.14), но они не играют какой-либо роли в классических выкладках, которые я привожу здесь (и которые могут восприниматься применительно к «основной» части системы, см. раздел 1.14). Поэтому я позволю себе до поры до времени игнорировать суперсимметрию, отложив рассуждения о ней до раздела 1.14. В сущности, все мои требования к $\mathcal{X}$ следующие: это пространство должно быть как минимум двумерным, что определенно предполагается в современной теории струн, однако мне понадобится, чтобы пространство-время $\mathcal{S}$ удовлетворяло некоторым уравнениям поля.

Выше (см. раздел 1.10) я отмечал, что, согласно теории струн, действительно должны существовать уравнения поля, которым удовлетворяет метрика, присваиваемая данному пространству-времени $\mathcal{S}$ с высшими измерениями. В качестве первого приближения можем предположить, что такое множество уравнений для $\mathcal{S}$ — это *вакуумные уравнения Эйнштейна* ${}^{10}\mathbf{G} = \mathbf{0}$, где ${}^{10}\mathbf{G}$ — тензор Эйнштейна, полученный из 10-метрики $\mathcal{S}$. Эти уравнения постулируются для пространства-времени $\mathcal{S}$, где находятся струны, с целью избежать *аномалии*, выходящей за пределы той аномалии, которая упоминалась в разделе 1.6 и из-за которой теоретики увеличили размерность пространства-времени. На самом деле ${}^{10}\mathbf{G}$ в уравнении ${}^{10}\mathbf{G} = \mathbf{0}$ — это всего лишь первый член в степенном ряде малой величины α', так называемой *струнной константы*. Это исчезающе малый параметр плоскости; обычно считается, что он лишь немного превышает квадрат планковской длины (см. раздел 1.5):

$$\alpha' \approx 10^{-68}\,\text{м}^2.$$

Итак, у нас есть уравнения поля для S, которые можно выразить в виде степенных рядов (раздел A.10):

$$\mathbf{0} = {}^{10}\mathbf{G} + \alpha'\mathbf{H} + \alpha'^2\mathbf{J} + \alpha'^3\mathbf{K} + \cdots,$$

где $\mathbf{H}$, $\mathbf{J}$, $\mathbf{K}$ и т. д. — выражения, построенные на основе римановой кривизны и ее различных высших производных. Однако из-за исключительно малого значения α' члены более высокого порядка обычно игнорируются в некоторых конкретных версиях теории струн (хотя есть сомнения, насколько такие действия правомерны, поскольку нам ничего не известно о схождении или итоговых свойствах этих рядов (ср. разделы A.10 и A.11)). В частности, можно без оговорок принять, что пространства Калаби — Яу, о которых шла речь ранее (см. также разделы 1.13 и 1.14), удовлетворяют шестимерному уравнению ${}^{6}\mathbf{G} = \mathbf{0}$, в результате соответствующее десятимерное уравнение ${}^{10}\mathbf{G} = \mathbf{0}$ соблюдается для произведения пространств $\mathcal{M} \times \mathcal{X}$ при условии, что стандартное эйнштейновское уравнение вакуума ${}^{4}\mathbf{G} = \mathbf{0}$ также полагается верным для $\mathcal{M}$ (это было бы логично для *вакуумного* «основного состояния» полей материи)[1]. При этом я беру эйнштейновские уравнения без лямбда-члена (см. разделы 1.1 и 3.1), поскольку космологическая постоянная будет пренебрежимой в рассматриваемых здесь масштабах.

В соответствии с вышеизложенным предположим, что вакуумные уравнения ${}^{10}\mathbf{G} = \mathbf{0}$ действительно соблюдаются для $\mathcal{S} = \mathcal{M} \times \mathcal{X}$. Нас интересует, что случится, если произвести небольшое возмущение в пространстве $\mathcal{X}$ с «дополнительными измерениями» (например, в пространстве Калаби — Яу). Здесь следует сделать важное замечание о природе тех возмущений, которые я собираюсь рассматривать. В сообществе теоретиков-струнников активно обсуждаются такие возмущения, которые могли бы деформировать пространство Калаби — Яу и превратить его в несколько иное пространство, немного изменив *модули*, о которых мы поговорим в разделе 1.16. Эти модули определяют конкретную форму многообразия Калаби — Яу в рамках конкретного топологического класса. Пример возмущений, меняющих такие модули, — *нулевые моды*, упоминавшиеся в разделе 1.10. В текущем разделе меня не особенно интересуют деформации такого рода, не выводящие нас за пределы семейства пространств Калаби — Яу. В традиционной теории струн принято полагать, что следует оставаться в рамках этого семейства, поскольку эти пространства считаются *стабильными*, ведь они удовлетворяют критериям суперсимметрии, ограничивающим такое многообразие шестью дополнительными пространственными измерениями. Приводимые соображения призваны показать, что лишь такие шестимерные пространства

---

[1] Эйнштейновский тензор ${}^{m,n}\mathbf{G}$ произведения $\mathcal{M} \times \mathcal{N}$ двух (псевдо-)римановых пространств, обладающих размерностями m и n соответственно, можно выразить как непосредственную сумму ${}^{m}\mathbf{G} \oplus {}^{n}\mathbf{G}$ соответствующих отдельных эйнштейновских тензоров ${}^{m}\mathbf{G}$ и ${}^{n}\mathbf{G}$, где (псевдо-)метрика $\mathcal{M} \times \mathcal{N}$ определяется как непосредственная сумма ${}^{m,n}\mathbf{g} = {}^{m}\mathbf{g} \oplus {}^{n}\mathbf{g}$ индивидуальных (псевдо-)метрик ${}^{m}\mathbf{g}$ и ${}^{n}\mathbf{g}$ пространств $\mathcal{M}$ и $\mathcal{N}$ (см. [Guillemin and Pollack, 1974]; подробнее эта тема рассмотрена у [Besse, 1987]). Следовательно, $\mathcal{M} \times \mathcal{N}$ имеет эйнштейновский тензор, обращающийся в нуль, тогда и только тогда, когда и $\mathcal{M}$ и $\mathcal{N}$ имеют эйнштейновский тензор, обращающийся в нуль.

удовлетворяют необходимым критериям суперсимметрии. Обычно стабильность формулируется как требование, чтобы небольшое возмущение, видоизменяющее пространство Калаби — Яу, оставляло его в рамках этого класса пространств; при этом не рассматривается возможность того, что подобное возмущение может вывести пространство из данного семейства, в конечном итоге превратив его в *сингулярное*, в котором не существует плавной метрики. На самом деле последняя возможность лавинообразной эволюции в такую сингулярную конфигурацию явно подразумевается при следующих рассуждениях.

Начнем с основного случая, где $\mathcal{M}$ не испытывает вообще никаких пертурбаций. Итак, мы утверждаем, что $\mathcal{M} = \mathbb{M}$, где $\mathbb{M}$ — плоское четырехмерное пространство Минковского, в котором действует *специальная* теория относительности (см. раздел 1.7). $\mathbb{M}$, будучи плоским, может быть выражено как произведение пространств:

$$\mathbb{M} = \mathbb{E}^3 \times \mathbb{E}^1$$

(см. раздел А.4, рис. А.25; в данном случае мы просто особым образом группируем координаты *x, y, z, t*: сначала (*x, y, z*), а затем *t*). Евклидово трехмерное пространство $\mathbb{E}^3$ — это обычное пространство (с координатами *x, y, z*), а одномерное евклидово пространство $\mathbb{E}^1$ — обычное время (координата *t*), причем последняя координата — просто копия вещественной линии $\mathbb{R}$. Если записать $\mathbb{M}$ таким образом, то целостное (невозмущенное) пространство-время $\mathcal{S}$ можно выразить в виде (считая $\mathbb{M}$ и $\mathcal{X}$ факторными пространствами $\mathcal{S}$):

$$\begin{aligned}\mathcal{S} &= \mathbb{M} \times \mathcal{X} = \\ &= \mathbb{E}^3 \times \mathbb{E}^1 \times \mathcal{X} = \\ &= \mathbb{E}^3 \times \mathcal{Z},\end{aligned}$$

просто перегруппировав координаты, где $\mathcal{Z}$ — семимерное пространство-время:

$$\mathcal{Z} = \mathbb{E}^1 \times \mathcal{X}$$

(координаты для $\mathcal{Z}$ устанавливаются в таком порядке: сначала *t*, а затем координаты $\mathcal{X}$).

Рассмотрим малую (но не бесконечно малую) пертурбацию шестимерного пространства $\mathcal{X}$ (например, Калаби — Яу), превращающегося в новое пространство $\mathcal{X}^*$, при $t = 0$, причем в данном случае мы можем предположить, что это возмущение распространяется в направлении $\mathbb{E}^1$ (при координате времени *t*), и в результате такого развития мы получаем семимерное пространство-время $\mathcal{Z}^*$. Для начала предположим, что эта пертурбация касается лишь $\mathcal{X}$, но не затрагивает внешнее евклидово трехмерное пространство $\mathbb{E}^3$. Такая картина отлично

согласуется с эволюционными уравнениями, но поскольку ожидается, что подобная пертурбация должна будет каким-то образом изменять 6-геометрию $\mathcal{X}^*$ с течением времени, мы не рассчитываем, что $\mathcal{Z}^*$ останется произведением пространств вида $\mathbb{E}^1 \times \mathcal{X}^*$, причем точная 7-геометрия $\mathcal{Z}^*$ будет определяться эйнштейновскими уравнениями ${}^7\mathbf{G} = \mathbf{0}$. Однако целостное пространство-время $\mathcal{S}$ все-таки должно в ходе своего развития остаться произведением $\mathbb{E}^3 \times \mathcal{Z}^*$, поскольку полные эйнштейновские уравнения ${}^{10}\mathbf{G} = \mathbf{0}$ будут соблюдаться и в этом произведении пространств, если $\mathcal{Z}^*$ удовлетворяет ${}^7\mathbf{G} = \mathbf{0}$, поскольку ${}^3\mathbf{G} = \mathbf{0}$ определенно соблюдается для плоского пространства $\mathbb{E}^3$.

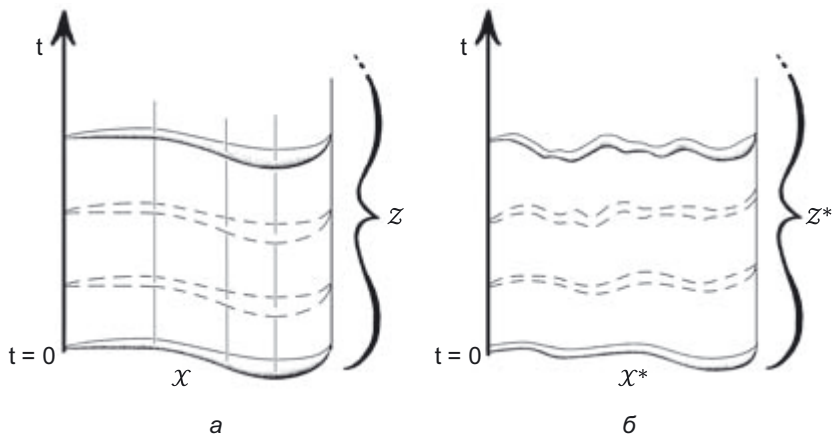


Рис. 1.33. Крошечное компактное пространство $\mathcal{X}$, когда пространство Калаби — Яу не изменяется с течением времени (*а*), но в случае когда в результате пертурбации оно принимает несколько иную форму $\mathcal{X}^*$, то в ходе эволюции превращается в нечто иное (*б*)

Шестимерное пространство $\mathcal{X}^*$ принимается в качестве отсчетной поверхности для эволюции $\mathcal{Z}^*$ при значении $t = 0$ (рис. 1.33). Затем, согласно уравнениям ${}^7\mathbf{G} = \mathbf{0}$, эта пертурбация распространяется в будущее (со значениями $t > 0$). Существуют определенные уравнения связей, которые должны удовлетворяться в $\mathcal{X}$, причем иногда возникает довольно тонкая проблема: как доказать на строгом математическом языке, что эти уравнения связей действительно будут удовлетворять нашим требованиям в каждой точке компактного пространства $\mathcal{X}^*$? Тем не менее функциональная свобода, которую мы рассчитываем получить при таких исходных пертурбациях $\mathcal{X}$, равна

$$\infty^{28\infty^6},$$

причем значение 28 в этом выражении получается путем подстановки $n = 7$ в выражение $n(n-3)$ и, таким образом, равняется числу исходных независимых компонентов на точку исходной $(n-1)$-плоскости для n-пространства с обращающимся в нуль тензором Эйнштейна. Значение 6 в данном случае — это размерность исходной шестимерной поверхности $\mathcal{X}^*$ [Wald, 1984]. Это касается пертурбаций самого $\mathcal{X}$ и *внешних* пертурбаций, связанных с тем, как $\mathcal{X}$ вплетено в $\mathcal{Z}$. Такая классическая свобода, естественно, многократно превосходит функциональную свободу $\infty^{k^{\infty^3}}$, которую мы ожидаем увидеть в физической теории, адекватно описывающей явления, происходящие в воспринимаемом нами трехмерном мире.

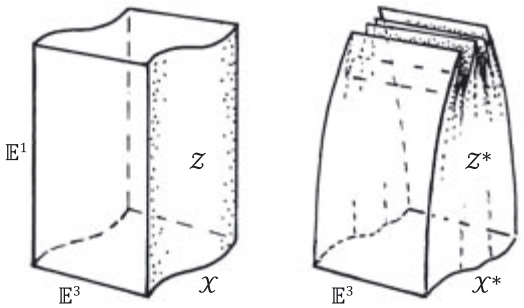


Рис. 1.34. Классическая нестабильность дополнительных измерений в теории струн. Эволюция $\mathcal{Z}^*$ возмущенного шестимерного пространства $\mathcal{X}^*$ с дополнительными измерениями с большой вероятностью приводит к сингулярности в соответствии с теоремой, сформулированной в 1970 году Хокингом и автором

Однако на самом деле ситуация еще серьезнее, поскольку практически все подобные пертурбации приводили бы к эволюции, которая давала бы в итоге *сингулярное* $\mathcal{Z}$ (рис. 1.34). Фактически это означает, что дополнительные измерения должны каким-то образом деградировать до такого состояния, в котором кривизны расходятся до бесконечности и дальнейшая эволюция классических уравнений становится невозможной. Такой вывод следует из математических *теорем о сингулярности*, доказанных в конце 1960-х годов. Среди них особого упоминания заслуживает та, которую мы тогда сформулировали со Стивеном Хокингом. Среди прочего данная теорема показала, что практически любое n-мерное пространство-время ($n \geq 3$), содержащее компактную пространственноподобную поверхность $(n-1)$ (здесь речь идет об исходном шестимерном пространстве Калаби — Яу $\mathcal{X}^*$), но не содержащее замкнутых времениподобных петель, должно развиться в пространственно-временную сингулярность, если эйнштейновский тензор ${}^n\mathbf{G}$ удовлетворяет энергетическому условию (неотрицательности), которое называется *сильное энергетическое условие* (здесь оно

определенно соблюдается, поскольку ${}^{7}\mathbf{G} = \mathbf{0}$ во всем пространстве $\mathcal{Z}^*$). Оговорку «почти» и отсутствие «замкнутых времениподобных петель» в данном случае можно игнорировать, поскольку возможные лазейки если и могут возникнуть, то лишь в исключительных обстоятельствах со значительно более низкой функциональной свободой, нежели та, что возникает при типичной пертурбации $\mathcal{Z}$.

Здесь следует сделать техническое замечание. На самом деле теорема не утверждает, что кривизны *расходятся до бесконечности*, а лишь гласит, что, как правило, эволюция может продолжаться только до определенной точки. Хотя и существуют альтернативные варианты, которые, в принципе, могут складываться в исключительных ситуациях, следует исходить из того, что продолжение эволюции оказывается невозможным именно потому, что кривизны *действительно* расходятся [Clarke, 1993]. Еще один важный момент в данном случае заключается в следующем: подразумеваемое здесь сильное энергетическое условие, конечно, автоматически соблюдается в уравнении ${}^{7}\mathbf{G} = \mathbf{0}$, но это определенно не гарантируется, если мы попробуем рассмотреть члены высшего порядка в степенных рядах α', упоминавшихся выше. Все-таки большинство актуальных разработок в теории струн, по-видимому, работают на том уровне, где эти члены высшего порядка из α' игнорируются, а $\mathcal{X}$ действительно считается пространством Калаби — Яу. Представляется, что смысл этой теоремы сингулярности таков: пока пертурбации дополнительных измерений можно трактовать в *классическом* смысле, что действительно представляется разумным на основе тех выводов, что были явно сделаны ранее, в разделе 1.10, следует ожидать сильной нестабильности в шести дополнительных пространственных измерениях, то есть они будут рушиться, *приближаясь* к состоянию сингулярности. Непосредственно перед такой катастрофой, возможно, придется всерьез учитывать члены α' высшего порядка или другие квантовые соображения. В зависимости от масштаба пертурбаций вполне можно полагать, что такое «разрушение» будет занимать ничтожные доли секунды, и при этом следует помнить, что *планковское время* (период, за который свет преодолевает планковскую длину; см. раздел 1.5) составляет порядка 10^{-43} секунд! Какую бы форму ни принимали эти дополнительные измерения после разрушения, это непременно отражалось бы на наблюдаемой физике радикальным образом. Все это не слишком напоминает удобную картинку десятимерного пространства-времени, которую теоретики-струнники предполагают увидеть в нашей Вселенной.

Здесь также следует отметить и другую проблему, а именно: вышеупомянутые пертурбации затрагивают *только* шесть дополнительных измерений, причем макроскопические измерения (в данном случае евклидово пространство $\mathbb{E}^3$) остаются без изменений. Действительно, при пертурбациях во всех пространственных измерениях девятимерного пространства $\mathbb{E}^3 \times \mathcal{X}$ было бы гораздо больше функциональной свободы (а именно $\infty^{70\infty^9}$), чем при пертурбациях, за-

трагивающих одно лишь измерение $\mathcal{X}$, функциональная свобода которых равна $\infty^{28\infty^6}$. Представляется возможным изменить вышеизложенные аргументы таким образом, чтобы все та же теорема [Hawking and Penrose, 1970] по-прежнему работала и снова (уже более замысловатым образом) приводила нас к заключению о сингулярности, но уже для всего пространства-времени [ПкР, см. раздел 31.46]. Обратите внимание: любые пертурбации четырехмерного пространства-времени, в принципе, сравнимые с теми, что мы рассматриваем для дополнительных шести измерений, катастрофическим образом сказались бы на обычной физике, поскольку такие крошечные искривления, касающиеся пространства $\mathcal{X}$, просто не заметны при наблюдаемых физических явлениях. Возникает неудобная проблема, которая в любом случае остается нерешенной в современной теории струн, а именно: каким образом искривления на столь различных масштабах могут сосуществовать, почти не затрагивая друг друга. Мы вновь вернемся к этой загвоздке в разделе 2.11.

1.12. Модность теории струн

Возможно, читатель уже озадачился вопросом, почему теория струн столь серьезно воспринимается огромным большинством в сообществе исключительно компетентных физиков-теоретиков, и в частности тех, чья работа непосредственно ведет нас к более глубокому пониманию базовой физики того мира, в котором мы живем. Если теория струн (и последующие разработки на ее основе) действительно выстраивает картину мира, в которой присутствуют высшие пространственные измерения и которая столь противоречит известной нам физике, то почему она остается столь модной в таком крупном сообществе исключительно сведущих физиков-теоретиков? Далее я покажу, *насколько* она модна в настоящий момент. Но признавая за ней такой статус, можно задаться вопросом: почему теоретики-струнники так невозмутимо относятся к аргументам против физической правдоподобности пространства-времени, которое содержало бы высшие измерения (например, к аргументам, кратко изложенным в разделах 1.10 и 1.11). Почему, в самом деле, такой исключительно модный статус остается фактически незыблемым в свете аргументов о ее неправдоподобности?

В сущности, в предыдущих двух разделах изложены те самые аргументы, которые я впервые выдвинул в заключительном выступлении в рамках семинара на конференции в честь 60-летия Стивена Хокинга, состоявшейся в британском Кембридже в январе 2002 года [Penrose, 2003]. На этой лекции присутствовали некоторые ведущие теоретики, занимающиеся теорией струн (и в частности, Габриэле Венециано и Майкл Грин), которые на следующий день поспорили со мной по поводу некоторых высказанных мною аргументов. Однако с тех

пор почти не было ни откликов, ни контраргументов — и, разумеется, никакого публичного опровержения изложенных мною идей. Возможно, самый показательный отклик я получил на следующий день за обедом от Леонарда Сасскинда — постараюсь воспроизвести его настолько дословно, насколько помню:

Вы совершенно правы, разумеется, но в корне заблуждаетесь!

Я не до конца понимаю, что он хотел сказать этой ремаркой, но полагаю, что он имел в виду следующее: хотя теоретики-струнники должны быть готовы признать, что дальнейшей разработке их теории препятствуют некоторые до сих пор не преодоленные математические сложности — практически обо всех сообщество струнников уже хорошо осведомлено, — такие проблемы считаются лишь техническими деталями, которые не должны мешать реальному прогрессу. Эти теоретики сказали бы, что такие детали, в сущности, не важны, поскольку фундаментально теория струн развивается в верном направлении. Они, разрабатывая ее, не должны ни тратить время на подобные математические тонкости, ни даже привлекать внимание к таким маловажным неувязкам на этапе разработки темы, чтобы не мешать общему движению струнного сообщества к осуществлению всех его потенциальных целей и не снижать его потенциал.

Такое пренебрежение общей математической непротиворечивостью кажется мне особенно вопиющим в теории, которая, так или иначе, развивается в основном на базе математики (вскоре я это объясню). Более того, некоторые возражения, которые я собираюсь высказать, далеко не сводятся к математическим препятствиям, мешающим непротиворечивой разработке струнной модели как правдоподобной физической теории — об этом речь пойдет в разделе 1.16. Даже заявленная *конечность* струнных вычислений, призванных заменить расходящиеся фейнмановские диаграммы, упомянутые в разделе 1.5, еще далеко не сформулирована математически [Smolin, 2006; см., в частности, с. 278–281]. Явная незаинтересованность в четкой математической демонстрации выразительно проявляется в таких комментариях, как приведенный ниже (по-видимому, принадлежит нобелевскому лауреату Дэвиду Гроссу):

Конечность теории струн столь очевидна, что если бы кто-нибудь представил мне математическое доказательство этого, мне было бы неинтересно его читать.

Абэй Аштекар, познакомивший меня с этой цитатой, был практически уверен, что автором ее является Гросс. Однако любопытно, что когда я читал на эту тему лекцию в Варшаве в 2005 году, Дэвид Гросс вошел в аудиторию в тот самый момент, когда я привел эту цитату! Поэтому я спросил, не он ли ее автор. Он не

стал этого отрицать, но признался, что теперь ему было бы интересно увидеть такое доказательство.

Надежда на то, что теория струн окажется *конечной*, то есть лишенной расходимостей традиционной квантовой теории поля, возникающих в результате стандартного анализа в терминах фейнмановских диаграмм (и других математических приемов), определенно была одной из фундаментальных движущих сил всей этой теории. В принципе, существует факт: при струнных расчетах, замещающих вычисления по фейнмановским диаграммам согласно рис. 1.16 в разделе 1.6, можно пользоваться «комплексной магией» римановых поверхностей (см. разделы А.10 и 1.6). Однако даже ожидаемая конечность отдельно взятой амплитуды (см. разделы 1.6 и 2.6), возникающая в результате конкретного топологического расположения струн, сама по себе не дает нам конечную теорию, поскольку каждая струнная топология представляет всего один член в *ряду* струнных представлений с возрастающей топологической сложностью. К сожалению, если каждый отдельный топологический член в действительности конечен — по-видимому, такого базового убеждения придерживаются теоретики-струнники, о чем свидетельствует вышеприведенная цитата, — весь ряд в целом, как ожидается, должен *расходиться*, что продемонстрировал сам Гросс [см. Gross and Periwal, 1988]. Как ни неуклюже все это выглядит с математической точки зрения, такая расходимость обычно кажется теоретикам-струнникам добрым знаком, который просто свидетельствует о том, что данное разложение в степенной ряд делается «вокруг неверной точки» (см. раздел А.10) и тем самым иллюстрирует определенное свойство струнных амплитуд, которое действительно было ожидаемо. Тем не менее такая несуразность, кажется, хоронит надежду на то, что теория струн могла бы дать нам конечную процедуру для расчета амплитуд в квантовой теории поля.

Итак, *насколько* же модна теория струн? Чтобы оценить популярность этой теории в качестве подхода к квантовой гравитации (как минимум по состоянию на 1997 год), можно обратиться к небольшому исследованию, представленному в лекции Карло Ровелли, которую он прочитал на Международном конгрессе по общей теории относительности и гравитации. Конгресс состоялся в индийском городе Пуна в декабре 1997 года, а сама лекция была посвящена различным подходам к квантовой гравитации, актуальным на тот момент. Следует отметить, что Ровелли был одним из основоположников конкурирующего подхода к квантовой гравитации, связанного с петлевыми переменными [Rovelli, 2004; см. также ПкР, глава 32]. Он не претендовал на роль объективного социолога, и, несомненно, можно усомниться в том, насколько серьезным может считаться такое исследование. Однако едва ли это важно, и здесь я не буду заострять на этом внимание. Он просто проверил лос-анджелесские архивы и посмотрел,

сколько статей по каждому из подходов к квантовой гравитации было опубликовано за предыдущий год. Получились следующие результаты:

- теория струн — 69,
- петлевая квантовая гравитация — 25,
- КТП в искривленных пространствах — 8,
- решеточный подход — 7,
- евклидова квантовая гравитация — 3,
- некоммутативная геометрия — 3,
- квантовая космология — 1,
- твисторы — 1,
- прочие — 6.

Итак, мы видим, что теория струн по популярности не просто с отрывом лидирует среди всех подходов к квантовой гравитации, но и значительно превосходит все остальные подходы, вместе взятые.

Оказалось, что в последующие годы Ровелли продолжал это исследование, немного ограничив выборку тем, но охватив все годы с 2000-го по 2012-й, и отслеживал относительную популярность всего трех подходов к квантовой гравитации: теории струн, петлевой квантовой гравитации и теории твисторов (рис. 1.35). Из рисунка видно, что теория струн по-прежнему уверенно удерживает позиции наиболее популярного из подходов — пик пришелся примерно на 2007 год, но спад с тех пор был небольшим. Основное изменение за этот период — устойчиво возрастающий интерес к петлевой квантовой гравитации. Начиная с 2004 года прослеживается заметный, но значительно более скромный рост интереса к теории твисторов, связанный, возможно, с теми факторами, о которых я расскажу в разделе 4.1. Однако не стоит придавать слишком большого значения этим тенденциям.

После того как я показал таблицу Ровелли за 1997 год на моей принстонской лекции 2003 года, меня заверили, что доля статей по теории струн к тому моменту была значительно выше, чем мне казалось. Действительно, именно в тот момент популярность теории струн сильно возросла. Кроме того, я подозревал, что моему собственному детищу, то есть теории твисторов (см. раздел 4.1), еще повезло получить в 1997 году свою «единицу» и что на момент лекции более правдоподобно выглядела оценка «ноль». Полагаю, что сегодня некоммутативная геометрия получила бы гораздо больше той «тройки», что имела тогда, но в последующем исследовании Ровелли эта тема не учитывалась. Разумеется, надо совершенно четко понимать, что подобные таблицы ничего не говорят о том,

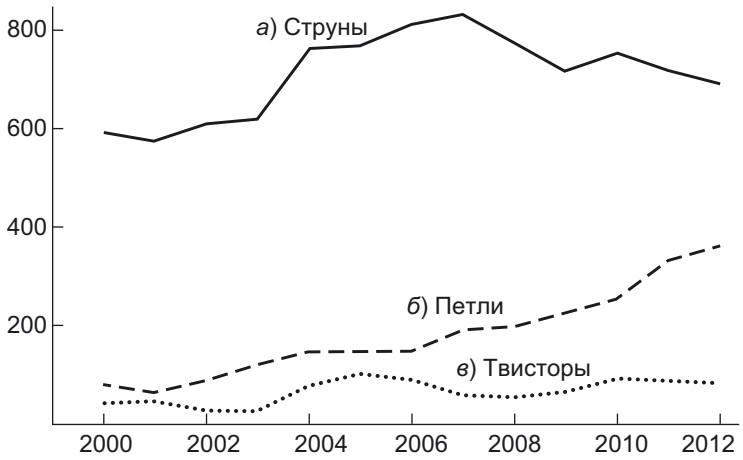


Рис. 1.35. Исследование Карло Ровелли по материалам лос-анджелесских архивов, отражающее популярность трех подходов к квантовой гравитации за 2000–2012 годы: теория струн, петлевая квантовая гравитация, теория твисторов

насколько та или иная гипотеза может быть близка к истинному устройству природы. Они лишь позволяют оценить, насколько модна та или иная картина мира. Более того, в главе 3 и в разделе 4.2 я расскажу, что, по моему личному мнению, *ни один* из современных подходов к квантовой гравитации не даст нам такую теорию, которая полностью согласовывалась бы с естественным сочетанием двух столпов современной физики: общей теории относительности и квантовой механики. Все дело в той ключевой причине, что квантовая гравитация, на мой взгляд, — совсем не то, что следует искать! Этот термин скорее подразумевает, что мы должны стремиться сделать существующую квантовую теорию применимой к гравитационному полю, тогда как я считаю, что следует изменить саму структуру квантовой механики, когда речь идет о включении в нее гравитации. Соответственно, полученная теория будет не строго квантовой теорией, а чем-то, что отличается от нынешних процедур квантования (см. раздел 2.13).

Однако стремление найти подходящую теорию квантовой гравитации очень реально. Многие физики, и в особенности ретивые молодые аспиранты, всерьез желают значительно продвинуться к очень достойной цели и объединить две великие революционные теории XX века: странную, но величественную квантовую механику и выдающуюся эйнштейновскую теорию, объясняющую гравитацию искривлением пространства-времени. Обычно эту цель именуют просто *квантовая гравитация* и пытаются достичь ее, применяя законы стандартной (квантовой) теории поля к теории гравитации (в разделах 2.13 и 4.2 я пред-

ставлю собственный, существенно иной ракурс такого объединения). Несмотря на то что можно обоснованно утверждать, что ни одна из современных теорий и близко не подводит нас к этой цели, поборники теории струн чувствуют себя довольно уверенно и продвигают собственное убеждение, согласно которому теория струн — единственный достойный вариант. Как выразился в 1999 году ведущий теоретик-струнник Джозеф Полчински:

Альтернатив нет... все хорошие идеи относятся к теории струн.

С другой стороны, необходимо помнить, что теория струн — плод всего одной научной школы в рамках теоретической физики. Это особая культура, выросшая на основе физики частиц и представлений о квантовой теории поля, где на первый план выходят проблемные вопросы, связанные с приведением расходящихся выражений к конечному результату. Эта культура значительно отличается от тех, которые по сути своей более непосредственно связаны с общей теорией относительности Эйнштейна. Здесь особая важность приписывается сохранению общих принципов, среди которых наиболее примечательны *принцип эквивалентности* (между эффектами ускорения и равитационного поля; см. разделы 3.7 и 4.2) и *принцип общей ковариантности* (см. разделы A.5 и 1.7), составляющие основу теории Эйнштейна. Так, подход к квантовой гравитации с использованием петлевых переменных принципиально основан на главенстве принципа общей ковариантности; при этом может показаться, что теория струн этот принцип практически полностью игнорирует!

Думаю, что соображения вроде описанного выше исследования Ровелли дают нам лишь самое приблизительное впечатление о преобладании теории струн и тех теорий, что были разработаны на ее основе (см. разделы 1.13 и 1.15), среди исследователей-теоретиков, пытающихся докопаться до оснований физики. На абсолютном большинстве физических факультетов и в физических НИИ по всему миру теоретики, скорее всего, будут уделять значительное внимание изучению теории струн либо одного из ее многочисленных ответвлений. Хотя в последние годы такое преобладание могло несколько уменьшиться, те студенты, которые действительно хотят исследовать основания физики, например квантовую гравитацию, в большинстве своем по-прежнему принимаются за теорию струн (или другие смежные дисциплины, связанные с высшими измерениями), зачастую жертвуя другими подходами, как минимум столь же многообещающими. Однако другие подходы далеко не так известны, и даже те студенты, которые не особо желают заниматься теорией струн, с трудом выходят на такие альтернативные варианты, в особенности из-за малочисленности подходящих потенциальных научных руководителей (хотя, по-видимому, петлевая квантовая гравитация в этом

отношении в последние годы завоевала серьезные позиции). Сами принципы развития карьеры в области теоретической физики (и, несомненно, во многих других научных областях) довольно тенденциозны и служат дальнейшему развитию областей, уже ставших модными. Такие факторы лишь повышают популярность теории струн, уже пользующейся значительным авторитетом.

Важный дополнительный фактор, помогающий распространению моды, — это финансирование. Комитеты, призванные оценивать относительную ценность исследовательских проектов в различных областях, с высокой вероятностью будут действовать с оглядкой на то, насколько велик интерес, который в настоящее время вызывает каждая из этих областей. На самом деле, если одна из этих областей исключительно популярна, то и многие члены комитета сами могли успеть ею позаниматься — возможно, даже отчасти посодействовали ее популярности. Следовательно, они будут выше оценивать исследования в модных областях, нежели в немодных. Все это неизбежно вызывает нестабильность и способствует гипертрофированному интересу к уже модным идеям за счет всех остальных. Более того, современные электронные телекоммуникационные технологии и межконтинентальное авиасообщение создают отличные условия для стремительного распространения тех идей, что уже стали модными, особенно в мире высокой конкуренции. Необходимость быстро получить результат ставит в выигрышное положение тех, кто может быстро развиваться, опираясь на готовые результаты предшественников, и работать в уже раскочегаренной области, по сравнению с теми, кто хочет вырваться из рамок и, возможно, долго и кропотливо развивать идеи, значительно отклоняющиеся от мейнстрима.

Однако я прихожу к выводу, что в настоящее время (по меньшей мере на некоторых физических факультетах в США) уже чувствуется, что достигнута своеобразная точка насыщения и что среди новых преподавателей, приглашаемых в университеты, должно расти число представителей других направлений. Возможно ли, что мода на теорию струн начинает спадать? На мой взгляд, число теоретиков-струнников долгие годы оставалось избыточным. Несомненно, в этой теории немало занимательных аспектов, которые заслуживают дальнейшего исследования. В частности, это касается связи данной теории со многими областями математики, где она определенно оказала очень положительный эффект. В то же время она настолько угнетающе повлияла на разработки в области фундаментальной физики, что результаты получились просто опустошительными и, на мой взгляд, затормозили развитие других областей, которые, в конечном счете, могли бы оказаться более успешными. Я думаю, что теория струн — примечательный пример, сравнимый, пожалуй, с некоторыми важнейшими заблуждениями прошлого, рассмотренными в разделе 1.2 (речь шла о том, как мода негативно повлияла на разработки в области фундаментальной физики).

При всем вышесказанном должен четко отметить, что развитие модных идей, тем не менее, может принести истинную пользу. Как правило, научные идеи остаются модными лишь в том случае, если они обладают математической связностью и хорошо подтверждаются наблюдениями. Однако насколько это верно для теории струн — вопрос как минимум спорный. Все-таки в случае с квантовой гравитацией считается нормальным, что любая эмпирическая проверка лежит далеко за пределами любого реалистичного эксперимента, который можно было бы сегодня поставить. Таким образом, исследователи должны практически целиком полагаться на теоретические рассуждения, почти не получая подсказок от природы. Причина, которой часто объясняют такой пессимизм, заключается в порядке планковской энергии, которая на уровне взаимодействия элементарных частиц недостижимо превышает возможности современных технологий (см. разделы 1.1, 1.5 и 1.10). Теоретики, занимающиеся квантовой гравитацией, не имея никакой надежды получить экспериментальное подтверждение или опровержение своих теорий, вынуждены полагаться на *математическую* составляющую, и именно ощущение математической силы и красоты служит основным критерием для суждения о предмете и правдоподобии той или иной гипотезы.

Подобные теории, проверка которых лежит далеко за пределами экспериментов, осуществимых в настоящее время, не подпадают под обычные научные критерии «эмпирической оценки», и оценка на уровне математики (плюс некоторая базовая физическая обоснованность) приобретает все большую важность. Разумеется, ситуация серьезно изменится, если удастся найти такую схему, которая не просто давала бы согласованную математическую структуру, но и позволяла бы прогнозировать новые физические явления, которые впоследствии в точности подтверждались бы наблюдениями. Как раз такую схему я предлагаю в разделе 2.13 (и в разделе 4.2), при этом объединение гравитационной и квантовой теорий потребует некоторых модификаций последней, но такой вариант вполне поддается экспериментальной проверке уже при помощи технологий, которые будут реализуемы в недалеком будущем. Если таким образом появится версия квантовой гравитации, которую можно будет адекватно экспериментально проверить, и такие эксперименты ее подтвердят, то вполне можно ожидать, что она по праву получит серьезное научное признание. Соответственно, это будет не «мода» в том смысле, как я описываю ее здесь, а подлинный научный прогресс. В теории струн мы ничего подобного не увидим.

Также можно ожидать, что при отсутствии однозначных экспериментов те гипотезы квантовой гравитации, которые не обладают математической связностью, вряд ли могли бы сохраниться, поэтому модность некоторой гипотезы может считаться признаком ее ценности. Однако мне кажется опасным чрезмерно доверять подобным чисто математическим суждениям. Обычно математиков не

слишком волнует правдоподобность и даже непротиворечивость физической теории как *вклада в наши представления о физическом мире*. Действительно, математик будет судить о таком вкладе в зависимости от того, насколько он помогает развивать новые математические концепции и мощные приемы для достижения математической истины.

Этот фактор был особенно важен для теории струн и, несомненно, сыграл важнейшую роль в поддержке ее модности. Действительно, идеи теории струн внесли колоссальный вклад в различные области чистой математики. Особенно характерен пример, приведенный в следующей цитате из электронного письма. Это письмо в начале 2000-х годов отправил мне заслуженный математик Ричард Томас из Имперского колледжа Лондона, отвечая на мой вопрос о математическом статусе одной сложной задачи, сформулированной на основе выкладок теории струн [см. Candelas et al., 1991]:

> Должен самым решительным образом подчеркнуть глубину некоторых подобных дуальностей — они постоянно удивляют нас новыми прогнозами. Они демонстрируют структуры, которые считались просто немыслимыми. Математики неоднократно уверенно утверждали, что подобные вещи невозможны, но такие люди, как Канделас, де ла Осса и др., доказали их неправоту. Каждый из сделанных прогнозов, который удалось надлежащим образом математически интерпретировать, оказался верным. Причем до сих пор это происходило *не* по каким-то *концептуальным* математическим причинам — мы не представляем, почему они верны, просто независимо вычисляем обе части и действительно находим по обе стороны одни и те же структуры, симметрии и ответы. Математик не может считать такие вещи совпадениями, они должны иметь высшую причину. Причина заключается в *предположении, что эта большая математическая теория описывает природу...*

Данная конкретная проблема, которую имеет в виду Томас, касается некоторых глубоких математических идей. Они возникли в результате того, каким образом была решена одна проблема, с которой столкнулась теория струн. Об этом повествует замечательная история, которую я расскажу ближе к концу следующего раздела.

1.13. М-теория

Особое достоинство теории струн, которое подчеркивалось на заре ее развития, заключалось в том, что она была призвана дать нам *единую* схему физики нашего мира. Однако эта надежда, которая провозглашалась на протяжении многих

лет, по-видимому, рухнула, когда оказалось, что существует *пять* различных видов теории струн. Они стали именоваться *Тип I, Тип IIA, Тип IIB, Гетеротическая O(32)* и *Гетеротическая* $E_8 \times E_8$ (здесь я даже не буду пытаться объяснять эти термины [Greene, 1999], хотя гетеротические модели вкратце обсуждались в разделе 1.9). Такое множество вариантов беспокоило теоретиков-струнников. Однако в 1995 году выдающийся теоретик Эдвард Виттен прочитал в Университете Южной Калифорнии исключительно важную лекцию, в которой обобщил замечательную совокупность идей. Согласно этим идеям, определенные преобразования, так называемые *дуальности*, указывают на неочевидную *эквивалентность* между различными теориями струн. Впоследствии эту лекцию объявили прологом «второй струнной революции» (первая началась с работ Грина и Шварца, упомянутых в разделе 1.9, когда введение суперсимметрии позволило снизить размерность пространства-времени с 26 до 10; см. разделы 1.9 и 1.14). Итак, согласно данной идее, за всеми этими явно различными вариантами теории струн лежит более глубокая и, возможно, единая теория — правда, пока не сформулированная в виде однозначной математической картины, — которую Виттен окрестил *М-теорией*. («М» в данном случае расшифровывается как «мастер», «матрица», «мистерия», «мать» или еще несколько вариантов в зависимости от предпочтений исследователя).

Одно из свойств М-теории таково: кроме одномерных струн (и их двумерных пространственно-временных историй), она требует рассматривать обладающие высшими измерениями структуры, которые обычно называют *бранами*. Брана сравнима с двумерной мембраной, но обладает p пространственными измерениями. Соответственно, такая p-брана имеет $p + 1$ пространственно-временных измерений. (На самом деле ранее подобные p-браны изучались другими специалистами независимо от М-теории [Becker et al., 2006].) Вышеупомянутые дуальности работают лишь потому, что браны с разными размерностями можно менять друг на друга одновременно с различными пространствами Калаби — Яу, привлекаемыми здесь для обеспечения дополнительных пространственных измерений. При этом вся концепция трактуется в более широком смысле, чем в теории струн, — естественно, поэтому и потребовалось новое название «М-теория». Можно отметить, что красивое исходное соотнесение струн с комплексными кривыми (то есть с римановыми поверхностями, которые упоминались в разделах 1.6 и 1.12 и описаны в разделе А.10), которая стала ключом к исходной привлекательности и успехам теории струн, отвергается для бран высокой размерности. С другой стороны, в таких идеях, несомненно, присутствует иная математическая красота и некоторая экстраординарная математическая сила (о чем можно заключить из ремарки Ричарда Томаса, процитированной в конце раздела 1.12), коренящаяся в этих примечательных дуальностях.

Здесь было бы полезно перейти к конкретным примерам и рассмотреть поразительную возможность применения одного из аспектов этих дуальностей — так называемой *зеркальной симметрии*. Эта симметрия каждому пространству Калаби — Яу ставит в соответствие другое пространство Калаби — Яу путем замены определеных параметров, называемых *числами Ходжа*, которые описывают конкретную «форму» каждого пространства Калаби — Яу. Пространства Калаби — Яу представляют собой своеобразные (вещественные) 6-многообразия, которые также можно интерпретировать как комплексные 3-многообразия, то есть в этих 6-многообразиях содержатся *комплексные структуры*. В принципе, *комплексное n-многообразие* (см. последнюю часть раздела А.10) — это просто аналог обычного вещественного n-многообразия (раздел А.5), где множество $\mathbb{R}$ вещественных чисел заменяется множеством $\mathbb{C}$ комплексных чисел (см. раздел А.9). Всегда можно предложить другую трактовку комплексного n-многообразия — как вещественного $2n$-многообразия, наделенного так называемой *комплексной структурой*. Однако лишь в благоприятных обстоятельствах вещественному $2n$-многообразию можно присвоить такую комплексную структуру, при которой оно может быть реинтерпретировано как комплексное n-многообразие (раздел А.10). Более того, в каждом пространстве Калаби — Яу есть и иные структуры, называемые *симплектическими* (именно такую структуру имеют *фазовые пространства*, упоминаемые в разделе А.6). Зеркальная симметрия позволяет проделывать очень странный математический трюк: *взаимозаменять* комплексные и симплектические структуры!

Тот вариант применения зеркальной симметрии, о котором мы поговорим здесь, возник в связи с проблемой, над которой некоторые специалисты по чистой математике (алгебраической геометрии) к тому моменту работали уже несколько лет. Два норвежских математика — Гейр Эллингсруд и Стейн Арильд Стрёмме — разработали метод подсчета рациональных кривых для конкретного типа комплексных 3-многообразий (так называемых квинтик, то есть многообразий, определяемых комплексным полиномиальным уравнением 5-й степени), которые на самом деле представляют собой примеры пространства Калаби — Яу. Вспомните (см. разделы 1.6 и А.10), что комплексная кривая — это так называемая *риманова поверхность*; такая комплексная кривая именуется *рациональной*, если топологически данная поверхность представляет собой *сферу*. В алгебраической геометрии рациональные кривые могут принимать формы с возрастающей «скрученностью», где простейшая форма — это обычная прямая линия (порядок 1), а следующая по сложности — комплексное коническое сечение (порядок 2). Далее имеем рациональные кубические кривые (порядок 3), квартики (порядок 4) и т. д., причем для каждого последующего порядка должно существовать точное, вычислимое, конечное число рациональных кривых. (*Порядок*

кривой, в настоящее время зачастую именуемый ее степенью, в том случае, если эта кривая находится в плоском окружающем n-пространстве, — это число точек, в которых она пересекает расположенную общим образом $(n-1)$-плоскость.) При помощи сложных компьютерных вычислений норвежцы нашли ряд чисел:

2875,

609250,

2682549425

для порядков 1, 2 и 3 соответственно, но продвинуться дальше оказалось очень непросто, поскольку неимоверно усложнялись те методы, которыми они пользовались при работе.

Узнав об этих результатах, эксперт по теории струн Филип Канделас (Philip Candelas) и его коллеги попытались применить процедуры зеркальной симметрии из М-теории, отметив, что могут иным способом сосчитать кривые в зеркальном пространстве Калаби — Яу. В этом дуальном пространстве можно не подсчитывать рациональные кривые, а воспользоваться гораздо более простым вычислением (при котором система рациональных кривых «отображается» на значительно более удобное семейство). Согласно зеркальной симметрии, такой метод должен был дать те же самые числа, что пытались вычислить Эллингсруд и Стрёмме. Канделас с коллегами нашли такую последовательность чисел:

2875,

609250,

317206375.

Поразительно, что первые два числа совпадают с данными норвежцев, хотя третье число странным образом получилось совершенно иным.

Сначала математики утверждали, что поскольку аргументы зеркальной симметрии просто взяты из каких-то физических гипотез без четкого математического обоснования, совпадение на порядках 1 и 2 принципиально должно быть простой случайностью, поэтому нет никаких причин доверять числам высшего порядка, полученным методом зеркальной симметрии. Однако впоследствии в компьютерной программе норвежцев обнаружилась ошибка, и после ее исправления получился ответ: 317206375 — в точности такой, как предполагала зеркальная симметрия! Более того, метод зеркальной симметрии легко расширяется, поэтому можно вычислить последовательность рациональных кривых соответственно для порядков 4, 5, 6, 7, 8, 9 и 10, получив следующие ответы:

242467530000,
229305888887625,
248249742118022000,
295091050570845659250,
375632160937476603550000,
503840510416985243645106250,
704288164978454686113488249750.

Несомненно, это был примечательный образец косвенного доказательства в пользу идеи о зеркальной симметрии, возникшей из желания продемонстрировать, что две теории струн, очевидно, отличавшиеся коренным образом, тем не менее в некотором глубоком смысле могут быть «одинаковы», когда два рассматриваемых различных пространства Калаби — Яу взаимно дуальны в вышеизложенном смысле. Последующие работы различных математиков[1] [Givental, 1996] в целом показали, что закономерность, принятая за обычную физическую гипотезу, на самом деле является строгой математической истиной. Однако ранее математики даже не догадывались о том, что идея вроде зеркальной симметрии может оказаться истинной, о чем убедительно свидетельствуют комментарии Ричарда Томаса, приведенные в конце раздела 1.13. Для математика, возможно, не знакомого с тонкостями физического смысла этих идей, подобное явление кажется подарком самой природы, может быть, даже мимолетным напоминанием о головокружительных днях на излете XVII века, когда волшебство дифференциального исчисления, разработанного Ньютоном и другими учеными, стало угадываться в устройстве природы, а также вовсю проявляться внутри самой математики.

Разумеется, среди нас, физиков-теоретиков, есть много таких, кто действительно верит: устройство природы с огромной точностью описывается математикой, которая, вероятно, обладает не только мощью, но и тонкой организацией, — это убедительно демонстрирует электромагнетизм Максвелла, гравитационная теория Эйнштейна и квантовый формализм Шрёдингера, Гейзенберга и Дирака. Следовательно, нас, скорее всего, также должны поразить достижения зеркальной симметрии, и нам стоит расценивать их как своеобразное свидетельство: физическая теория, породившая столь мощную и утонченную математику, с большой вероятностью будет представлять некую глубокую ценность *и* для физики. Однако нужно быть очень осторожными с такими выводами. Известно множество примеров мощных и впечатляющих математических теорий, которые никто и не пытался всерьез увязать с устройством физического мира. Характер-

[1] Концевич, Гивенталь, Лян, Лю, Яу.

ный пример — чудесное математическое достижение Эндрю Уайлса, который, во многом опираясь на имеющиеся наработки коллег и не без помощи Ричарда Тейлора, в 1994 году наконец решил задачу возрастом 350 лет, известную под названием *Великая теорема Ферма*. Уайлс определил (и это был ключевой элемент его доказательства), что можно получить некое подобие эффекта зеркальной симметрии, а именно рассчитать две последовательности чисел, которые будут получены при помощи двух совершенно разных математических процедур, но при этом окажутся идентичными. В случае Уайлса идентичность двух последовательностей чисел формулировалась в виде утверждения, известного под названием *гипотеза Таниямы — Симуры*. Уайлс смог успешно доказать Великую теорему Ферма, воспользовавшись его же методами, чтобы сформулировать недостающую часть рассуждений (полное доказательство было сделано немного позже, в 2001 году, Брейлем, Конрадом, Даймондом и Тейлором, опиравшимися на методы, разработанные Уайлсом; см. [Breuil et al., 2001]). В чистой математике есть и множество других подобных достижений; при этом очевидно, что для совершенно новой физической теории одной такой математики явно недостаточно, несмотря на всю ее филигранность, сложность и на некоторые подлинно волшебные свойства, которыми математика может обладать. Важнейшую роль играют физическая обоснованность и экспериментальное подтверждение — только так мы можем убедиться, что теория действительно имеет какую-то связь с реальным устройством физического мира. Эти вопросы в самом деле являются ключевыми для проблемы, которую мы рассмотрим далее; эта проблема сыграла важнейшую роль при разработке теории струн.

1.14. Суперсимметрия

До сих пор я позволял себе такую роскошь и игнорировал ключевой результат суперсимметрии — именно суперсимметрия позволила Грину и Шварцу уменьшить размерность пространства-времени в теории струн с 26 до 10. Суперсимметрия играет основополагающую роль в современной теории струн и во многих других случаях. На самом деле суперсимметрия — это феномен, оказавшийся важным во многих физических областях, довольно далеких от теории струн. Действительно, суперсимметрию можно считать очень модной идеей современной физики, и, следовательно, она сама по себе заслуживает серьезного обсуждения в этой главе! Да, стимул к разработке этой идеи во многом обусловлен требованиями теории струн, но свой модный статус она в значительной степени приобрела сама по себе.

Итак, что же такое суперсимметрия? Чтобы объяснить это, мне придется вернуться к обсуждению основ физики частиц, о которой мы говорили в разделах 1.3

и 1.6. Как вы помните, существуют различные семейства частиц, обладающих массой, например лептоны и адроны, а также другие частицы, например безмассовый фотон. На самом же деле существует еще более простая классификация частиц всего на два типа, пересекающаяся с теми, что встречались нам выше. Это подразделение частиц на фермионы и бозоны, вкратце упомянутое в разделе 1.6.

Один из способов показать разницу между фермионами и бозонами — это первые уподобить тем элементарным частицам, которые известны нам из классической физики (электронам, протонам, нейтронам и т. д.), а вторые сравнить с переносчиками взаимодействий между частицами (так, фотоны переносят электромагнитное взаимодействие, а W-бозоны и Z-бозоны — слабое взаимодействие; сущности под названием *глюоны* переносят сильное взаимодействие). Однако это не самое четкое различие, особенно потому, что существуют еще и частицеподобные пионы, каоны и другие бозоны, упоминавшиеся в разделе 1.3. Более того, многие атомы с хорошим приближением можно рассматривать как бозоны, поскольку во многих ситуациях такие составные объекты действительно ведут себя как отдельные частицы. Бозонные атомы не слишком отличаются от фермионных: и те и другие напоминают классические частицы.

Однако давайте оставим в стороне проблему составных объектов и вопрос о том, допустимо или нет рассматривать их как отдельные частицы. Поскольку интересующие нас объекты можно приравнивать к отдельным частицам, разница между фермионом и бозоном сводится к так называемому *принципу запрета Паули*, применимому только к фермионам. Согласно этому принципу, два фермиона не могут одновременно находиться в одном и том же квантовом состоянии, а два бозона — могут. Грубо говоря, принцип Паули постулирует, что два идентичных фермиона не могут находиться ровно друг на друге, поскольку воздействуют один на другой и фактически отталкиваются, стоит им чрезмерно сблизиться. В свою очередь, бозоны обладают определенным сродством к другим бозонам своего типа и могут располагаться один на другом (что и происходит в агрегатном состоянии, представляющем собой множество бозонов и известном под названием *конденсат Бозе — Эйнштейна*). Такие конденсаты описаны в работе Кеттерле [Ketterle, 2002]. Более подробную информацию см. у Форда [Ford, 2013].

Вскоре я вернусь к этому весьма любопытному аспекту квантово-механических частиц и попытаюсь пояснить эту достаточно зыбкую характеристику, которая, конечно же, дает весьма приблизительное представление о разнице между бозонами и фермионами. Несколько более четкое различие прослеживается, если посмотреть на частоту вращения частицы. Странно, что любая (невозбужденная) квантовая частица может вращаться только с определенной фиксированной

скоростью, свойственной частице данного конкретного типа. Частоту вращения стоит воспринимать не как угловую скорость, а скорее как *момент импульса* — такой мерой вращения обладает объект, движущийся без влияния внешних сил, причем эта величина не изменяется в ходе вращения объекта. Представьте себе мяч для бейсбола или крикета, который летит по воздуху и вращается. Или фигуристку, которая крутится на одном коньке. В каждом из этих случаев мы увидим, что спин (то есть момент импульса) самоподдерживается, и так продолжалось бы до бесконечности при отсутствии каких-либо внешних сил (например, силы трения).

Пример с фигуристкой, пожалуй, более хорош, поскольку мы сами можем убедиться: угловая скорость уменьшается, если спортсменка раскинет руки, и увеличивается, если она прижмет их к телу. Однако при таком процессе сохраняется постоянный *момент импульса*, который при заданной угловой скорости будет выше, если распределение масс (скажем, положение рук фигуристки) будет дальше от оси вращения, и меньше, если ближе (рис. 1.36). Таким образом, вытягивание рук по швам должно компенсироваться увеличением скорости вращения, чтобы момент импульса мог оставаться постоянным.

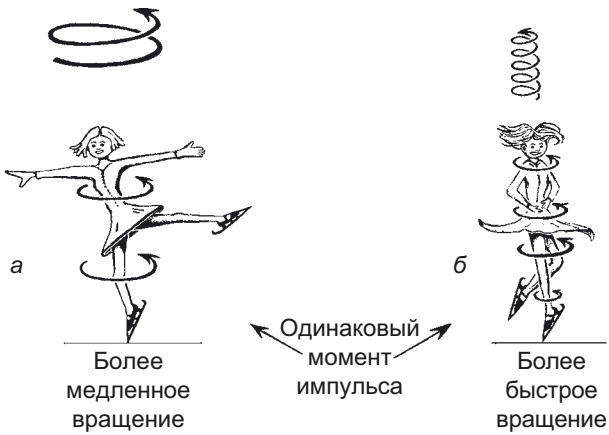


Рис. 1.36. Момент импульса при физических процессах сохраняется. Это иллюстрирует пример с кружащейся фигуристкой, которая может увеличить скорость вращения, просто прижав руки к телу. Дело в том, что угловое движение на большем расстоянии дает больший вклад в момент импульса, чем на меньшем

Итак, существует феномен момента импульса, применимый ко всем компактным изолированным телам. Он также применим к отдельным квантовым частицам, но на квантовом уровне законы природы немного странные, и требуется время,

чтобы к ним привыкнуть. Если говорить об отдельной квантовой частице, то для каждого типа частиц обнаруживается следующее: *величина* момента импульса всегда равна одному и тому же фиксированному числу, в какой бы ситуации эта частица ни оказалась. Направление оси вращения не обязано в разных ситуациях оставаться одинаковым, но направление спина изменяется странным квантово-механическим образом, о чем мы подробнее поговорим в разделе 2.9. На данный момент достаточно знать, что если рассмотреть распределение возможных значений проекций спина на какое-то выбранное направление, то для бозона эти значения будут кратными $\hbar$, где $\hbar$ — редуцированная постоянная Планка h, называемая также *постоянной Дирака* (см. раздел 2.11):

$$\hbar = \frac{h}{2\pi}.$$

Следовательно, значение проекции спина бозона на любое заданное направление должно быть равно одному из значений:

$$\ldots, -2\hbar, -\hbar, 0, \hbar, 2\hbar, 3\hbar, \ldots$$

Однако у фермионов значение спина в любом заданном направлении будет отличаться от этих чисел на $1/2\hbar$, то есть относиться к последовательности:

$$\ldots, \ -\frac{3}{2}\hbar, \ -\frac{1}{2}\hbar, \ \frac{1}{2}\hbar, \ \frac{3}{2}\hbar, \ \frac{5}{2}\hbar, \ \frac{7}{2}\hbar, \ \ldots$$

(поэтому такое значение всегда будет полуцелым кратным $\hbar$). Это любопытное свойство квантовой механики на практике мы подробнее рассмотрим в разделе 2.9.

Существует знаменитая теорема, доказанная в контексте КТП, — так называемая теорема о связи спина со статистикой [Streater and Wightman, 2000], фактически постулирующая эквивалентность двух описаний сущности бозонов и фермионов. Точнее, с математической точки зрения она получается значительно более обширной, чем упоминавшийся выше принцип запрета Паули, — она описывает своеобразную статистику, которой должны подчиняться бозоны и фермионы. Сложно объяснить ее удовлетворительным образом, не углубляясь в квантовый формализм еще глубже, чем уже было сделано в этой главе. Тем не менее я постараюсь хотя бы приблизительно передать ее суть.

Как вы помните, квантовые амплитуды, упоминавшиеся в разделе 1.4 (см. также разделы 2.3–2.9), — это комплексные числа, которые мы пытаемся получить при вычислениях КТП (см. раздел 1.5), причем на базе этих чисел вычисляются вероятности квантовых измерений (по правилу Борна; см. раздел 2.8). В любом

квантовом процессе амплитуда будет функцией всех параметров, описывающих все квантовые частицы, вовлеченные в процесс. Амплитуду также можно понимать как значение шрёдингеровской волновой функции, которую мы обсудим в разделах 2.5–2.7. Если P_1 и P_2 — две идентичные частицы, вовлеченные в процесс, то такая амплитуда (или волновая функция) ψ будет функцией $\psi(\mathbf{Z}_1, \mathbf{Z}_2)$ соответствующих совокупностей параметров $\mathbf{Z}_1$ и $\mathbf{Z}_2$ для этих двух частиц. (Здесь я использую одну жирную букву $\mathbf{Z}$ для охвата всех параметров каждой частицы: координаты, момента, спина и т. д.) Нижний индекс (1 или 2) служит ссылкой на конкретную частицу. Имея n частиц P_1, P_2, P_3, ..., P_n (неважно, идентичных или нет), мы будем иметь n таких наборов параметров: $\mathbf{Z}_1$, $\mathbf{Z}_2$, $\mathbf{Z}_3$, ..., $\mathbf{Z}_n$. Итак, ψ есть функция всех этих переменных:

$$\psi = \psi(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n).$$

Теперь, если частицы $\mathbf{Z}_1$ и $\mathbf{Z}_2$ относятся к одному и тому же типу и эти частицы — бозоны, мы всегда обнаруживаем симметрию:

$$\psi(\mathbf{Z}_1, \mathbf{Z}_2, \ldots) = \psi(\mathbf{Z}_2, \mathbf{Z}_1, \ldots),$$

поэтому если поменять частицы P_1 и P_2 местами, это никак не скажется на амплитуде (или волновой функции). Но если частицы $\mathbf{Z}_1$ и $\mathbf{Z}_2$ относятся к одному и тому же типу и эти частицы — фермионы, то имеем:

$$\psi(\mathbf{Z}_1, \mathbf{Z}_2, \ldots) = -\psi(\mathbf{Z}_2, \mathbf{Z}_1, \ldots),$$

поэтому если частицы P_1 и P_2 поменять местами, то изменится и знак амплитуды (или волновой функции). Можно отметить, что если обе частицы (P_1 и P_2) находятся в одном и том же состоянии, то $\mathbf{Z}_1 = \mathbf{Z}_2$, то есть $\psi = 0$ (поскольку ψ равно отрицательному ψ). По правилу Борна (см. раздел 1.4) получаем, что $\psi = 0$ подразумевает нулевую вероятность. Это выражение описывает принцип Паули: невозможно найти два таких фермиона идентичного типа, которые будут находиться в одном и том же состоянии. Если все n частиц идентичны, то для n бозонов получаем симметрию, распространяющуюся на перестановку слагаемых в любой паре:

$$\psi(\ldots, \mathbf{Z}_i, \ldots, \mathbf{Z}_j, \ldots) = \psi(\ldots, \mathbf{Z}_j, \ldots, \mathbf{Z}_i, \ldots),$$

а для n фермионов — антисимметрию в любой паре:

$$\psi(\ldots, \mathbf{Z}_i, \ldots, \mathbf{Z}_j, \ldots) = -\psi(\ldots, \mathbf{Z}_j, \ldots, \mathbf{Z}_i, \ldots).$$

Симметрия или антисимметрия, выраженная соответственно в первом и во втором вышеприведенных уравнениях, объясняет статистические различия между

бозонами и фермионами. Если «подсчитать» число разных состояний, в которых может находиться множество однотипных бозонов, то перемену двух бозонов местами не следует считать возникновением нового состояния. Этот метод подсчета дал начало так называемой *статистике Бозе — Эйнштейна* (или просто *статистике Бозе*, от фамилии которого происходит и термин *бозон*). С фермионами происходит все то же самое, за исключением странного изменения знака амплитуды. Поэтому к ним применима *статистика Ферми — Дирака* (или просто *статистика Ферми*, в честь которого был назван *фермион*). Такая ситуация влечет за собой множество квантово-механических последствий, наиболее очевидное из которых — принцип запрета Паули. Отметим, что, как с бозонами, так и с фермионами, взаимное изменение мест двух одинаковых частиц никак не влияет на квантовое состояние (всего лишь меняется знак волновой функции, что не сказывается на физическом состоянии, поскольку при умножении на -1 фактически происходит домножение на фазовый множитель $\times e^{i\theta}$, где $\theta = \pi$; см. раздел 1.8). Соответственно, квантовая механика требует, чтобы две частицы одного и того же вида *были* идентичны! Этот пример показывает важность эйнштейновского возражения против исходного варианта калибровочной теории, предложенного Вейлем, где «калибровка» на самом деле означает изменение шкалы (см. раздел 1.8).

Все это — стандартная квантовая механика, имеющая множество следствий, прекрасно подтвержденных наблюдениями. Однако многие физики считают, что должна существовать симметрия нового вида, позволяющая преобразовывать бозоны в фермионы и напоминающая те симметрии, что соотносят лептоны друг с другом и порождают калибровочную теорию слабых взаимодействий. Либо она может напоминать симметрии, связывающие друг с другом различные кварки и лежащие в основе калибровочной теории сильных взаимодействий (см. раздел 1.3 и последний абзац раздела 1.8). Такая новая симметрия не может быть обычной, поскольку статистика для двух этих различных семейств частиц различается. Соответственно, физики обобщили обычную симметрию до симметрии нового рода, которую стали называть *суперсимметрией* [Kane and Shifman, 2000]. В суперсимметрии симметричные состояния бозонов преобразуются в антисимметричные состояния фермионов, и наоборот. При этом задействуются странные «числа» — так называемые *генераторы суперсимметрии*, обладающие следующим свойством: при перемножении двух таких чисел, скажем $\boldsymbol{\alpha}$ и $\boldsymbol{\beta}$, получим результат, отличающийся *знаком* от итога их же перемножения в обратном порядке:

$$\boldsymbol{\alpha\beta} = -\boldsymbol{\beta\alpha}$$

(на самом деле некоммутативность операций, когда $\mathbf{AB} \neq \mathbf{BA}$, — обычное явление в квантовом формализме; см. раздел 2.13). Именно этот минус позволяет

преобразовывать статистику Бозе — Эйнштейна в статистику Ферми — Дирака, и наоборот.

Для того чтобы лучше рассказать о некоммутирующих величинах, я должен немного подробнее пояснить общий формализм квантовой механики (и КТП). В разделе 1.4 мы познакомились с понятием квантового состояния системы; такие состояния (**Ψ**, **Φ** и т. д.) подчиняются законам комплексного векторного пространства (см. разделы A.3 и A.9). Важную роль тут играет теория *линейных операторов* — мы вернемся к ней позже (см., в частности, разделы 1.16, 2.12, 2.13 и 4.1). Такой оператор **Q**, воздействующий на квантовые состояния **Ψ**, **Φ** и т. д., имеет свойство, называемое *линейностью*, это означает, что:

$$\mathbf{Q}(w\Psi + z\Phi) = w\mathbf{Q}(\Psi) + z\mathbf{Q}(\Phi),$$

где w и z — (постоянные) комплексные числа. К квантовым операторам относятся, например, операторы координаты и импульса — **x** и **p** — и оператор энергии **E**, о котором мы поговорим в разделе 2.13, а также оператор спина (см. раздел 2.12). В стандартной квантовой механике *измерения* всегда выражаются в терминах линейных операторов. Это будет объяснено в разделе 2.8.

Генераторы суперсимметрии, например **α** и **β**, также являются линейными операторами, но их задача в КТП — воздействовать на другие линейные операторы, именуемые *операторами рождения и уничтожения*, которые играют центральную роль в алгебраическом устройстве КТП. Символом **a** часто обозначают оператор уничтожения, а символом $\mathbf{a}^\dagger$ — соответствующий оператор рождения. Если у нас есть конкретное квантовое состояние **Ψ**, то $\mathbf{a}^\dagger\mathbf{\Psi}$ будет состоянием, полученным из **Ψ** путем добавления частицы, находящейся в состоянии, обозначаемом $\mathbf{a}^\dagger$; аналогично **aΨ** будет состоянием **Ψ**, но без конкретной частицы (предполагается, что такая операция удаления частицы возможна; в противном случае получим просто $\mathbf{a\Psi} = \mathbf{0}$). Генератор суперсимметрии, например **α**, в таком случае повлияет на оператор рождения (или уничтожения) бозона, преобразуя его в соответствующий оператор уничтожения (или рождения) фермиона, и наоборот.

Обратите внимание: если в выражении $\boldsymbol{\alpha\beta} = -\boldsymbol{\beta\alpha}$ положить $\boldsymbol{\beta} = \boldsymbol{\alpha}$, то можно сделать вывод, что $\boldsymbol{\alpha}^2 = \mathbf{0}$ (поскольку $\boldsymbol{\alpha}^2 = (-\boldsymbol{\alpha})^2$). Соответственно, мы никогда не сможем возвести генератор суперсимметрии в степень (> 1). Возникает любопытный эффект: если предположить, что есть конечное число N генераторов суперсимметрии **α**, **β**, ..., **ω**, то любое алгебраическое выражение X можно записать без степеней этих величин:

$$X = X_0 + \alpha X_1 + \beta X_2 + \cdots + \omega X_N + \alpha\beta X_{12} + \cdots + \alpha\omega X_{1N} + \cdots + \alpha\beta\cdots\omega X_{12\cdots N},$$

поэтому в сумме получится всего 2^N членов (по одному на каждый возможный набор разных элементов в наборе генераторов суперсимметрии). Это выражение явно демонстрирует единственный вид зависимости от генераторов суперсимметрии, которая может возникать, хотя значения некоторых X в правой части могут быть равны нулю. Первый член X_0 иногда именуется *телом*, а остальные $\alpha X_1 + \cdots + \alpha\beta\cdots\omega X_{12\cdots N}$, среди которых присутствует как минимум один генератор суперсимметрии, — *душой*. Обратите внимание: как только часть выражения переходит в душу, ее уже невозможно перенести обратно в тело, умножая на другие такие выражения. Следовательно, тело любого алгебраического выражения существует особняком, давая нам совершенно корректное *классическое* вычисление, в котором мы просто не учитываем душу. Так можно обосновать алгебраические и геометрические соображения, например, из раздела 1.11, где суперсимметрия попросту игнорируется.

Требование суперсимметрии определенным образом направляет нас при выборе физической теории. Условие, согласно которому предлагаемая теория должна быть суперсимметричной, — очень сильное. Такое ограничение придает теории определенный баланс между ее бозонными и фермионными составляющими. Эти составляющие соотносятся друг с другом при помощи суперсимметричной операции (то есть операции, составляемой при помощи генераторов суперсимметрии, таких как X, упомянутый выше). Считается, что это ценный актив для создания КТП, предназначенной для правдоподобного моделирования природы, чтобы такая теория не страдала от неуправляемых расходимостей. Если мы требуем, чтобы теория была суперсимметричной, то значительно повышаем шансы на то, что она будет перенормируемой (см. раздел 1.5), и всерьез претендуем на то, что она позволит нам получить конечные (нерасходящиеся) результаты. При суперсимметрии расходимости фермионной и бозонной частей теории фактически взаимоуничтожают друг друга.

По-видимому, именно в этом и заключается одна из основных причин (кроме связи с теорией струн), по которой суперсимметрия приобрела такую популярность в физике частиц. Однако если бы природа в самом деле была суперсимметрична (скажем, с *одним* генератором суперсимметрии), то любой элементарной частице соответствовала бы другая — так называемый *суперсимметричный партнер*, — обладающая той же массой, что и первая частица. Таким образом, каждая пара суперсимметричных партнеров состояла бы из бозона и фермиона с одинаковой массой. Электрону сопутствовал бы бозон *сэлектрон*, а бозонный *скварк* сопутствовал бы кварку любого типа. Еще должны были бы существовать безмассовые *фотино* и *гравитино*, то есть фермионы, которые сопутствовали бы фотону и гравитону соответственно, а также *уино* и *зино* — парные вышеупомянутым W- и Z-бозонам. В действительности вся ситуация

тревожнее, чем относительно простой случай с единственным генератором суперсимметрии. Если бы существовало N генераторов суперсимметрии, где $N > 1$, то элементарные частицы были бы не просто спарены таким образом, но и присутствовали в количестве $2N$ в каждой суперсимметричной совокупности частиц (мультиплете) взаимных партнеров. Мультиплет наполовину состоял бы из бозонов, наполовину — из фермионов, и все эти частицы имели бы одинаковую массу.

Учитывая, каким угрожающим образом множатся все эти разновидности элементарных частиц (и, пожалуй, даже термин «элементарный» начинает казаться абсурдным), позволю читателю облегченно выдохнуть: на самом деле ни одного такого суперсимметричного множества частиц пока наблюдать не удалось! Однако этот эмпирический факт не смущает сторонников суперсимметрии, поскольку принято утверждать, что должен существовать какой-то механизм ее нарушения, приводящий к серьезному отклонению от точной суперсимметрии для частиц, наблюдаемых в природе. При этом массы частиц в каждом мультиплете могут очень сильно различаться. Таким образом, все эти суперсимметричные партнеры (соответствующие *единственному* найденному до сих пор члену каждой группы) должны обладать массами, недостижимыми на каких-либо уже имеющихся в нашем распоряжении ускорителях частиц!

Разумеется, сохраняется возможность, что все эти частицы, предсказанные суперсимметрией, все-таки существуют, а наблюдать их не удается по той простой причине, что они слишком массивны. Можно надеяться, что, когда БАК снова станет эксплуатироваться на полную мощность и заработает активнее, можно будет получить четкие свидетельства за или против суперсимметрии. Однако существуют и другие варианты суперсимметричных теорий, не дающие однозначного ответа, на каком уровне природы могут существовать механизмы нарушения суперсимметрии. На момент написания книги не было никаких доказательств существования суперсимметричных партнеров, однако ситуация по-прежнему далека от научного идеала, к которому предположительно стремятся большинство теоретиков: в идеале предлагаемая научная гипотеза, чтобы считаться подлинно научной, должна быть *фальсифицируемой* (этот знаменитый критерий выдвинул в 1963 году философ науки Карл Поппер). Остается неловкое ощущение, что даже если суперсимметрия на самом деле *ложна* и в природе отсутствует, а значит, *никаких* суперсимметричных партнеров никогда не удастся найти ни на БАК, ни на более современном и мощном ускорителе, то сторонники суперсимметрии все равно станут утверждать, что дело не в том, что реальные элементарные частицы не обладают суперсимметрией, а в том, что уровень, на котором нарушается суперсимметрия, должен быть *еще* выше современного, и для его наблюдения потребуется еще более мощная машина!

В действительности ситуация с научной опровержимостью, пожалуй, не так плоха. Последние результаты работы БАК, в частности примечательное открытие бозона Хиггса, который так долго не удавалось найти, не просто не дают никаких доказательств существования суперсимметричных партнеров какой-либо известной частицы, но и исключают наиболее простые модели суперсимметрии, с которыми ранее связывались большие надежды. Теоретические и эмпирические преграды вполне могут оказаться слишком серьезными, что не позволит довести до ума какую-либо теорию суперсимметрии из тех, что предлагались ранее. Все это может вывести теоретиков на новые и более многообещающие идеи о том, как могут соотноситься друг с другом семейства бозонов и фермионов. Также следует отметить, что модели, предусматривающие *более одного* генератора суперсимметрии, например очень популярная среди теоретиков *суперсимметричная теория Янга — Миллса*, где $N = 4$, еще дальше от согласия с результатами наблюдений, чем варианты с единственным генератором суперсимметрии.

Тем не менее идея суперсимметрии остается очень популярной среди физиков-теоретиков, и, как мы убедились, она также является ключевой составляющей теории струн. Вообще, сам выбор пространства Калаби — Яу в качестве предпочтительного многообразия $\mathcal{X}$ для описания дополнительных пространственных измерений (см. разделы 1.10 и 1.11) связан с тем, что оно обладает суперсимметричными свойствами. Можно выразить данное требование иначе: сказать, что в $\mathcal{X}$ существует так называемое (ненулевое) *спинорное поле*, остающееся постоянным во всем $\mathcal{X}$. Термин *спинорное поле* означает один (не обязательно единственный) из основных видов физических полей (см. разделы A.2 и A.7), при помощи которого можно описать волновую функцию фермиона (см. разделы 2.5 и 2.6; подробнее о спинорных полях см., например, [Penrose and Rindler, 1984]. Случай для высших измерений описан в приложении к [Penrose and Rindler, 1986]).

Это постоянное спинорное поле может выступать в качестве генератора суперсимметрии, и с его помощью можно выразить суперсимметричную природу всех высших измерений пространства-времени. Оказывается, что такое требование суперсимметрии гарантирует, что общая энергия пространства-времени будет нулевой. Считается, что такая нулевая энергия должна быть основным состоянием всей Вселенной, причем есть мнение, что это состояние достигается именно благодаря ее суперсимметричности. По-видимому, в основе такого аргумента лежит следующая идея: пертурбации такого базового состояния с нулевой энергией должны повышать энергию Вселенной, поэтому пространственно-временная структура такой слегка возмущенной Вселенной будет вновь возвращаться в исходное нулевое состояние просто путем отдачи приобретенной таким образом энергии.

Однако должен сказать, что такая аргументация серьезно меня смущает. Учитывая мои соображения из раздела 1.10 и мысли, высказанные ранее в этой главе (о том, что тело любой суперсимметричной геометрии можно выделить и выразить на языке классической геометрии), кажется очень уместным считать такие пертурбации источником *классических* возмущений. Опираясь на выводы из раздела 1.11, следует признать, что абсолютное большинство таких классических возмущений будут приводить к *пространственно-временным сингулярностям* в крошечную долю секунды! По меньшей мере возмущения дополнительных пространственных измерений быстро возрастут так сильно, что фактически станут сингулярными еще до того, как смогут вступить в игру какие-либо члены высшего порядка из струнной константы α'. Согласно такой картине, пространство-время не уляжется в стабильное суперсимметричное базовое состояние, а скатится в сингулярность! Не вижу никаких разумных причин надеяться на то, что такую катастрофу можно было бы обратить вспять, независимо от суперсимметричной природы такого состояния.

1.15. AdS/CFT-соответствие

Хотя я и не уверен, что многие (или хотя бы некоторые) профессиональные струнные теоретики позволяли себе отступить от своих основных целей из-за аргументов вроде изложенных выше — речь об аргументах, рассмотренных в разделах 1.10 и 1.11, а также в конце раздела 1.14, и вообще о проблемах, связанных с функциональной свободой, затронутых в разделах А.2, А.8 и А.11, — в самые последние годы они устремились в существенно иные научные сферы, нежели те, что я описывал выше. Тем не менее проблемы избыточной функциональной свободы остаются очень актуальными, и мне кажется правильным обсудить в заключительной части этой главы наиболее значительные подобные разработки. В разделе 1.16 я очень кратко опишу таинственную территорию, куда нас, по-видимому, ведет тропа теории струн. Эту территорию называют *миры на бранах*, *ландшафт* и *болота*. Значительно больший математический интерес и соблазнительные связи с различными иными областями физики представляет так называемое *AdS/CFT-соответствие*, также именуемое *голографической гипотезой* или *дуальностью Малдасены*.

AdS/CFT-соответствие (см. [Ramallo, 2013; Zaffaroni, 2000; Susskind and Witten, 1998]) зачастую называется *голографическим принципом*. Я сразу должен подчеркнуть, что это не установленный принцип, а набор интересных феноменов, которые действительно получают некоторую эмпирическую математическую поддержку, но, на первый взгляд, как будто противоречат определенным серьезным аспектам функциональной свободы. В широком смысле идея голографиче-

ского принципа такова: существуют две физические теории абсолютно разных типов, одна из которых (как оказывается, вариант теории струн) определяется в некотором $(n+1)$-мерном регионе пространства-времени, так называемом *объемлющем пространстве*, а другая (более традиционная квантовая теория поля) определяется на n-мерной *границе* этого региона. На первый взгляд может сложиться впечатление, что такое соответствие маловероятно с точки зрения функциональной свободы, поскольку в теории объемлющего пространства, по-видимому, должна быть функциональная свобода $\infty^{A\infty^n}$ для некоторого A, тогда как может *показаться*, что в теории для границы региона будет гораздо меньше функциональной свободы — $\infty^{B\infty^{n-1}}$ для некоторого B. Такая ситуация возникала бы, если бы в обоих случаях мы имели дело с теориями пространства-времени более или менее обычных типов. Чтобы лучше понять основополагающие причины, по которым предполагается наличие такого соответствия, а также возможные сложности, которые могут быть связаны с предложенной гипотезой, будет полезно сначала рассмотреть некоторый контекст этой идеи.

Одна из самых ранних предпосылок этой идеи связана с хорошо обоснованным свойством термодинамики черных дыр — мы гораздо подробнее обсудим это свойство в главе 3. Речь идет о фундаментальной формуле Бекенштейна — Хокинга, описывающей энтропию черной дыры (этой формулой мы плотнее займемся в разделе 3.6). Эта формула свидетельствует о том, что энтропия черной дыры пропорциональна площади ее поверхности. Итак, грубо говоря, энтропия объекта, находящегося в абсолютно произвольном («термализованном») состоянии, в принципе, равна общему числу степеней свободы у данного объекта. (Эта идея точнее описывается мощной обобщенной формулой Больцмана, о которой мы подробнее поговорим в разделе 3.3.) Странное свойство этой формулы для энтропии черной дыры заключается в том, что если бы у нас было некое обычное классическое тело, состоящее из множества крошечных молекул или других элементарных локализованных составляющих, то число степеней свободы, потенциально доступных такому телу, было бы пропорционально объему данного тела. Поэтому можно было бы ожидать, что если тело находится в максимально разогретом состоянии, то есть обладает максимальной энтропией, данная энтропия будет величиной, пропорциональной объему, а не площади поверхности. Однако из формулы Бекенштейна — Хокинга следует, что все, что происходит внутри черной дыры, оставляет какой-то след на ее поверхности, и этот след в каком-то смысле эквивалентен происходящим внутри черной дыры процессам. Итак, следуя этому аргументу, мы имеем дело с примером голографического принципа в случае с информацией о степенях свободы внутри черной дыры; эта информация каким-то образом кодируется на границе («горизонте») черной дыры.

Такой обобщенный аргумент приводится в одной ранней работе по теории струн [Strominger and Vafa, 1996], где предпринимались попытки дать для формулы Бекенштейна — Хокинга обоснование в стиле Больцмана, подсчитав число степеней свободы струны, окруженной некоторой поверхностью. При этом сначала задавалось небольшое значение гравитационной постоянной, так чтобы поверхность не являлась границей черной дыры, а затем постоянная увеличивалась, пока поверхность не превращалась в горизонт событий. На тот момент теоретики-струнники считали этот результат большим прогрессом в понимании энтропии черных дыр, поскольку ранее никогда не удавалось проследить прямую связь между формулами Больцмана и Бекенштейна — Хокинга. Однако против такого аргумента выдвигалось множество возражений (аргумент действительно в некоторых отношениях был ограниченным и нереалистичным). Конкурирующий подход предложили сторонники применения петлевых переменных в квантовой гравитации [Ashtekar et al., 1998, 2000]. Этот подход, в свою очередь, также оказался сопряжен с собственными (казалось, менее существенными) трудностями. На мой взгляд, честнее было бы сказать, что до сих пор нет убедительной и однозначной процедуры, которая позволяла бы вывести формулу Бекенштейна — Хокинга для черной дыры из общего больцмановского определения энтропии. Тем не менее аргументы в пользу корректности формулы энтропии черной дыры хорошо разработаны другими способами и убедительны, поэтому на самом деле не требуют *прямого* обоснования по Больцману.

По моему личному мнению, неоправданно сопрягать внутреннюю область черной дыры с «объемлющим пространством» и считать, что в недрах черной дыры постоянно существуют «степени свободы» (см. раздел 3.5). Такое представление не согласуется с причинной обусловленностью внутри черной дыры. В черной дыре присутствует сингулярность, которая, как считается, способна уничтожать информацию, поэтому существование баланса, поиском которого заняты теоретики-струнники, на мой взгляд, является заблуждением. Аргументы из подхода с использованием петлевых переменных, по-моему, гораздо лучше обоснованы, чем ранние аргументы из теории струн, но пока не удается убедительно согласовать их математически с формулой для энтропии черной дыры.

Теперь давайте обратимся к той версии голографического принципа, которая связана с AdS/CFT-соответствием. На данный момент это по-прежнему не доказанная гипотеза, впервые предложенная Хуаном Малдасеной в 1997 году [Maldacena, 1998] и серьезно поддержанная Эдвардом Виттеном [Witten, 1998], а неустановленный математический принцип. Однако считается, что есть немало математических доказательств в пользу реального точного математического соответствия, связывающего две очень разные с виду физические модели. Идея такова: можно продемонстрировать, что некая сравнительно понятная теория

(в данном случае — теория струн), определяемая в $(n + 1)$-мерном объемлющем регионе $\mathcal{D}$, на самом деле *эквивалентна* гораздо более понятной теории (здесь имеется в виду более традиционная КТП) на n-мерной *границе* $\partial\mathcal{D}$ этого региона. Хотя корни этой идеи произрастают из глубоких проблем физики черных дыр, термин *голографический* происходит от хорошо знакомой сегодня всем нам *голографии*. Поэтому принято утверждать, что кажущаяся противоречивость информации о размерностях имеет право на существование, поскольку некоторые феномены такого общего характера уже известны на голограммах, ведь голографическая информация фактически представляет собой трехмерное изображение, закодированное на двумерной плоскости. Так и появились термины *голографическая гипотеза* и *голографический принцип*. Однако на самом деле реальная голограмма не иллюстрирует такой принцип, поскольку получаемый на голограмме трехмерный эффект зачастую является стереографическим, при котором два плоских изображения (воспринимаемых двумя глазами) создают впечатление глубины. Соответственно, они возникают из функциональной свободы $\infty^{2\infty^2}$, а не ∞^{∞^3}. Тем не менее таким образом получается хорошее приближение кодирования трех измерений, а при некоторой аккуратности и изобретательности такое приближение можно улучшить, создав не только объемное изображение, но и иллюзию движения. Дополнительная информация фактически скрыта в высокочастотных данных, которые глаз напрямую не воспринимает ['t Hooft, 1993; Susskind, 1994].

В конкретной версии этого принципа, именуемой *AdS/CFT-соответствием*, область $\mathcal{D}$ является пятимерным пространством-временем, которое именуется *анти-де-ситтеровской космологией* $\mathcal{A}^5$. В разделах 3.1, 3.7 и 3.9 мы увидим, как эта модель соотносится с обширным классом моделей, обычно называемых моделями FLRW, причем предполагается, что некоторые из них могут весьма точно соответствовать реальной структуре нашей четырехмерной Вселенной. Более того, де-ситтеровское пространство — это модель, которая, согласно теории и наблюдениям, должна хорошо аппроксимировать далекое будущее нашей Вселенной (см. разделы 3.1, 3.7 и 4.3). С другой стороны, четырехмерное *анти*-де-ситтеровское пространство $\mathcal{A}^4$ не является правдоподобной моделью реальной Вселенной, поскольку космологическая постоянная Λ в ней имеет знак, противоположный тому, который наблюдается в реальности (см. разделы 1.1, 3.1 и 3.6). По-видимому, этот эмпирический факт не мешает теоретикам-струнникам истово верить в полезность $\mathcal{A}^5$ для анализа природы нашей Вселенной.

Как я подчеркивал в предисловии к этой книге, физические модели зачастую исследуются в поисках подсказок, которые помогли бы расширить наши общие представления, и при этом такие модели совсем не обязательно должны быть реалистичными с физической точки зрения. Правда, именно в данном случае

существовала отнюдь не иллюзорная надежда на то, что значение Λ действительно может оказаться отрицательным. Хуан Малдасена сначала выдвинул в 1997 году свою гипотезу AdS/CFT, и вскоре после этого были произведены наблюдения [Perlmutter et al., 1998; Riess et al., 1998], убедительно доказывавшие положительность Λ, тогда как Малдасена считал, что Λ должна быть отрицательной. Еще даже в 2003 году, когда мне довелось беседовать на эту тему с Эдвардом Виттеном, оставалась надежда на то, что наблюдения допускают отрицательность Λ.

Согласно гипотезе AdS/CFT, адекватная теория струн в $\mathcal{A}^5$ в определенном смысле полностью эквивалентна более традиционной калибровочной теории (см. разделы 1.3 и 1.8) на четырехмерной конформной границе $\partial\mathcal{A}^5$ пространства $\mathcal{A}^5$. Однако, как отмечалось ранее (раздел 1.9), современные идеи теории струн требуют, чтобы пространственно-временное многообразие было десятимерным, а не пятимерным, как $\mathcal{A}^5$. Вот как решается эта проблема: считается, что теория струн должна применяться не к пятимерному пространству $\mathcal{A}^5$, а к десятимерному пространственно-временному многообразию: $\mathcal{A}^5 \times \mathcal{S}^5$ (см. рис. A.25 в разделе A.7 или раздел 1.9, где объясняется смысл операции «×», где $\mathcal{S}^5$ — это пятимерная сфера с космологическим радиусом; рис. 1.37). В данном случае рассматривается теория струн типа IIB, но здесь я не буду подробно останавливаться на различиях между разными типами теории струн.

Важно, что, поскольку $\mathcal{S}^5$ имеет космологический размер (и квантовые соображения в данном случае неактуальны), она должна обладать функциональной свободой, доступной в рамках «множителя» $\mathcal{S}^5$, и такая свобода определенно

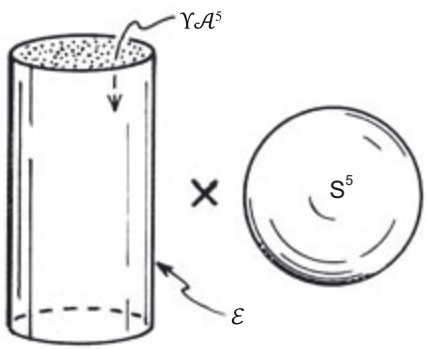


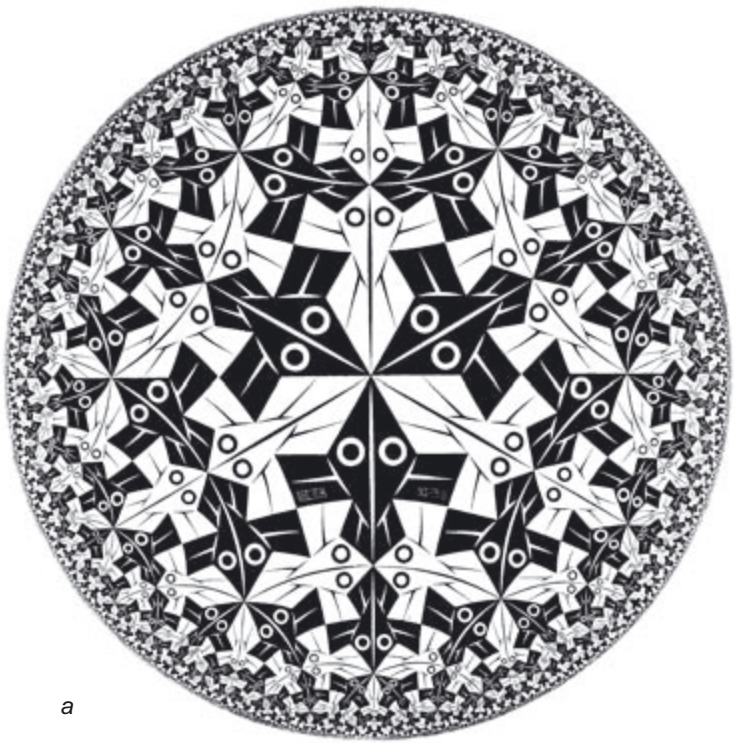
Рис. 1.37. Гипотеза AdS/CFT относится к лоренцеву 10-многообразию, представляющему собой произведение пространств $\mathcal{A}^5 \times \mathcal{S}^5$, где $\mathcal{A}^5$ — это анти-де-ситтеровское пятимерное пространство, а $\mathcal{S}^5$ — пространственноподобная пятимерная сфера. Здесь $\Upsilon\mathcal{A}^5$ — это «развернутая» версия $\mathcal{A}^5$, а $\mathcal{E}$ (статичная эйнштейновская Вселенная) — развернутая версия компактифицированного пространства Минковского

полностью поглощает любую динамическую свободу, доступную в $\mathcal{S}^5$, *если* такая динамика будет соответствовать традиционной динамике трехмерного пространства, предлагаемой AdS/CFT для границы $\partial\mathcal{A}^5$. Не требуется апеллировать к аргументам из раздела 1.10, связанным с возможными квантовыми препятствиями, способными помешать возбуждению такой избыточной функциональной свободы. В $\mathcal{S}^5$ нет никаких предпосылок для подавления этих огромных степеней свободы, и это красноречиво свидетельствует о том, что модель AdS/CFT не отражает структуру нашей Вселенной каким-либо непосредственным образом.

В представлении AdS/CFT $\mathcal{S}^5$ просто переносится на $\partial\mathcal{A}^5$ — конформную границу $\mathcal{A}^5$, и таким образом получается своеобразная граница $\partial\mathcal{A}^5 \times \mathcal{S}^5$ для $\mathcal{A}^5 \times \mathcal{S}^5$, но она еще далеко не является *конформной* границей $\mathcal{A}^5 \times \mathcal{S}^5$. Чтобы это объяснить, я должен дать вам представление о том, что же такое конформная граница, и для этого отсылаю читателя к рис. 1.38 *а*, где вся гиперболическая плоскость (вернемся к этому в разделе 3.5) изображена конформно точным образом. В данном случае конформная граница — это внешний круг. На рисунке представлена знаменитая красивая ксилография голландского художника М. К. Эшера — довольно точное представление гиперболической плоскости (изначально его предложил Эудженио Бельтрами в 1868 году, но фигура стала известна под названием *диск Пуанкаре*). Прямые линии в этой геометрии представлены в виде дуг окружности, встречающихся с ограничивающей окружностью под прямым углом (рис. 1.38 *б*). В этой *неевклидовой* планиметрии множество линий («параллелей»), проведенных через точку P, не пересекаются с линией *a*, а сумма углов треугольника α, β, γ меньше π (меньше 180°). Рисунки 1.38 *а* и *б* можно представить и с дополнительными высшими измерениями — например, конформно изобразить трехмерное гиперболическое пространство как внутреннюю часть обычной сферы S^2. «Конформный», в принципе, означает, что все очень мелкие фигуры, например рыбьи плавники, в этом изображении представлены очень точно: чем меньше фигура, тем точнее представление, хотя различные экземпляры таких же мелких фигур могут различаться по размеру (а рыбьи глаза остаются круглыми вплоть до самого края). Некоторые мощные идеи конформной геометрии вскользь упоминаются в разделе A.10 и — в контексте пространства-времени — в конце раздела 1.7 и в начале раздела 1.8. (Мы вернемся к конформным границам в разделах 3.5 и 4.3.) Оказывается, что в случае $\mathcal{A}^5$ его конформная граница $\partial\mathcal{A}^5$, в сущности, может трактоваться как конформная копия обычного пространства-времени Минковского $\mathbb{M}$ (см. разделы 1.7 и 1.11), пусть и «компактифицированная» определенным способом, который мы вскоре обсудим. Идея AdS/CFT заключается в том, что в определенном варианте теории струн для пространства-времени $\mathcal{A}^5$ тайны математической природы этой теории

можно разгадать при помощи именно этой гипотезы, поскольку калибровочные теории в пространстве Минковского очень хорошо разработаны.

Также существует проблема с «множителем» $\times \mathcal{S}^5$, который и должен привносить основную часть функциональной свободы. В контексте общих идей голографии $\mathcal{S}^5$ в основном игнорируется. Как упоминалось ранее, $\partial\mathcal{A}^5 \times \mathcal{S}^5$ определенно не является конформной границей $\mathcal{A}^5 \times \mathcal{S}^5$, поскольку «расплющивание» бесконечных областей $\mathcal{A}^5$ для «достижения» $\partial\mathcal{A}^5$ неприменимо к $\mathcal{S}^5$, тогда как при конформном расплющивании это явление должно в равной мере затрагивать все измерения. Для обработки информации в $\mathcal{S}^5$ остается только прибегнуть к *анализу мод*, то есть полностью закодировать ее в виде последовательности чисел (в формализме AdS/CFT такая последовательность именуется *башней*). Как будет указано в конце раздела A.11, это хороший способ затушевать проблемы функциональной свободы!



a

Рис. 1.38 a. На гравюре «Предел круга I» М. К. Эшера используется конформное представление гиперболической плоскости, предложенное Бельтрами. Бесконечность этой плоскости превращается в круговую границу

Те из читателей, кто еще не потерял нить рассуждений, наверняка уже слышат тревожный звонок по поводу гипотезы AdS/CFT: ведь если теория, определяемая на четырехмерной границе $\partial\mathcal{A}^5$, выглядит как самая обычная четырехмерная теория поля, то она должна основываться на величинах с функциональной свободой $\infty^{A\infty^3}$ (для некоторого положительного целого числа A), тогда как в пятимерных недрах $\mathcal{A}^5$ следовало бы ожидать несравнимо бóльшей функциональной свободы $\infty^{B\infty^4}$ (для любого B), если бы можно было считать, что недра также описывает обычная теория поля (см. разделы A.2 и A.8). Вероятно, в таком случае было бы очень сложно поверить в предполагаемую эквивалентность между двумя этими теориями. Однако здесь есть несколько осложняющих проблем, которые также необходимо учитывать.

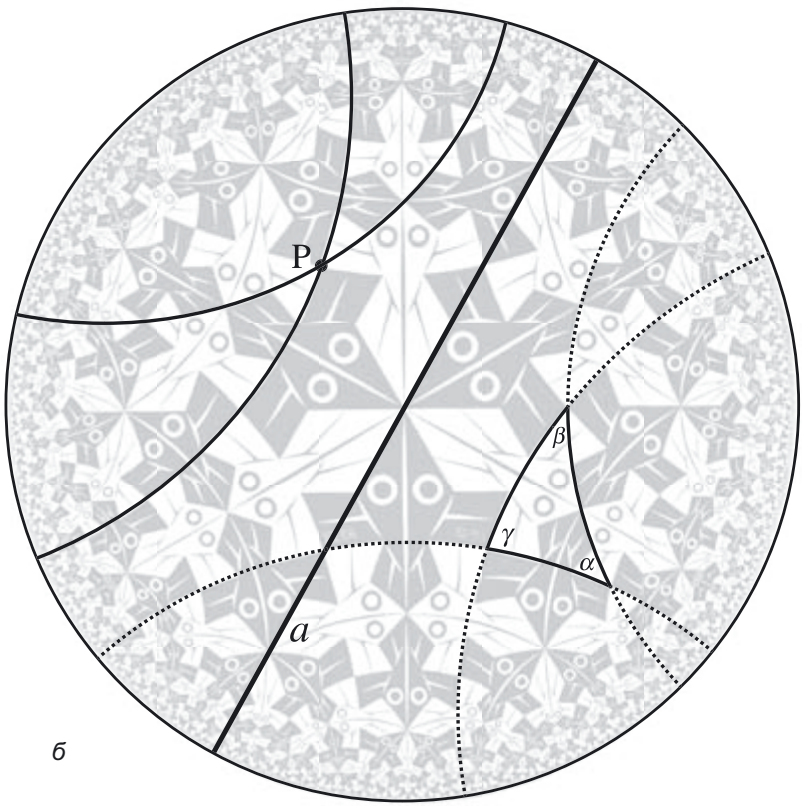


Рис. 1.38 б. Прямые линии в этой геометрии представлены в виде дуг окружности, пересекающихся с ограничивающей окружностью под прямым углом. Через P проходит множество «параллелей», не пересекающихся с линией a, а сумма углов треугольника α, β, γ меньше π (меньше 180°)

Прежде всего, необходимо учитывать, что теория для недр позиционируется как теория струн, а не как обычная КТП. В ответ на такое условие сразу хочется довольно обоснованно указать, что в теории струн функциональная свобода должна быть гораздо *больше*, чем в теории, описывающей точечные частицы, поскольку если говорить о *классической* функциональной свободе, у струны будет гораздо больше петель, чем у отдельных точек. Однако это приводит нас к весьма ошибочной оценке масштабов функциональной свободы в теории струн. Лучше считать, что теория струн — просто еще один способ представить обычную физику (в конце концов, это одна из целей теории струн), поэтому она подводит нас все к тому же общему представлению о функциональной свободе (в определенном смысле к классическому пределу), какой мы получили бы и в обычной классической теории поля в $(n + 1)$-мерном пространстве-времени, то есть к рассмотренной выше общей форме $\infty^{B\infty^4}$ (здесь я не учитываю колоссальную свободу в $\mathcal{S}^5$). Форма $\infty^{B\infty^4}$ определенно служит ответом на вопрос о функциональной свободе классической эйнштейновской гравитации в объемлющем пространстве, следовательно, именно эту форму мы предположительно и должны были бы получить на основе адекватного классического предела для теории струн в объемлющем пространстве.

Однако с классическим пределом связана еще одна проблема. Более подробно она будет рассмотрена в совершенно другом ракурсе в разделе 2.13 (и в разделе 4.2). Возможно, между этими случаями есть какая-то связь, но я не пытаюсь здесь ее прослеживать. Тем не менее возможно, что в случае с объемлющим пространством и границей мы просто рассматриваем разные пределы, и это привносит еще один осложняющий фактор. Он может быть связан с загадкой голографического принципа для ситуации, когда функциональная свобода на границе, по-видимому, может иметь лишь форму $\infty^{A\infty^3}$, что неизмеримо меньше $\infty^{B\infty^4}$ (второе значение мы рассчитываем встретить в объемлющем пространстве). Разумеется, нельзя отвергать такую возможность, что гипотеза AdS/CFT попросту неверна, несмотря на, казалось бы, уже обнаруженные сильные частные доказательства в пользу близкого родства между теориями объемлющего пространства и границы. Например, может оказаться, что любое решение граничных уравнений действительно проистекает из решения уравнений для объемлющего пространства, но найдется целая масса «объемлющих» решений, у которых не будет соответствия среди «граничных». Подобные ситуации будут возникать, если мы рассмотрим некую пространственноподобную трехмерную сферу S^3 на $\partial\mathcal{A}^5$, охватывающую пространственноподобный четырехмерный шар D^4 в $\partial\mathcal{A}^5$, и посмотрим на уравнения, представляющие собой трех- и четырехмерный вариант уравнения Лапласа. Каждое решение для S^3 получается из уникального решения для D^4 (см. раздел A.11), но многие решения для D^4 *не являются* решениями в S^3 $(= \partial D^4)$. На гораздо более детализированном уровне

обнаруживаем, что некоторые решения актуальных уравнений, называемые *состояниями BPS* (Богомольного — Прасада — Соммерфилда), характеризующиеся определенными симметричными и суперсимметричными свойствами, обнаруживают удивительно точное соответствие между BPS-состояниями поверхности и внутренней части сферы. Однако можно спросить: в какой степени эти частные случаи иллюстрируют общую ситуацию, в которой целиком задействована функциональная свобода?

Еще один момент, заслуживающий рассмотрения (как отмечено в разделе A.8), таков: наши соображения о функциональной свободе, в сущности, *локальны*, поэтому глобально вышеупомянутых проблем, связанных с классической версией AdS/CFT, может и не оказаться. Глобальные ограничения порой могут радикально сокращать число решений для классических уравнений теории поля. Чтобы прояснить этот момент для AdS/CFT-соответствия, нужно разобраться с одним вопросом, который освещается в литературе несколько путано, а именно с тем, что означает термин «глобально» в случае такого соответствия. На самом деле речь идет о двух разных вариантах геометрии. В обоих случаях имеются совершенно валидные (хотя и несомненно запутанные) примеры конформной геометрии, где *конформность* в пространственно-временном контексте является свойством *семейства нулевых (световых) конусов* (см. разделы 1.7 и 1.8). Я буду различать две эти геометрии как *свернутую* и *развернутую* версии $\mathcal{A}^5$, а также его конформной границы $\partial\mathcal{A}^5$. Здесь символы $\mathcal{A}^5$ и $\partial\mathcal{A}^5$ будут относиться к *свернутым* версиям, а $\Upsilon\mathcal{A}^5$ и $\Upsilon\partial\mathcal{A}^5$ — к *развернутым* версиям. Технически $\Upsilon\partial\mathcal{A}^5$ является *универсальным накрытием* $\mathcal{A}^5$. Концепции, рассматриваемые здесь, объяснены (надеюсь, что доходчиво) на рис. 1.39. Свернутую версию $\mathcal{A}^5$ проще

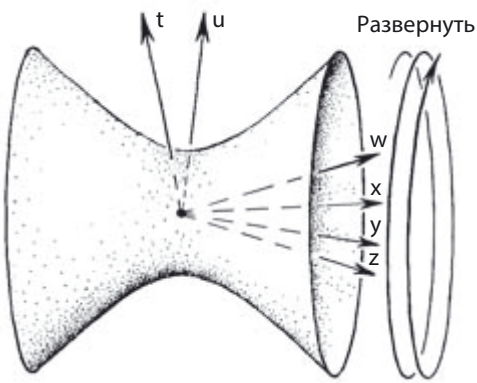


Рис. 1.39. Анти-де-ситтеровское пространство $\mathcal{A}^5$ содержит замкнутые времениподобные кривые. Их можно удалить, «развернув» $\mathcal{A}^5$, то есть повернув его в плоскости (t, u) и тем самым сформировав универсальное накрытие $\Upsilon\mathcal{A}^5$

всего получить при помощи соответствующих алгебраических уравнений;[1] она имеет топологию $S^1 \times \mathbb{R}^4$. Развернутая версия имеет топологию $\mathbb{R}^5$ ($= \mathbb{R} \times \mathbb{R}^4$), где каждый круг S^1 в $S^1 \times \mathbb{R}^4$ развернут (для этого его обходят неограниченное число раз) в прямую ($\mathbb{R}$). Физическая причина, по которой мы хотим сделать такое «развертывание», заключается в том, что эти круги являются *замкнутыми времениподобными мировыми линиями*, которые обычно считаются неприемлемыми в любой модели пространства-времени, позиционируемой как реалистичная (дело в парадоксальных действиях наблюдателей, обладающих такими мировыми линиями, и потенциальной возможности того, что наблюдатель силой своей свободной воли смог бы изменить события, которые уже считаются свершившимися в прошлом). В таком случае после развертывания повышается вероятность того, что такая модель окажется реалистичной.

Конформная граница $\mathcal{A}^5$ представляет собой так называемое *компактифицированное четырехмерное пространство Минковского* $\mathbb{M}^{\#}$. Эту границу можно (конформно) трактовать как обычное четырехмерное пространство Минковского $\mathbb{M}$ из специальной теории относительности (см. раздел 1.7; рис. 1.23), к которому прилегает собственная конформная граница $\mathcal{I}$ (см. рис. 1.40). Однако

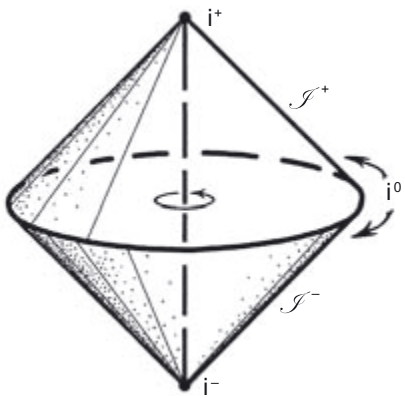


Рис. 1.40. Пространство Минковского показано вместе с конформной границей, состоящей из двух нулевых трехмерных гиперповерхностей, $\mathcal{I}^+$ (будущая нулевая бесконечность) и $\mathcal{I}^-$ (прошлая нулевая бесконечность), и трех точек, i^+ (будущая времениподобная бесконечность), i^0 (пространственноподобная бесконечность) и i^- (прошлая времениподобная бесконечность)

[1] $\mathcal{A}^5$ — это 5-квадрика $t^2 + u^2 - w^2 - x^2 - y^2 - z^2 = R^2$ в шестимерном пространстве $\mathbb{R}^6$ вещественных координат (t, u, w, x, y, z) с метрикой $ds^2 = dt^2 + du^2 - dw^2 - dx^2 - dy^2 - dz^2$. Развернутые версии $\Upsilon\mathcal{A}^5$ и $\Upsilon\mathbb{M}^{\#}$ — это *универсальные накрытия* $\mathcal{A}^5$ и $\mathbb{M}^{\#}$ соответственно; см. [Alexakis, 2012].

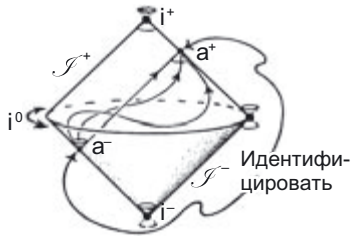


Рис. 1.41. Для того чтобы построить компактифицированное четырехмерное пространство Минковского с топологией $S^1 \times S^3$, мы приравниваем $\mathcal{I}^+$ к $\mathcal{I}^-$ точка за точкой, как показано на рисунке, чтобы a^- на $\mathcal{I}^-$ идентифицировалось с a^+ на $\mathcal{I}^+$, где любая нулевая геодезическая в $\mathbb{M}$, имеющая в прошлом конечную точку a^- на $\mathcal{I}^-$, получает конечную точку a^+ на $\mathcal{I}^+$ в будущем. Дополнительно потребуется идентифицировать все три точки i^-, i^0 и i^+

мы соответствующим образом «свертываем» его, приравнивая «будущую» часть его конформной границы $\mathcal{I}^+$ к «прошлой» части $\mathcal{I}^-$, приравнивая для этого бесконечный обращенный в будущее отрезок любого светового луча (a^+) (нулевую геодезическую) в $\mathbb{M}$ к бесконечному отрезку этого луча, обращенному в прошлое (a^-) (рис. 1.41). Развернутое граничное пространство $\Upsilon\mathbb{M}^\#$ (универсальное накрытие $\Upsilon\mathbb{M}^\#$) оказывается конформно эквивалентным статической эйнштейновской Вселенной $\mathcal{E}$ (рис. 1.42; см. также раздел 3.5, рис. 3.23), то есть пространственной трехмерной сферой, не изменяющейся со временем: $\mathbb{R} \times S^3$. Часть этого пространства, конформная пространству Минковского, с прилегающей к ней конформной границей $\mathcal{I}$ (для случая с двумя измерениями) показана на рис. 1.43 (см. также рис. 3.23).

Развернутые версии этих пространств, по-видимому, не налагают значительных глобальных ограничений на классические решения уравнений поля. В принципе, нас должны волновать такие вещи, как обнуление общего заряда, в случае с максвелловскими уравнениями для источников, возникающих при компактификации пространственных направлений (так мы получаем S^3 для эйнштейновской Вселенной, о чем шла речь выше). Не вижу, по каким причинам могли бы возникать еще какие-то топологические ограничения для классических полей на развернутых $\Upsilon\mathcal{A}^5$ и $\Upsilon\mathbb{M}^\#$, поскольку не возникает никаких дополнительных ограничений для эволюции во времени. Но компактификация темпоральной составляющей, которая обеспечивается путем свертывания открытого временного направления в $\Upsilon\mathcal{A}^5$ и $\Upsilon\mathbb{M}^\#$, для получения $\mathcal{A}^5$ и $\mathbb{M}^\#$ определенно приведет к радикальному сокращению числа классических решений, поскольку после процедуры свертывания сохранятся лишь те решения, периодичность которых согласуется с этой компактификацией [Jackiw and Rebbi, 1976].

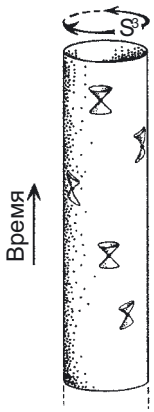


Рис. 1.42. Статическая эйнштейновская модель Вселенной $\mathcal{E}$ — это пространственная трехмерная сфера, не изменяющаяся со временем; топологически представляет собой $\mathbb{R} \times S^3$

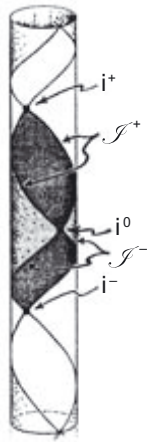


Рис. 1.43. Это изображение (пусть и всего лишь двумерное) указывает, как можно трактовать пространство Минковского с его конформной границей в качестве замкнутого сегмента эйнштейновской статической модели $\mathcal{E}$. Этот пример поясняет, каким образом i^0 может представлять собой отдельную точку

Следовательно, я полагаю, что именно *развернутые* варианты $\Upsilon\mathcal{A}^5$ и $\Upsilon\partial\mathcal{A}^5$ — это пространства, релевантные для гипотезы AdS/CFT. Как же в таком случае избавиться от явной несогласованности значений функциональной свободы в двух теориях? Весьма возможно, что ответ может заключаться в одном свойстве этого соответствия, которое я пока не затрагивал. Дело в том, что теория граничного поля Янга — Миллса на самом деле не является стандартной теорией поля (даже если не учитывать, что в ней четыре генератора суперсимметрии), поскольку ее группа калибровочной симметрии должна рассматриваться в тех пределах, где ее размерность стремится к бесконечности. С точки зрения функциональной свободы ситуация немного напоминает рассмотрение «башен» гармоник, которые должны определять, что происходит в S^5-пространствах. Дополнительная функциональная свобода может оставаться «скрытой» в бесконечности этих гармоник. Аналогично тот факт, что для работоспособности AdS/CFT-соответствия размер калибровочной группы доводится до *бесконечности*, вполне позволяет решить явный конфликт с функциональной свободой.

Итак, кажется несомненным, что AdS/CFT-соответствие открывает огромное поле для исследования, связанное со многими активно развивающимися областями теоретической физики. На этом поле удается проследить неожиданные

связи между столь несхожими дисциплинами, как физика конденсированных сред, физика черных дыр и физика элементарных частиц. С другой стороны, существует странный контраст между такой замечательной многоплановостью и изобилием идей — и в то же время нереалистичностью той картины мира, которая при этом получается. Чтобы избавиться от получаемого неверного знака космологической константы, требуется четыре генератора суперсимметрии, в то время как мы пока что не наблюдали ни одного. Нам нужна такая группа калибровочной симметрии, которая будет распространяться на бесконечное множество параметров, а не на три, требуемые для физики элементарных частиц, а у объемлющего пространства при этом будет слишком много измерений! Будет особенно интересно посмотреть, к чему нас все это приведет.

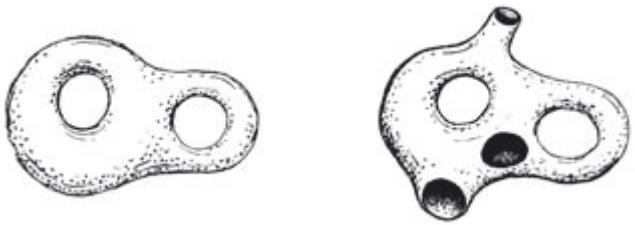


Рис. 1.44. Изображения римановых поверхностей: с ручками (*а*) и с ручками и отверстиями (*б*). (Примечание: в литературе те сущности, которые я называю «ручками», иногда ошибочно именуются «отверстиями»)

1.16. Миры на бранах и ландшафт

Теперь давайте обсудим проблему миров на бранах. В разделе 1.13 я упомянул, что сущности, известные под названием «*p*-браны» — аналоги струн, обладающие высшими измерениями, — наряду со струнами нужны для описания различных дуальностей, характеризующих М-теорию. Сами струны (1-браны) могут существовать в двух принципиально разных формах: *замкнутые*, которые можно описывать как обычные компактные римановы поверхности (см. рис. 1.44 *а* и раздел А.10), и *открытые*, которым соответствуют римановы поверхности с отверстиями в них (рис. 1.44 *б*). Кроме того, предположительно существуют структуры под названием *D-браны*, играющие в теории струн иную роль. D-браны считаются классическими структурами в (надразмерном) пространстве-времени и предположительно возникают из огромных скоплений простейших струнных и *p*-бранных составляющих, но трактуются как классические решения уравнений супергравитации с соблюдением требований симметрии и суперсимметрии.

Считается, что D-браны играют важную роль в качестве тех точек, где должны находиться «концы» открытых струн (то есть отверстия) (рис. 1.45).

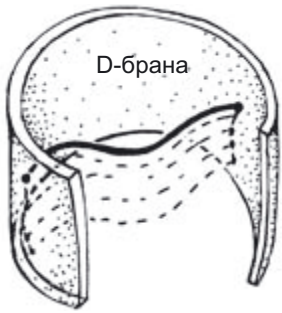


Рис. 1.45. Схема D-браны. Это (классическая) область, где будут находиться концы струн («отверстия»)

Концепция *мира на бранах* представляет собой серьезное (что, впрочем, нечасто признается) отступление от исходной трактовки теории струн, описанной в разделе 1.6. В такой трактовке пространство-время с высшими измерениями (локально) считается произведением пространств $\mathcal{M} \times \mathcal{X}$, причем непосредственно воспринимаемое нами пространство-время — это четырехмерное пространство $\mathcal{M}$, а шестимерное пространство $\mathcal{X}$ содержит крошечные невидимые дополнительные измерения. Согласно этой исходной точке зрения, наблюдаемое четырехмерное пространство-время является примером известного в математике *факторного пространства*, где можно получить пространство $\mathcal{M}$, если отобразить каждый экземпляр $\mathcal{X}$ из $\mathcal{M} \times \mathcal{X}$ в точку:

$$\mathcal{M} \times \mathcal{X} \to \mathcal{M}$$

(см. рис. 1.32 *а* в разделе 1.10). Но с другой стороны, согласно концепции мира на бранах, ситуация складывается с точностью до наоборот, поскольку наблюдаемая четырехмерная Вселенная трактуется как *подпространство* окружающего десятимерного пространства-времени $\mathcal{S}$ и идентифицируется как определенная четырехмерная D-брана $\mathcal{M}$ *внутри* него:

$$\mathcal{M} \to \mathcal{S}$$

(см. рис. 1.32 *б*). Мне эта идея представляется очень странной, поскольку большая часть пространства-времени $\mathcal{S}$ видится совершенно нерелевантной для нашего опыта. Тем не менее это следует считать определенным прогрессом, поскольку функциональная свобода в таком случае должна быть гораздо менее избыточной, чем ранее. Есть, однако, и недостаток: нормальное детерминирован-

ное распространение полей в будущее, к которому мы привыкли в традиционной физике, должно полностью утратиться, поскольку информация будет постоянно утекать из подпространства $\mathcal{M}$ в окружающее надразмерное пространство-время. Следует отметить, что подобное полностью противоречит нормальной детерминированной эволюции классических полей в пространстве-времени, которое мы ощущаем. В картине мира на бранах функциональная свобода классических полей, которую мы должны непосредственно наблюдать, теперь имела бы вид $\infty^{B\infty^4}$, а не значительно меньшее значение $\infty^{A\infty^3}$, которое мы наблюдаем на практике. Это все равно слишком много. На самом деле воспринимать такую картину всерьез мне еще сложнее, чем даже исходную из раздела 1.6.

Наконец, переходим к проблеме *ландшафта*, а также *болот* — этот вопрос в отличие от других, упомянутых мною ранее, по-видимому, действительно волнует некоторых теоретиков-струнников! В разделе 1.10 я затрагивал нулевые моды пространств Калаби — Яу, для возбуждения которых не требуется какой-либо энергии. Эти моды возбуждения не связаны с избыточной функциональной свободой, присущей дополнительным пространственным измерениям, но влекут конечномерные изменения параметров, именуемых *модулями*, которые характеризуют *формы* пространств Калаби — Яу, откуда берутся дополнительные пространственные измерения. Деформация этих модулей дает нам огромное число альтернативных теорий струн, задаваемых так называемыми *альтернативными вакуумами*.

Феномен *вакуума* в квантовой теории поля очень важен, и ранее в этой книге я его не обсуждал. Ситуация такова, что для описания КТП требуются *две* составляющие. Одна из них — *алгебра* операторов этой теории, например операторов рождения и уничтожения, о которых шла речь в разделе 1.14. Другая — это выбор вакуума, на который эти операторы в конечном итоге будут действовать, создавая состояния с участием все более и более многочисленных частиц (за что отвечают операторы рождения частиц). В КТП часто обнаруживается, что одна и та же алгебра операторов вполне может допускать различные «неэквивалентные» вакуумы, поэтому если начать с конкретной разновидности вакуума, то не удастся перейти к другой средствами все той же исходной алгебры. Это означает, что теория, выстраиваемая на базе одного вакуума, описывает совершенно иную Вселенную, чем если бы изначально был выбран другой неэквивалентный вакуум, а квантовые суперпозиции между состояниями из разных КТП не допускаются. Этот факт также сыграет очень важную роль в разделах 3.9, 3.11 и 4.2. В теории струн мы таким образом получаем неимоверное множество неэквивалентных струнных теорий (или М-теорий).

Это разительно отличается от исходного положения теории струн, которая претендовала на роль *единой* физической теории. Мы помним, какой успех прочили

М-теории: объединение пяти, казалось бы, совершенно разных возможных типов теории струн. Мне кажется, этот явный успех ныне совершенно потерялся на фоне вышеупомянутого умножения различных теорий струн (или М-теорий). Точное их число на данный момент неизвестно, но приводится оценка ~10^{500} [Douglas, 2003; Ashok and Douglas, 2004]. Они возникают из-за существования колоссального множества неэквивалентных вакуумов, которое кажется вполне допустимым! В ходе попыток справиться с этой проблемой возникла точка зрения, согласно которой все различные вселенные сосуществуют, образуя «ландшафт» таких возможностей. Среди этого необозримого множества, казалось бы, нереализованных математических возможностей есть и много таких, которые лишь *кажутся* возможностями, но с математической точки зрения получаются *внутренне противоречивыми*. Они и образуют так называемое *болото*. Теперь, по-видимому, формируется следующая идея: если мы стремимся объяснить явный «подбор» значений, заданных природой для различных модулей, наблюдаемых в нашей Вселенной, то должны выстраивать следующую аргументацию: можно очутиться лишь в такой Вселенной, значения модулей в которой порождают константы, совместимые с химией, физикой и космологией, необходимыми для развития разумной жизни. Это пример так называемого *антропного принципа*, о котором мы поговорим в разделе 3.10. Эта местность, в которую нас завела столь грандиозная теория, мне кажется очень унылой. Антропный принцип должен помочь понять некоторые, пожалуй, явно случайные совпадения между некоторыми фундаментальными константами природы, но в целом его объяснительная сила весьма ограниченна. Я подробнее обращусь к этой проблеме в разделе 3.10.

Какой вывод можно сделать из этих последних разделов относительно грандиозных амбиций тех исходных, подлинно привлекательных струнно-теоретических идей? Гипотеза AdS/CFT действительно продемонстрировала много интригующих и подчас неожиданных соответствий между различными областями, вызывающих у физиков неподдельный интерес (например, связи между черными дырами и физикой твердого тела; см. также раздел 3.3 и [Cubrovic et al., 2009]). Такие соответствия могут быть и в самом деле занимательными, особенно в математической сфере, но в целом ситуация значительно отошла от первых ожиданий, которые связывались с теорией струн, когда со всей серьезностью предполагалось, что эта тема приблизит нас к гораздо более полному пониманию глубоких секретов природы. Что же касается идеи миров на бранах, то мне кажется, что от нее веет неким отчаянием, когда наше существование ютится на крошечном утесе с малым числом измерений и у нас нет никакой надежды заглянуть в безбрежные дали надразмерных непостижимых процессов. Ландшафт — еще хуже, поскольку на нем у нас нет никаких разумных оснований рассчитывать даже на то, что мы могли бы найти координаты такого относительно безопасного утеса, на котором мы только и могли бы существовать!

ГЛАВА 2

Вера

2.1. Квантовое откровение

Согласно *Краткому оксфордскому словарю*, *вера* — это убеждение, основанное на авторитете. Мы привыкли, что авторитет в значительной степени влияет на наше мышление, будь то авторитет родителей в нашем детстве, либо авторитет учителей в школьные годы, либо авторитет представителей таких уважаемых профессий, как врач, юрист, ученый, телеведущий, либо авторитет представителей правительства или международных организаций, либо, наконец, иерархов религиозных институтов. Так или иначе, авторитет влияет на наше мнение, и информация, которую мы получаем от авторитета, зачастую превращается в убеждения, в которых мы всерьез уже никогда не сомневаемся. Действительно, зачастую мы даже не испытываем сомнений относительно достоверности большей части информации, получаемой из авторитетных источников. Более того, авторитеты часто влияют на наше поведение и на то, как мы позиционируем себя в обществе; в свою очередь, авторитет, которым можем обладать мы сами, повышает статус нашего личного мнения и позволяет нам влиять на убеждения окружающих.

Зачастую такое поведенческое влияние — это аспект культуры, и подчиняться ему требует наше воспитание, чтобы избежать ненужных трений. Однако возникает более серьезная проблема, если задаться вопросом: а что на самом деле *истинно*? Действительно, один из идеалов науки заключается в том, что *нельзя* просто принимать что-либо на веру и что наши убеждения должны, хотя бы время от времени, сверяться с окружающей реальностью. Разумеется, у нас наверняка не будет возможности или ресурсов, чтобы подвергнуть такой проверке значительное большинство наших убеждений. Тем не менее мы должны стараться по меньшей мере непредвзято смотреть на вещи. Вполне возможно, что порой в нашем распоряжении окажется лишь рассудок, проницательность, объективность и здравый смысл. Однако эти качества не стоит недооценивать.

Если им довериться, то логично будет предположить, что научные взгляды — это не результат хитроумно сплетенного заговора по навязыванию лжи. Например, сегодня мы располагаем массой волшебных устройств, в которых есть чуть-чуть волшебства, — это телевизоры, мобильные телефоны, планшеты iPad, навигаторы с системой глобального позиционирования (GPS), не говоря уже о реактивных самолетах, чудодейственных лекарствах и т. д., — все это может нас убедить, что в большинстве заявлений, основанных на научном понимании мира, есть нечто глубоко истинное. Эти заявления удалось сделать благодаря строгой проверке фактов научным методом. Соответственно, хотя культура естественных наук не породила новый авторитет, она является авторитетом, который, по меньшей мере в принципе, подвергается постоянному пересмотру. Следовательно, вера в научный авторитет, которую мы испытываем, не слепая, и мы всегда должны быть готовы к неожиданному изменению взглядов, выражаемых авторитетными учеными. Более того, нас не должно удивлять, что некоторые научные взгляды являются предметом серьезных противоречий.

Разумеется, слово *вера* чаще используется в отношении к *религиозным* доктринам. В этом контексте вполне могут приветствоваться дискуссии по базовым вопросам, а некоторые детали официальной доктрины могут с годами аккуратно корректироваться, чтобы соответствовать меняющимся обстоятельствам. Тем не менее обычно в религии существует корпус догматических положений, которые (как минимум в случае основных мировых религий) могут существовать на протяжении длительного времени, порядка тысячелетий. В каждом случае происхождение догматов, на которых зиждется вера, возводится к выдающейся личности (личностям), обладающей исключительной моральной чистотой, силой характера, мудростью и даром убеждения. Хотя можно ожидать, что в тумане времени интерпретация и детали исходных заветов могут немного измениться, никто не поспорит, что центральная идея дошла до нас практически неискаженной.

Все это совсем не напоминает тот путь, по которому развивались научные знания. Тем не менее ученым ничего не стоит поддаться самоуспокоению и воспринимать устоявшиеся научные формулировки как незыблемые. На самом деле мы не раз наблюдали различные значительные изменения в научных убеждениях, и при таких переменах хотя бы отчасти опровергались те мнения, которые ранее пользовались уверенной поддержкой. Однако такие изменения лишь со скрипом признавались теми, кто придерживался старых убеждений, и обычно им удавалось закрепиться только после появления новых очень веских эмпирических доказательств. Один из таких примеров — открытая Кеплером эллиптическая траектория движения планет; это открытие позволило отказаться от более ранних идей о системе круговых орбит. Эксперименты Фарадея и уравнения Максвелла стали первыми вестниками новых великих перемен в научной кар-

тине мира, касавшихся природы материи. Оказалось, что отдельные, дискретные частицы, описываемые теорией Ньютона, нужно дополнить непрерывными электромагнитными полями. Еще более поразительными были две великие революции XX века: появление теории относительности и квантовой механики. Я рассматривал некоторые примечательные идеи специальной и общей теории относительности в разделах 1.1, 1.2, и в особенности в разделе 1.7. Но даже эти революции, какими бы впечатляющими они ни казались, практически полностью блекнут перед ошеломляющими откровениями квантовой теории. Именно *квантовое откровение* и будет темой второй части книги.

В разделе 1.4 мы уже познакомились с самым странным свойством квантовой механики: одно из следствий принципа квантовой суперпозиции заключается в том, что частица может находиться в двух местах сразу! Это определенно уводит нас от удобных ньютоновских представлений об отдельных частицах, каждая из которых находится в строго определенном месте. Очевидно, что такое, казалось бы, безумное описание реальности, которое дает квантовая теория, не было бы всерьез воспринято уважаемыми учеными, если бы не множество эмпирических доказательств, подкрепляющих ее положения. Дело не только в этом — ведь как только привыкнешь к квантовому формализму и освоишь множество тонких математических процедур, которые с ним связаны, сами собой начинают появляться объяснения бесчисленных физических явлений, многие из которых прежде оставались абсолютной тайной.

Квантовая теория объясняет феномен химических связей, цвет и физические свойства металлов и других веществ, подробно описывает природу дискретного спектра частот света, излучаемого при нагревании конкретных элементов и их соединений (спектральные линии), стабильность атомов (где классическая теория предсказывает катастрофический коллапс с излучением всей энергии, когда электроны стремительно падают по спирали на атомное ядро); она также объясняет сверхпроводники, сверхтекучие жидкости, конденсаты Бозе — Эйнштейна. В биологии квантовая теория объясняет удивительную дискретность наследуемых признаков (открытую Грегором Менделем около 1860 года и в целом объясненную Шрёдингером в его совершившей научный прорыв работе «Что такое жизнь?», опубликованной в 1943 году [см. Schrödinger, 2012], еще до того как стала пристально изучаться молекула ДНК). В космологии квантовая теория объясняет микроволновое реликтовое излучение, пронизывающее всю нашу Вселенную (о нем мы подробно поговорим в разделах 3.4, 3.9 и 4.3), имеющее спектр абсолютно черного тела (см. раздел 2.2), который сформировался в результате квантового по своей сути процесса. Многие современные физические приборы критическим образом зависят от квантовых явлений, а для их конструирования требуется глубокое понимание основ квантовой механики.

Лазеры, CD- и DVD-плееры, ноутбуки — все эти устройства принципиальным образом зависят от квантовой составляющей, равно как и сверхпроводящие магниты, разгоняющие элементарные частицы практически до скорости света по 27-километровым туннелям Большого адронного коллайдера в Женеве. Список можно продолжать почти бесконечно. Итак, квантовую теорию действительно следует воспринимать всерьез и признать, что она дает убедительное описание физической реальности, значительно выходящее за рамки классических представлений, веками столь прочно коренившихся в научном мире до появления этой замечательной теории.

Объединив квантовую теорию поля со специальной теорией относительности, мы получаем квантовую теорию поля, играющую важнейшую роль, например, в современной физике частиц. Как вы помните из раздела 1.5, квантовая теория поля дает значение магнитного момента электрона с точностью до 10 или 11 значащих цифр после того, как будет выполнена процедура перенормировки, позволяющая справиться с расходимостями. Существуют и другие примеры, красноречиво свидетельствующие об исключительной точности, присущей квантовой теории поля, если правильно ее применять.

Обычно считается, что квантовая теория обладает большей глубиной, чем предшествовавшая ей классическая. Несмотря на то что, как правило, квантовая механика обычно используется для описания сравнительно мелких сущностей, например атомов и образующих их элементарных частиц, область ее применения не ограничивается такими элементарными составляющими материи. Очень большие совокупности электронов нередко отвечают за странное и подлинно квантово-механическое поведение сверхпроводников, а также за поведение атомов водорода (в количестве около 10^9 штук) в конденсатах Бозе — Эйнштейна [Greytalk et al., 2000]. Более того, эффекты квантовой запутанности сегодня воспроизведены на расстояниях более 143 км [Xiao et al., 2012], когда пары фотонов, разделенных такими промежутками, тем не менее могут считаться цельными квантовыми объектами. Существуют и такие измерения, которые позволяют определить диаметр далекой звезды; они основаны на том, что два фотона, излучаемые с противоположных сторон звезды, автоматически спутываются друг с другом в силу статистики Бозе — Эйнштейна, упоминавшейся в разделе 1.14. Этот эффект впечатляющим образом зафиксировали и объяснили Роберт Хэнбери Браун и Ричард К. Твисс (*эффект Хэнбери Брауна — Твисса*) в 1956 году, когда они измерили диаметр Сириуса, расположенного на расстоянии 2,4 млн км от Земли, и выявили существование квантовой запутанности на таких расстояниях [см. Hanbury Brownand Twiss, 1954, 1956 a, b]! Следовательно, можно сказать, что квантовые эффекты никоим образом не ограничены микроскопическими расстояниями;

нет никаких причин полагать, что дальность действия таких эффектов вообще чем-то ограничена. Более того, в общепринятом смысле до сих пор *нет* никаких наблюдений, которые бы *противоречили* прогнозам этой теории.

Соответственно, догматика квантовой механики кажется весьма обоснованной, поскольку зиждется на огромном массиве исключительно веских доказательств. Работая с достаточно простыми системами, в которых можно выполнять детальные вычисления и ставить достаточно точные эксперименты, мы обнаруживаем невероятно точное соответствие между теорией и результатами наблюдений. Более того, процедуры квантовой механики успешно применяются на самых разных масштабах; как упоминалось выше, квантовые эффекты прослеживаются и на уровне элементарных частиц, и между атомами и молекулами, квантовая запутанность фиксируется на расстояниях около 150 км, а также на отрезках в миллионы километров (диаметр звезды). Существуют точные квантовые эффекты, релевантные даже в масштабах всей Вселенной (см. раздел 3.4).

Эта догматика возникла не из пророчеств одной исторической личности, а из ожесточенных дискуссий многих выдающихся физиков-теоретиков, каждый из которых обладал феноменальными способностями и проницательностью. Это Планк, Эйнштейн, де Бройль, Бозе, Бор, Гейзенберг, Шрёдингер, Борн, Паули, Дирак, Йордан, Ферми, Вигнер, Бете, Фейнман и многие другие. Все они работали над математическими формулировками, руководствуясь результатами опытов, выполненных еще более многочисленными первоклассными экспериментаторами. Поразительно, что в этом отношении истоки квантовой механики значительно отличаются от начал общей теории относительности, чья теоретическая формулировка практически полностью является плодом размышлений Альберта Эйнштейна[1], а также обошлась почти без всяких наблюдений, кроме тех, что вписываются в теорию Ньютона. (Представляется, что Эйнштейн был хорошо осведомлен об уже известной ранее аномалии в движении Меркурия[2], что определенно могло повлиять на его ранние размышления, однако прямых доказательств в пользу такой версии нет.) Возможно, что в формулировке квантовой механики принимало участие такое множество теоретиков потому,

---

[1] Однако для вывода строгой математической формулировки этой теории Эйнштейну пришлось обратиться за помощью к коллеге Марселю Гроссману. Также следует отметить, что специальная теория относительности, в свою очередь, была сформулирована силами нескольких ученых, которые наряду с Эйнштейном также серьезно содействовали ее появлению, — это, в частности, Фогт, Фитцджеральд, Лоренц, Лармор, Пуанкаре, Минковский и др. [см. *Pais*, 1982].

[2] Общая теория относительности была окончательно сформулирована в 1915 году; Эйнштейн упоминает о перигелии Меркурия в письме от 1907 года (см. сноску 6 на с. 90 в главе о Дж. Ренне и Т. Зауэре в книге Goenne [1999]).

что эта теория по природе своей абсолютно не поддается интуиции. Однако с математической точки зрения она характеризуется замечательной красотой, а глубокая согласованность между математикой и физическими явлениями зачастую бывает столь же завораживающей, сколь и неожиданной.

С учетом всего этого, пожалуй, неудивительно, что догма квантовой механики при всей ее странности часто воспринимается как абсолютная истина, и считается, что любое явление природы обязательно должно согласовываться с этой догмой. Действительно, квантовая механика образует универсальную основу, которая, казалось бы, должна быть применима к любому физическому процессу независимо от его масштабов. Поэтому не приходится удивляться, что среди физиков укоренилась глубокая *вера* в то, что все явления природы должны подчиняться квантовой механике. Соответственно, исключительные странности, связанные с этим «предметом веры», на уровне повседневного опыта должны просто считаться такими феноменами, которые еще предстоит уложить в голове, как-то к ним привыкнуть и попытаться понять.

Особенно выделяется квантово-механический вывод, на который я уже указывал в разделе 1.4, а именно: квантовая частица может существовать в таком состоянии, когда она одновременно находится в двух разных местах. Несмотря на то что, согласно теории, аналогичный принцип должен применяться к любому макроскопическому объекту — даже к кошке, упомянутой в разделе 1.4, которая, казалось бы, должна была бы проходить через две дверцы одновременно, — ничего подобного мы никогда не наблюдаем, а также не имеем никаких причин полагать, что подобное явление когда-либо могло бы произойти в макроскопическом мире, даже если кот совершенно скрыт от нашего взгляда, когда проникает через дверцу (дверцы). Проблема кота Шрёдингера (правда, в исходной шрёдингеровской версии кот находился в суперпозиции двух состояний — «жив» и «мертв» [Schrödinger, 1935]) будет постоянно преследовать нас в ходе рассуждений, изложенных в последующих разделах этой части книги (см. разделы 2.5, 2.7 и 2.13). Мы убедимся в том, что, несмотря на нынешнюю убежденность в правильности квантовой физики (такая вера даже не обсуждается), от этих вопросов не так-то легко отмахнуться, и это в самом деле должно принципиально ограничивать нашу веру в квантовую физику.

2.2. Уравнение Макса Планка $E = hv$

Здесь и далее я должен буду гораздо более предметно говорить о структуре квантовой механики. Давайте начнем с основной первопричины, по которой

сложилось убеждение, что нам требуется еще что-то, кроме классической физики. Рассмотрим, в каких обстоятельствах выдающийся немецкий ученый Макс Планк в 1900 году постулировал принцип, представлявшийся с точки зрения всей известной на тот момент физики неслыханной дерзостью [Planck, 1901]. Однако тогда всей дерзости этого утверждения, пожалуй, по-настоящему не понимал ни сам Планк, ни кто-либо из его современников. На тот момент Планка интересовала ситуация, в которой материя и излучение, заключенные в неотражающей полости, находятся в равновесии с нагретой оболочкой полости (рис. 2.1) при заданной температуре. Он открыл, что этот материал должен испускать и поглощать электромагнитное излучение дискретными порциями согласно знаменитой ныне формуле

$$E = h\nu.$$

Здесь E — порция энергии, о которой только что шла речь; ν — частота излучения; h — фундаментальная константа природы, сегодня известная под названием *постоянная Планка*. Впоследствии стало понятно, что формула Планка описывает базовую связь между энергией и частотой, которая, согласно законам квантовой механики, соблюдается *везде и всегда*.

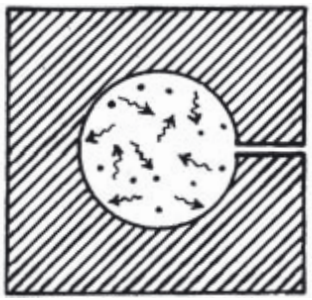


Рис. 2.1. Полость, подобная абсолютно черному телу, с черной внутренней поверхностью; в ней заключены вещество и электромагнитное излучение, которые пребывают в равновесии с нагретой оболочкой полости

Планк пытался объяснить наблюдаемое соотношение между интенсивностью и частотой излучения, показанное на рис. 2.2 сплошной линией. Это соотношение называется *спектром абсолютно черного тела*. Он возникает в условиях, когда вещество и энергия, взаимодействующие друг с другом, находятся в равновесии.

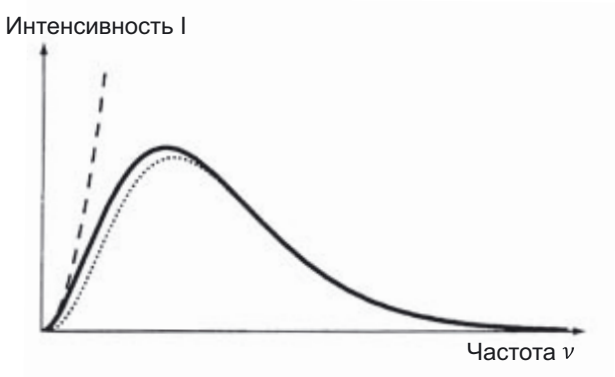


Рис. 2.2. Сплошная линия соответствует наблюдаемому соотношению между интенсивностью излучения абсолютно черного тела I и частотой излучения ν, которое описывается знаменитой формулой Планка. Штриховая линия соответствует формуле Рэлея — Джинса, а пунктирная — аппроксимации Вина

Ранее высказывались предположения о том, каким должно быть такое соотношение. Одно из них, именуемое *формулой Рэлея — Джинса*, описывает интенсивность I как функцию частоты ν следующим выражением:[1]

$$I = 8\pi k c^{-3} T \nu^2,$$

где T— температура, c — скорость света, а k — физическая константа, известная под названием *постоянная Больцмана*, о важности которой мы еще поговорим, особенно в части 3. Эта формула (ее график показан на рис. 2.2 штриховой линией) была основана на чисто классической интерпретации электромагнитного поля (по Максвеллу). Другое приближение, *аппроксимация Вина* (которую можно вывести из представления о том, что электромагнитное излучение состоит из произвольно движущихся классических безмассовых частиц), описывается формулой:

$$I = 8\pi h c^{-3} \nu^3 e^{-h\nu/kT}$$

(пунктирная линия на рис. 2.2).

Немало потрудившись, Планк установил, что может вывести очень точную формулу такого соотношения между интенсивностью и частотой излучения,

[1] В различных формулах этой части книги в начале выражения стоит множитель 8π, вместо которого иногда записывают просто множитель 2. Это чистая условность, связанная с определением термина *интенсивность* в этих выражениях.

которая, оказывается, хорошо согласуется с формулой Вина при больших значениях частоты и с формулой Рэлея — Джинса при малых значениях частоты:

$$\frac{8\pi h c^{-3}\nu^{3}}{e^{h\nu/kT}-1},$$

но для этого пришлось бы сделать очень странное предположение, что испускание и поглощение излучения веществом действительно всегда происходит дискретными порциями согласно вышеуказанной формуле $E = h\nu$.

По-видимому, Планк поначалу сам не осознал всей революционности своего предположения, и вопрос на некоторое время повис в воздухе — до тех пор, пока Эйнштейн [см. Pais, 1982; Stachel, 1995] четко не осознал, что и электромагнитное излучение как таковое должно состоять из отдельных порций энергии, подчиняющихся формуле Планка. Позже эти порции были названы *фотонами*. В действительности существует очень простая причина (которую явно не учитывали ни Планк, ни Эйнштейн), лежащая в основе равновесной природы излучения абсолютно черного тела и предполагаемой корпускулярной природы материальных тел, по которой и электромагнитное излучение (то есть свет) также должно иметь корпускулярную природу: а именно принцип, известный как *теорема о равнораспределении энергии*. Этот принцип утверждает, что по мере приближения конечной системы к равновесному состоянию энергия равномерно распределяется по всем степеням свободы этой системы.

Такой вывод из принципа равнораспределения энергии можно считать еще одним упражнением на тему функциональной свободы (см. разделы A.2, A.5, 1.9, 1.10 и 2.11). Предположим, что наша система состоит из N отдельных частиц, находящихся в равновесии с континуальным электромагнитным полем (причем частицы и поле могут обмениваться энергией благодаря тому, что некоторые частицы имеют электрический заряд). Функциональная свобода частиц составит ∞^{6N}, причем для простоты допустим, что мы имеем дело с классическими точечными частицами. Поэтому «6» означает, что у каждой частицы есть три степени свободы положения и три степени свободы импульса (*импульс*, в сущности, равен произведению массы на скорость; см. разделы 1.5, A.4 и A.6). С другой стороны, нам могут понадобиться некоторые параметры для описания *внутренних* степеней свободы — тогда мы заменим «6» большим целым числом. Например, если «частица» *неправильной* формы вращается классическим образом, то у нее будет еще шесть степеней свободы: три для ориентации в пространстве и три для направления и величины момента импульса (ср. раздел 1.14), что в сумме дает 12 степеней свободы на частицу и ∞^{12N} для числа степеней свободы всей классической системы, состоящей из N частиц. В разделе 2.9 будет показано, что в квантовой механике эти числа немного иные, но мы все равно

получаем функциональную свободу вида ∞^{kN} для некоторого целого k. Однако в случае с непрерывным электромагнитным полем у нас будет несравнимо большая функциональная свобода $\infty^{4\infty^3}$, что непосредственно следует из обсуждения в разделе А.2 (отдельно для электрического и магнитного полей).

Что в таком случае говорит нам принцип равнораспределения энергии о классической системе, где состоящее из частиц вещество и континуальное электромагнитное поле находятся в состоянии равновесия? Оказывается, по мере приближения к равновесию все больше и больше энергии будет распределяться по этим бесчисленным степеням свободы, заключенным в поле, и в итоге в поле перейдет вся энергия, распределенная по степеням свободы частиц. Коллега Эйнштейна Пауль Эренфест назвал такое явление *ультрафиолетовой катастрофой*, поскольку описанное состояние наступает на высокочастотном (то есть ультрафиолетовом) конце спектра, где в конце концов произойдет катастрофическое утекание степеней свободы в электромагнитное поле. Бесконечный рост на графике, представленном штриховой линией на рис. 2.2 (формула Рэлея — Джинса), как раз и иллюстрирует это явление. Однако если принять, что и поле состоит из отдельных частиц (что наиболее явно проявляется на высоких частотах), катастрофы удается избежать и становятся возможны согласованные равновесные состояния (я вернусь к этому вопросу в разделе 2.11 и постараюсь подробнее осветить возникающую в данном случае проблему функциональной свободы).

Из этих общих рассуждений понятно, что речь идет не просто об одном из свойств электромагнитного поля. Любая система, состоящая из континуальных полей, взаимодействующих с дискретными частицами, будет сталкиваться с такими же сложностями, как только речь зайдет о приближении к равновесию. Поэтому небезосновательно ожидать, что спасительное планковское соотношение $E = h\nu$ также может соблюдаться и для других полей. Соблазнительно предположить, что это свойство физических систем *универсально*. И действительно, для квантовой механики именно так оно и есть.

Однако глубокое значение планковской работы в этой области оставалось практически непонятым, пока в 1905 году Эйнштейн не опубликовал ныне знаменитую работу [Stachel, 1995, p. 177], где выдвинул экстраординарную идею: когда это уместно, электромагнитное поле действительно должно трактоваться как *система* частиц, а не как непрерывное целое. Это было настоящим шоком для физического сообщества, так как к тому моменту сложилось общепринятое мнение, что электромагнитные поля можно подробно описать красивой системой уравнений Максвелла (см. раздел 1.2). Наиболее убедительное свойство уравнений Максвелла состоит в том, что они, казалось бы, позволяют полностью описать свет как самораспространяющиеся электромагнитные волны, причем

такая волновая интерпретация объясняет многие специфические свойства света, в частности поляризацию и интерференцию. Также эти уравнения позволили спрогнозировать существование других типов электромагнитного излучения, отличного от видимого света, например радиоволн (частота которых гораздо ниже) и рентгеновских лучей (частота которых значительно выше). Чтобы утверждать, что свет в конечном итоге следует трактовать как совокупность частиц — в полном соответствии с предположением Ньютона, сделанным еще в XVII веке и, казалось бы, убедительно опровергнутым Томасом Янгом в начале XIX века, — требовалась, мягко говоря, изрядная дерзость. Пожалуй, еще более примечателен факт, что в том же 1905 году (который Эйнштейн назвал «годом чудес» [Stachel, 1995, p. 161–164 об $E= mc^2$ и p. 99–122 о теории относительности]) сам Эйнштейн написал свою еще более знаменитую статью о специальной теории относительности, в основу которой положил незыблемо верные максвелловские уравнения!

В довершение всего, даже *в той самой* статье, где Эйнштейн излагал свои идеи о корпускулярной природе света, он написал, что «теория Максвелла, пожалуй, никогда не будет заменена никакой другой» [Stachel, 1995, p. 177]. Хотя это, казалось бы, и противоречило цели всей статьи, мы, с учетом нашего более полного понимания квантовых полей, убеждаемся, что эйнштейновское корпускулярное представление об электромагнетизме в некотором глубинном смысле *не* противоречит максвелловской теории поля, поскольку современная квантовая теория электромагнитного поля возникает в результате общего процесса квантования поля, и эта процедура применяется именно к *теории Максвелла*! Также можно отметить, что и сам Ньютон понимал: у частиц света должны быть некоторые характерные *волновые* свойства [см. Newton, 1730]. Однако даже в эпоху Ньютона существовали определенные глубокие причины разделять точку зрения о корпускулярной природе распространения света. На мой взгляд, они определенно импонировали Ньютону в его рассуждениях [Penrose, 1987b, p. 17–49], и я считаю, что у Ньютона были очень веские причины наделять свет как корпускулярными, так и волновыми свойствами — даже сегодня эти причины сохраняют актуальность.

Следует четко отметить, что такое корпускулярное представление о физических полях (поскольку рассматриваемые частицы на самом деле являются бозонами и, следовательно, демонстрируют свойства, обеспечивающие эффект Хэнбери — Твисса, о котором шла речь в разделе 2.1) нисколько не нарушает их полностью «полевого» поведения на сравнительно низких частотах (то есть при большой длине волны) — именно такие свойства мы и наблюдаем. В некотором смысле квантовые составляющие природы нельзя с абсолютной точностью трактовать как частицы или как поля; вместо этого их лучше считать более таинственными

промежуточными сущностями (допустим, корпускулярно-волновыми), проявляющими как свойства волны, так и свойства частицы. Элементарные компоненты, кванты, подчиняются закону Планка $E = h\nu$. В целом, когда частота, а значит, и количество энергии на один квант, очень велики (соответственно длина волны мала), корпускулярная природа совокупности таких объектов начинает доминировать, поэтому мы получим достоверную картинку, предположив, что вся совокупность состоит из частиц. Но при низкой частоте (и соответственно малой энергии каждой отдельной частицы) речь идет о сравнительно длинных волнах (и неимоверном количестве низкоэнергетических частиц). В таком случае обычно хорошо работает классическое представление о «поле».

Как минимум, именно такова ситуация с бозонами (раздел 1.14). В случае с фермионами предельно длинные волны на самом деле не напоминают поле, поскольку тучи частиц, которые могли бы такое поле образовать, начинают мешать друг другу в соответствии с принципом Паули (см. раздел 1.14). Однако в определенных ситуациях (например, в сверхпроводнике) электроны (а это фермионы) могут образовывать так называемые *куперовские пары*, которые ведут себя примерно так же, как и отдельные бозоны. Вместе такие «бозоны» генерируют в сверхпроводнике сверхток, который может поддерживаться неопределенно долго без подпитки извне. Этот ток также обладает некоторыми когерентными свойствами, характерными для классического поля (правда, он в большей степени напоминает конденсат Бозе — Эйнштейна, вкратце упомянутый в разделе 2.1).

Универсальность планковской формулы $E = h\nu$ подсказывает, что корпускулярно-волновая природа полей должна проявляться и в противоположном случае: когда у объектов, которые мы привыкли считать частицами, наблюдаются свойства поля (или волны). Соответственно, и формула $E = h\nu$ должна в некотором виде применяться к таким обычным частицам, частота которых ν будет зависеть от их энергии, согласно формуле $\nu = E/h$. Оказывается, бывает и так, причем именно такую гипотезу выдвинул Луи де Бройль в 1923 году. Согласно теории относительности, покоящаяся частица массой m должна иметь энергию $E = mc^2$ (знаменитая формула Эйнштейна), так что, опираясь на отношение Планка, де Бройль присваивает этой частице естественную частоту $\nu = mc^2/h$, как уже отмечалось в разделе 1.8. Но, будучи в движении, частица также приобретает импульс p, и, опять же по требованию теории относительности, этот импульс должен быть обратно пропорционален естественной длине волны λ, присущей этой частице, где

$$\lambda = \frac{h}{p}.$$

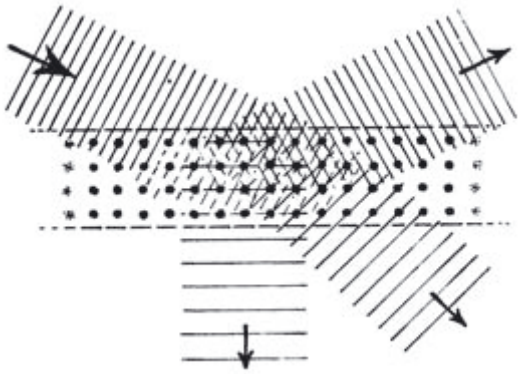


Рис. 2.3. Волновая природа электронов демонстрируется в эксперименте Дэвиссона — Джермера при бомбардировке электронами кристаллической мишени. Опыт показывает, что рассеяние или отражение зависит от соотношения периода кристаллической структуры и рассчитанной де Бройлем длины волны электрона

Эта формула де Бройля уже убедительно подтверждена неисчислимыми экспериментами, демонстрирующими, что частица с импульсом *p* будет демонстрировать интерференционные эффекты, как если бы она была волной с длиной λ. Одним из наиболее четких ранних примеров такого рода был опыт Дэвиссона — Джермера, поставленный в 1927 году. В ходе эксперимента электронами бомбардируют кристаллическую мишень. Опыт показывает, что рассеяние или отражение электронов от мишени зависит от соотношения периода кристаллической структуры и рассчитанной де Бройлем длины волны электрона (рис. 2.3). С другой стороны, сделанное ранее Эйнштейном предположение о корпускулярном характере света объяснило результаты наблюдений Филиппа Ленарда, проведенных им в 1902 году и связанных с *фотоэлектрическим эффектом*. Данное явление заключается в том, что если направить высокочастотный свет на металл, то свет высекает из металла электроны, обладающие конкретной энергией. Эта энергия зависит только от длины волны света, но, что удивительно, не зависит от его интенсивности. Такие результаты на тот момент казались загадочными, но именно их и объясняет гипотеза Эйнштейна (в итоге именно за это он получил в 1921 году Нобелевскую премию по физике) [Pais, 1982]. Более прямое и важное подтверждение эйнштейновской гипотезы обнаружил в 1923 году Артур Комптон. В его опыте квантами рентгеновских лучей бомбардировались заряженные частицы, и эти кванты действительно реагировали точно как безмассовые частицы, сегодня именуемые фотонами, в соответствии со стандартной релятивистской динамикой. В расчетах используется все та же формула де Бройля, однако на этот раз она читается в другом

направлении: импульс присваивается отдельным фотонам с длиной волны λ по формуле $p = h/\lambda$.

2.3. Корпускулярно-волновой парадокс

До сих пор мы все еще так практически и не подступились к самой структуре квантовой механики. Каким-то образом нам придется свести воедино волновые и корпускулярные аспекты наших простейших квантовых составляющих и сформировать более четкую картину. Чтобы лучше понять эти проблемы, давайте рассмотрим два разных (идеализированных) эксперимента. Они очень похожи, но в одном из них подчеркивается корпускулярная природа квантовой волны-частицы, а в другом — волновая. Для удобства я буду называть этот объект просто *частица*, или, точнее, *фотон*, поскольку на практике такие эксперименты действительно чаще всего ставятся на фотонах. Однако необходимо помнить, что все описанное с тем же успехом применимо и к электрону, и к нейтрону, и к любой другой волне-частице. В моем описании я не буду останавливаться на каких-либо технических сложностях, которые могут возникать при практической реализации этих экспериментов.

В каждом эксперименте в качестве источника света используется, скажем, лазер, расположенный в точке L на рис. 2.4. Мы направляем отдельно взятый фотон на полупосеребренное зеркало, расположенное в точке M. Правда, в реальных экспериментах вряд ли использовался бы отражатель с таким серебрением, как обычное настольное зеркальце. (В таком квантово-оптическом эксперименте могли бы использоваться более качественные зеркала, специально рассчитанные на активное взаимодействие с волновой природой нашего фотона, например, на уровне интерференционных эффектов, но в данном контексте это неважно.) Технически такое устройство можно назвать *светоделитель*. В этих экспериментах светоделитель должен располагаться под углом 45° относительно лазерного луча, тогда ровно половина попадающего на него света будет отражаться под прямым углом, а половина — проходить сквозь него.

В эксперименте 1, показанном на рис. 2.4 *a*, у нас два детектора: один находится в точке A, то есть на пути проходящего луча (поэтому LMA — прямая линия), а другой — в точке B, то есть на пути отраженного луча (поэтому LMB — прямой угол). Я исхожу из того (для простоты изложения), что каждый детектор абсолютно исправен, то есть он срабатывает тогда и только тогда, когда на него попадает фотон. Более того, я предполагаю, что остальное оборудование в эксперименте также идеально, поэтому фотоны не теряются в результате поглощения, отклонения или каких-либо других накладок. Я также полагаю, что лазер

оснащен специальным средством, позволяющим фиксировать каждый «пуск» фотона.

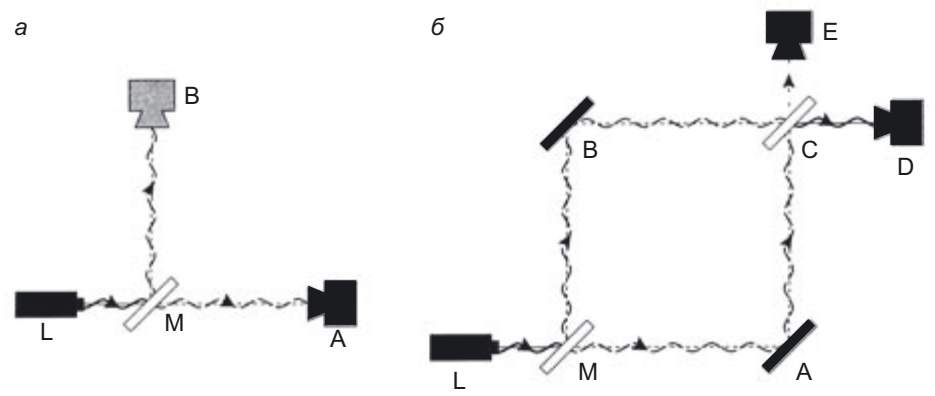


Рис. 2.4. Корпускулярно-волновые свойства фотона, испускаемого лазером, расположенным в точке L и направленным на светоделитель в точке M; *а*) эксперимент 1: корпускулярные свойства фотона фиксируются детекторами в точках A и B, каждый выпущенный фотон фиксируется ровно на одном детекторе; *б*) эксперимент 2 иллюстрирует волновые свойства фотона (интерферометр Маха — Цендера) с зеркалами A и B, детекторами, расположенными в точках D и E, и вторым светоделителем — в точке C; выпущенный фотон попадает только в точку D

Итак, в эксперименте 1 при каждом пуске фотона этот фотон регистрируется либо на детекторе A, либо на детекторе B, но не на обоих сразу. Вероятность любого результата — 50 %. Здесь мы наблюдаем *корпускулярные* свойства фотона. Фотон летит одним или другим путем, результаты этого эксперимента согласуются с «корпускулярной» посылкой о том, что в момент попадания фотона на светоделитель должно выясниться, отразится этот фотон или проскочит дальше. На первый и на второй вариант приходится по 50 % случаев.

Теперь обратимся ко второму эксперименту, показанному на рис. 2.4 *б*. Здесь мы ставим на место детекторов A и B полностью посеребренные зеркала, расположенные под углом 45°. Поэтому всякий раз луч, падающий на зеркало, отражается на второй светоделитель C, идентичный первому и опять же расположенный под углом 45° (итак, все зеркала и все светоделители установлены параллельно друг другу). Теперь два детектора стоят в точках D и E (где прямая CD параллельна LMA, а прямая CE параллельна MB), а MACB — прямоугольник (на рисунке показан как квадрат). Такая система называется *интерферометром Маха — Цендера*.

Итак, что же происходит, когда лазер испускает фотон? Может показаться, что, в соответствии с экспериментом 1, фотон должен вылететь из светоделителя в точке M и с 50 %-й вероятностью избрать либо путь MA (то есть отразиться по линии AC), либо путь MB (то есть отразиться по линии BC). Соответственно, светоделитель в точке C с 50 %-й вероятностью получит фотон, пришедший по пути AC, и с равной вероятностью может отправить его на детектор D или E, либо с 50 %-й вероятностью получит фотон, пришедший по пути BC, который, аналогично, с равной вероятностью может отправить на детектор D или E. Таким образом, детекторы D и E имеют по 50 % шансов (25 % + 25 %) получить фотон.

Однако на практике все *иначе*! Бесчисленные эксперименты на таких или аналогичных установках показывают, что фотон попадает на детектор D со 100 %-й вероятностью, а вероятность попадания на детектор E нулевая! Это полностью противоречит корпускулярной трактовке свойств фотона, описанной в эксперименте 1. С другой стороны, результат эксперимента 2 выглядит гораздо логичнее, если уподобить фотон волне. Тогда можно изобразить, что происходит на светоделителе M в случае применения описанной установки (при условии, что эксперимент идеален). Волна разделяется на два более мелких цуга, один из которых идет по плечу MA, а другой — по плечу MB. Они отражаются соответственно от зеркал A и B, поэтому на второй светоделитель C одновременно приходят две волны: одна — с AC, другая — с BC. Обе они в точке C делятся на составляющие CD и CE, но чтобы рассмотреть, как они складываются, нужно уделить внимание фазовым соотношениям. Мы видим (предполагается, что длина плеча в обоих случаях одинакова), что две волны, сходящиеся по линии CE, находятся в противофазе и полностью *гасят* друг друга, поскольку гребни одной волны накладываются на впадины другой. Напротив, волны, сходящиеся по линии CD, *усиливают* друг друга, так как их гребни и впадины совпадают. Таким образом, суммарная волна после второго светоделителя пойдет в направлении CD, а в направлении CE ничего не окажется. Эта картина соответствует 100 % обнаружению фотонов в D и 0 % в E, причем именно это и наблюдается на практике.

Можно сделать вывод, что волну-частицу логичнее сравнивать с так называемым *волновым пакетом*, похожим на небольшой всплеск колебательных волноподобных явлений, но локализованных в очень ограниченной области, поэтому в крупном масштабе возникает небольшое точечное возмущение, напоминающее частицу (см. рис. 2.11 в разделе 2.5, где изображен волновой пакет в соответствии с представлениями стандартного квантового формализма). Однако с точки зрения квантовой механики такая картинка почти ничего не объясняет — по целому ряду причин. Во-первых, формы волн, часто ис-

пользуемых в подобных экспериментах, совсем не напоминают упомянутые волновые пакеты, поскольку единственный фотон может иметь длину волны, которая окажется больше размеров всего аппарата. Гораздо важнее, что такие иллюстрации и близко не объясняют, что происходит в эксперименте 1, для чего нам снова придется вернуться к рис. 2.4 *а*. Чтобы наша модель с волновыми пакетами согласовывалась с результатами эксперимента 2, фотонная волна должна разделиться на два более мелких волновых пакета в светоделителе в точке M, причем один пакет отправится в A, а другой — в B. Чтобы воспроизвести результаты эксперимента 1, нам, по-видимому, придется считать, что детектор в точке A, получив такой волновой пакет меньшего размера, должен с 50 %-й вероятностью срабатывать в ответ на это событие (именно по такому принципу он регистрирует попадания фотонов). То же касается детектора в точке B — он также должен с 50 %-й вероятностью регистрировать попадание фотона. Да, все складно, но эти выкладки не соответствуют действительности, поскольку оказывается, что хотя эта модель и прогнозирует равные шансы поймать фотон в точках A и B, половина этих попаданий *вообще не случается*. Ведь данная гипотеза неверно предполагает, что в 25 % случаев *оба* детектора зарегистрируют попадание фотона, а в 25 % случаев ни один детектор ничего не зарегистрирует! Два таких «комбинированных» результата никогда не возникают, поскольку в ходе эксперимента ни один фотон не теряется и не дублируется. Подобное описание фотона как отдельного волнового пакета попросту не работает.

На самом деле квантовые физические процессы устроены гораздо тоньше. Волна, описывающая квантовую частицу, не похожа ни на водную рябь, ни на звуковую волну, которая соответствовала бы некоему *локальному* возмущению в окружающей среде, поэтому эффект от воздействия волны на детектор в одной точке никак не отражался бы на втором детекторе, установленном совсем в другом месте. На примере эксперимента 1 видно, что волновая ипостась фотона после того, как он будет «разрезан» в светоделителе на два синхронных луча, так и остается *одиночной* частицей, несмотря на такое разделение. По-видимому, волна описывает своеобразное распределение вероятностей обнаружения частицы в различных местах. Такое описание немного точнее передает поведение этой волны, и такую волну действительно иногда называют *вероятностной*. Однако эта картина по-прежнему несовершенна, поскольку вероятности могут иметь только положительное (или нулевое) значение и не могут обнулять друг друга — в таком случае не удается объяснить полное отсутствие откликов в E в эксперименте 2.

Иногда предпринимаются попытки объяснить такие свойства в контексте «волн вероятности». При этом допускается, что в некоторых точках значения

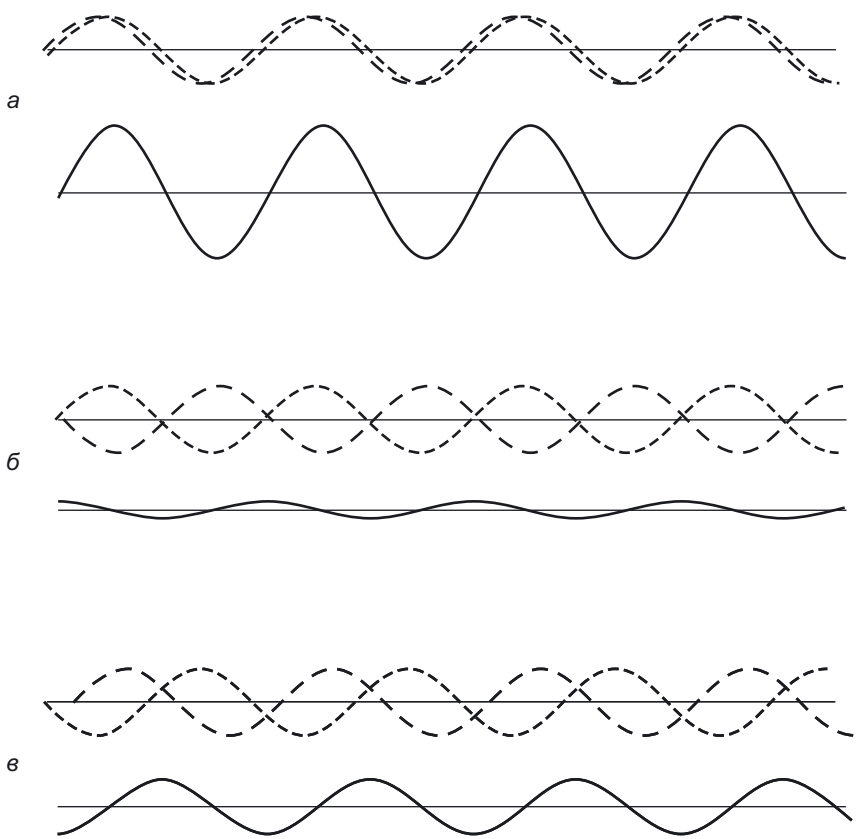


Рис. 2.5. Сумма двух мод волны с равными значениями амплитуды и частоты (здесь они показаны в виде штриховых кривых), причем эти моды могут усиливать друг друга (*а*), гасить друг друга (*б*) или давать некий промежуточный результат (*в*) в зависимости от того, в какой фазе друг относительно друга находятся эти моды

вероятности каким-то образом могут становиться *отрицательными*, так, чтобы волны могли погасить друг друга. Однако в действительности квантовая теория устроена иначе (рис. 2.5). На самом деле нужно сделать еще один шаг и понять, что амплитуды волн могут выражаться *комплексными числами* (см. раздел А.9). Ведь такие комплексные амплитуды уже вскользь упоминались в разделах 1.4, 1.5 и 1.13. Комплексные числа критически важны для всей структуры квантовой механики. Они тесно связаны с вероятностями, но *не являются* вероятностями (и, естественно, быть таковыми не могут, поскольку вероятности выражаются вещественными числами). Тем не менее роль комплексных чисел в квантовом формализме — гораздо более масштабная, и в этом мы вскоре убедимся.

2.4. Квантовые и классические уровни: C, U и R

В 2002 году меня пригласили выступить с лекцией в датском городе Оденсе в академии Ганса Христиана Андерсена. Близился двухсотлетний юбилей Андерсена, родившегося в Оденсе в 1805 году. Думаю, меня позвали туда потому, что ранее я написал книгу «Новый ум короля», название которой намекает на сказку Андерсена «Новое платье короля». Однако я хотел рассказать о чем-нибудь другом, поэтому задумался, есть ли у Андерсена другая сказка, которая позволила бы проиллюстрировать идеи, занимавшие меня в последнее время. Эти идеи преимущественно касались основ квантовой механики. После некоторых размышлений мне показалось, что сказка «Русалочка» поможет мне проиллюстрировать многие вопросы, которые я собирался обсудить.

Взгляните на рис. 2.6, на котором я изобразил русалочку, сидящую на скале. Половина ее тела скрыта под водой. В нижней части рисунка показано, что происходит под водой — там творится настоящая кутерьма, плавают разные диковинные твари, напоминающие инопланетян, но, пожалуй, не лишенные собственной особой прелести. Так я демонстрирую странный и чужеродный мир квантовой механики. В верхней части картинки видим более привычный мир, где отдельные объекты четко отделены друг от друга; это физические тела, каждое из которых функционально самодостаточно. Это мир классической физики, законы которого нам более привычны и которые мы успели изучить до пришествия квантовой механики, считая, что они в точности описывают поведение вещей. Русалочка сочетает в себе оба мира — она полудева-полурыба. Русалочка связывает два эти взаимно несовместимых мира (рис. 2.7). Кроме того, Русалочка кажется таинственной и волшебной, поскольку ее способность связывать эти миры, по-видимому, нарушает законы обоих. Более того, поскольку ей знаком подводный мир, она по-особому воспринимает и наш мир, на который взирает словно свысока, удобно расположившись на скале.

Большинство физиков — и я в том числе — считают, что различные множества физических явлений должны подчиняться не принципиально разным законам, а некоей всеобъемлющей системе универсальных законов (или общих принципов), в которую вписываются *все* физические процессы.

С другой стороны, многие философы и, бесспорно, очень заметная партия физиков, полагают, что вполне могут существовать принципиально различные уровни реальности, на каждом из которых действуют свои законы физики; таким образом, нет необходимости искать какую-то всеобъемлющую систему, которая бы описывала их все (см., например, [Cartwright, 1997]). Разумеется, когда условия окружающей среды действительно радикально отличаются от

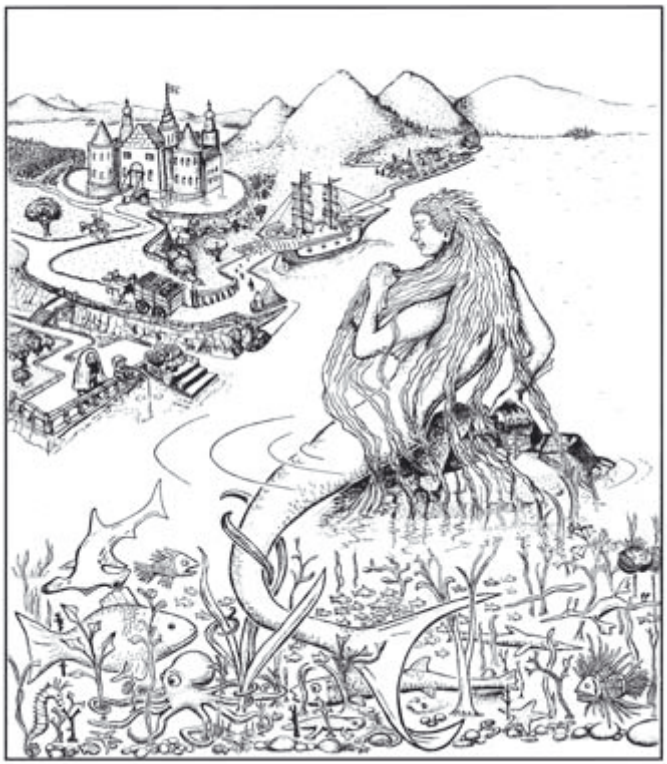


Рис. 2.6. Магия и тайны квантовой механики, которые автор проиллюстрировал, вдохновленный сказкой Ганса Христиана Андерсена «Русалочка»

привычных нам, логично ожидать, что там вступают в свои права и иные физические законы — или по меньшей мере необычные *проявления* универсальных законов. Действительно, в любой насущной ситуации мы определенно должны уделять максимальное внимание наиболее релевантным законам и вполне можем проигнорировать остальные. Тем не менее любые *фундаментальные* законы, которые мы игнорируем в силу незначительности их воздействия, все равно могут косвенно влиять на ситуацию. Как минимум среди физиков это мнение общепринято. Мы ожидаем — хотя, возможно, это всего лишь *вера,* — что физика должна представлять собой единое *целое*, и даже если непонятно, какова роль конкретного физического закона, у него все равно должно быть свое место в общей картине и он может быть очень важен для ее целостности.

Соответственно, два мира, изображенные на рис. 2.7, не следует считать совершенно чуждыми друг другу. Однако в современном понимании квантовой теории

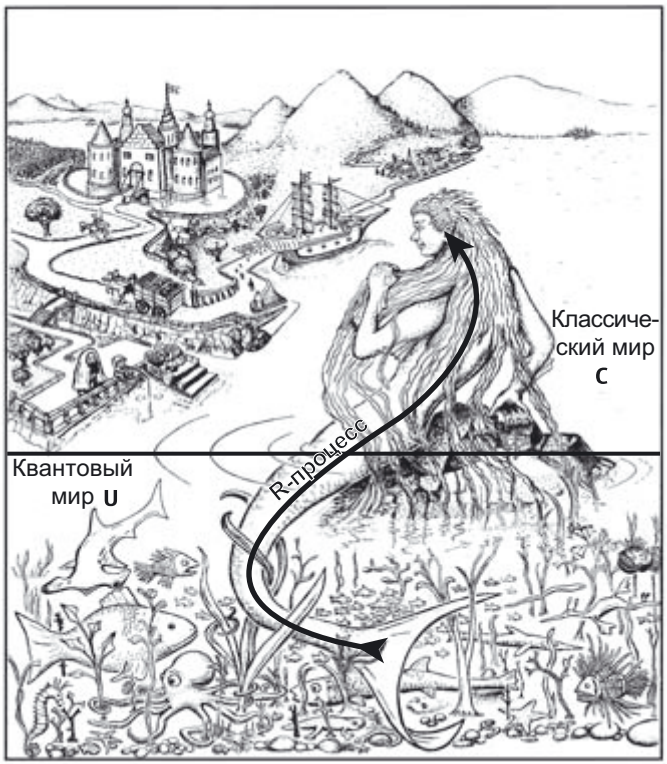


Рис. 2.7. В верхней части рисунка — знакомый нам мир классической физики **C**, состоящий из отдельных физических явлений. В нижней части — чуждый квантовый мир **U**, где царит квантовая запутанность. Русалочка объединяет оба мира, и ее я сравниваю с таинственным **R**-процессом, позволяющим квантовым сущностям влиять на устройство привычного мира

и ее связи с макромиром нам просто удобно интерпретировать явления именно так, словно они относятся к разным мирам и поэтому подчиняются разным законам. На практике мы действительно склонны использовать одну совокупность законов на так называемом *квантовом уровне*, а другую на уровне *классической физики*. Граница между двумя этими уровнями никогда не была совершенно четкой, и существует распространенное убеждение, что классическая физика — в любом случае всего лишь удобное приближение к «истинной» квантовой физике, которая *полностью* выполняется на уровне ее же фундаментальных компонентов. Считается, что классическое приближение обычно является исключительно точным, если речь идет о процессе с участием огромного количества квантовых частиц. Далее я покажу (в частности, в раз-

делах 2.13 и 4.2), что все-таки если слишком строго придерживаться такой удобной точки зрения, она оборачивается изрядными сложностями. Но пока я постараюсь с ними мириться.

Следовательно, в принципе, квантовую физику следует считать очень точной на уровне «малых» объектов, но понимать, что гораздо проще постижимая классическая физика весьма хорошо описывает свойства «больших» тел. Но в этом контексте слова «большой» и «малый» нужно употреблять очень осторожно, поскольку, как отмечалось в разделе 2.1, в некоторых обстоятельствах квантовые эффекты могут проявляться на очень больших расстояниях (определенно свыше 143 км). Далее, в разделах 2.13 и 4.2, я изложу точку зрения, согласно которой одно лишь расстояние — недостаточный критерий, чтобы охарактеризовать явление как «вот теперь — классическое», но пока для нас не столь важно учитывать подобный специфический критерий.

Итак, давайте пока не будем отходить от общепринятой точки зрения, что физические явления делятся на квантовые и классические лишь для удобства и что есть тела в некотором смысле «малые» — настолько, что их можно описывать динамическими уравнениями квантовой теории, а есть «большие», свойства которых исключительно точно описываются классической динамикой. Так или иначе, именно такая точка зрения почти всегда принимается на практике, и она поможет нам понять, как именно используется квантовая теория. Обычный мир исключительно строго подчиняется классическим законам Ньютона, дополняемым законами Максвелла, описывающими непрерывные электромагнитные поля. Законы Максвелла, в свою очередь, дополняются законом Лоренца, описывающим взаимодействие отдельных заряженных частиц с электромагнитными полями (см. разделы 1.5 и 1.7). Если речь пойдет о материи, которая очень быстро движется, то потребуется учитывать законы специальной теории относительности, а если в процессе также участвуют достаточно сильные гравитационные потенциалы, то в дело вступает общая теория относительности Эйнштейна. Все эти законы образуют согласованное целое, в рамках которого физические явления подчиняются дифференциальным уравнениям (см. раздел А.11), причем очень *детерминированно* и *локально*. Свойства пространства-времени выводятся из данных, которые могут быть указаны для любого конкретного момента времени (причем в общей теории относительности «в определенный момент времени» понимается как «на определенной исходной пространственноподобной поверхности»; см. раздел 1.7). В этой книге я решил не вдаваться в какие-либо детали дифференциального исчисления — нам всего лишь требуется знать, что эти дифференциальные уравнения описывают поведение системы в будущем (или прошлом) именно на основании ее состояния (и движения) в любой момент времени. Такую классическую эволюцию я обозначу буквой **C**.

Квантовый мир, в свою очередь, эволюционирует во времени — такое развитие я обозначу буквой **U**, означающей *унитарную* эволюцию. Такая эволюция описывается иным уравнением, которое называется *уравнением Шрёдингера*. В данном случае речь по-прежнему идет о детерминированной и локальной эволюции во времени (подчиняющейся дифференциальному уравнению; см. раздел A.11), которая касается математической сущности, именуемой *квантовым состоянием*. Квантовое состояние было введено в квантовую теорию для описания системы в любой момент времени. Этот детерминизм очень напоминает классический, но существуют разнообразные ключевые различия между процессом **U** и классическим эволюционным процессом **C**. На самом деле некоторые из этих различий, в частности отдельные следствия *линейности*, о которых мы еще поговорим в разделе 2.7, влекут последствия, столь несовместимые с нашими представлениями о фактических свойствах мира, что исчезает всякий смысл продолжать придерживаться **U** при описании реальности, после того как появляется возможность перейти к макроскопическим альтернативам. Однако в квантовой теории мы избираем третий путь и прибегаем к *квантовому измерению* — эту процедуру я обозначу буквой **R** (от слова *редукция*). Именно здесь Русалочка и играет свою важнейшую роль, выступая в качестве необходимого звена, связывающего квантовый мир с повседневным, который мы непосредственно воспринимаем. Процедура **R** (см. раздел 2.8) полностью отличается от детерминированной эволюции (случаев **C** или **U**) и является *вероятностным* актом. Как мы убедимся в разделе 2.10, ей присущи любопытные нелокальные свойства, ниспровергающие все наши привычные представления о классических законах.

Чтобы составить некоторое впечатление о роли **R**, рассмотрим, как работает счетчик Гейгера — широко известный прибор для обнаружения заряженных частиц, образующихся при радиоактивном распаде. Любая подобная отдельная частица будет считаться квантовым объектом и подчиняться законам квантового уровня **U**. Однако счетчик Гейгера — классический измерительный прибор, который, подобно увеличительному стеклу, позволяет заметить воздействия квантовой частицы на классическом уровне: регистрация каждой частицы сопровождается характерным щелчком. Поскольку мы непосредственно воспринимаем такой щелчок, мы считаем его явлением из знакомого нам классического мира, то есть результатом классических вибраций звуковых волн, которые можно весьма точно описать классическими (ньютоновскими) уравнениями гидродинамики (аэродинамики). Короче говоря, эффект **R** замещает непрерывную эволюцию, которую дает нам **U**, и является внезапным *скачком* к одному из возможных классических **C**-описаний. В примере со счетчиком Гейгера до щелчка все такие альтернативы должны проистекать из различных составляющих квантового состояния частицы, которое формируется в ходе **U**-эволюции, причем частица может находиться в одном месте или в другом, двигаться так или иначе, и все

эти варианты существуют в виде квантовой суперпозиции, рассмотренной в разделе 1.4. Однако когда в дело вступает счетчик Гейгера, альтернативы из квантовой суперпозиции превращаются в различные возможные классические результаты, причем *щелчок* может произойти в тот или иной момент времени, и этим моментам присваиваются вероятностные значения. Вот и все, что остается при наших расчетах от **U**-эволюции.

Та процедура, которая обычно используется в таких ситуациях, весьма точно согласуется с подходом Нильса Бора и представителей его копенгагенской школы. Невзирая на философскую подоплеку, которую Бор пытался приписать своей *копенгагенской интерпретации* квантовой механики, она на самом деле дает очень прагматичный вариант трактовки квантового измерения **R**. Грубо говоря, представление о «реальности» в основе этого прагматизма таково: измерительный прибор (например, только что рассмотренный счетчик Гейгера) и его произвольное окружение могут рассматриваться как элементы настолько крупной и сложной системы, что было бы просто нецелесообразно пытаться дать ей точную характеристику в соответствии с законами **U**. Поэтому мы лучше будем описывать прибор (и окружающую среду, в которой он находится) как *классическую* по сути систему, классические эффекты которой предположительно дают весьма точное приближение «верного» квантового поведения, так что наблюдаемые явления, следующие за квантовым измерением, можно в дальнейшем очень точно описать на уровне классических законов **C**. Тем не менее переход от **U** к **C**, как правило, невозможен *без* привлечения вероятностей, поэтому детерминизм, присутствующий в уравнениях **U** (и **C**), нарушается (рис. 2.8), и квантовое описание, как правило, не обходится без «скачка» согласно устройству **R**. Считается, что усложнение, присущее любой «реальной» физике, лежащей в основе измерительного процесса, можно считать столь масштабным, что предположительно верное описание, соответствующее **U**, будет совершенно

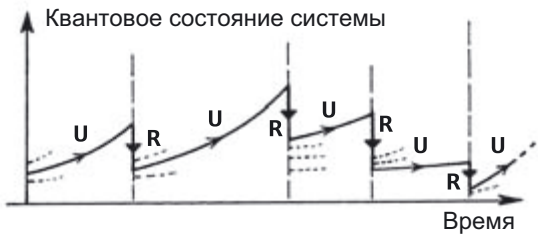


Рис. 2.8. По-видимому, именно так устроен мир квантовой теории: периоды детерминированной **U**-эволюции перемежаются с точечными моментами вероятностных **R**-действий, причем каждое из них воспроизводит некое классическое явление

оторвано от практики, и в лучшем случае можно надеяться на некоторое приблизительное описание, которое будет характеризоваться вероятностными, а не детерминированными свойствами. Соответственно, вполне можно полагать, что законы **R** адекватно описывают квантовую ситуацию. Однако, как мы убедимся в разделе 2.12, и особенно 2.13, с такой точкой зрения связаны глубоко мистические загадки, и оказывается очень сложно признать, что причудливые **U**-законы, примененные непосредственно к макроскопическим телам, могут провоцировать в этих телах **R**-подобные или **C**-подобные явления. Именно здесь и начинают во всей красе вырисовываться серьезные сложности с «интерпретацией» квантовой механики.

Таким образом, копенгагенская интерпретация Бора, прежде всего, совершенно не требует наделять квантовый уровень неким «реализмом». На самом деле процедуры **U** и **R** в ней трактуются просто как способ составить основу *вычислительной процедуры*, дающей математическое описание развития системы во времени и позволяющей в момент измерения воспользоваться **R** для расчета вероятностей различных возможных исходов измерения. **U**-действия квантового мира, хотя они и не считаются физически реальными, по-видимому, требуется «держать в уме» и использовать только таким прагматическим способом — для вычислений, поэтому затем при помощи **R** можно узнать лишь те вероятности, которые нам нужны. Более того, тот *скачок*, который обычно претерпевает квантовое состояние, когда задействуется **R**, не считается настоящим физическим процессом, но представляет собой лишь изменение *знаний*, которыми обладает физик, получающий в результате измерения дополнительную информацию.

Мне представляется, что изначальная ценность такой копенгагенской интерпретации заключалась в следующем: она позволяла физикам прагматично работать с квантовой механикой, получая при этом множество удивительных результатов и, помимо того, избавляла от обременительной необходимости глубоко разбираться, что же «действительно происходит» в квантовом мире, а также соотносить его с классическим миром, который мы, казалось бы, непосредственно воспринимаем. Однако сегодня нам мало такой интерпретации. Уже есть множество значительно более свежих идей и экспериментов, которые позволяют нам непосредственно прозондировать природу квантового мира и в значительной степени подтверждают, насколько реальны *странные* описания, формулируемые в рамках квантовой теории. Мы определенно ощутим силу такой квантовой «реальности», если попытаемся докопаться до возможных границ догматики, принятой в квантовой механике.

В нескольких следующих разделах (2.5–2.10) мы исследуем основные элементы величественного математического аппарата квантовой механики. Это система,

которая с невероятной точностью соответствует устройству реального мира. Многие следствия этой системы совершенно нелогичны, а некоторые откровенно противоречат нашим ожиданиям, которые мы крепко усвоили на опыте всей нашей жизни в классическом мире. Более того, некоторые из этих экспериментов показывают, что в противовес выстраданной нами классической интуиции эффекты квантового мира обнаруживаются не только на субмикроскопических отрезках, но могут простираться на значительные расстояния (текущий рекорд, зафиксированный на Земле, — 143 км). Распространенная научная *вера* в формализм квантовой механики и в самом деле серьезно подкреплена наблюдаемыми фактами!

Наконец, в разделах 2.12 и 2.13 я обосную мой тезис о том, что тотальная вера в такой формализм неуместна. Хотя и признается, что в квантовой теории поля (КТП) существует проблема с расходимостями, что и стало одним из первых мотивов для создания теории струн (см. раздел 1.6), выдвигаемые мною аргументы касаются только лишь более фундаментальных и всеобъемлющих законов квантовой механики как таковой. В этой книге я не углубляюсь в обсуждение нюансов КТП, за исключением упомянутых в разделах 1.3, 1.5, 1.14 и 1.15. Однако в разделе 2.13, а затем в разделе 4.2 я выдвигаю свои собственные идеи относительно ее изменения, которое я считаю ключевым, и объясняю, почему существует явная необходимость освободиться от этой пресловутой тотальной веры в догму стандартного квантового формализма.

2.5. Волновая функция точечной частицы

В чем же тогда *заключается* этот стандартный квантовый формализм? Как вы помните, в разделе 1.4 я упоминал так называемый *принцип суперпозиции*, который универсальным образом применяется в квантовых системах. Там мы рассматривали ситуацию, когда некие частицы одна за другой пропускаются через две очень близкие параллельные щели, расположенные перед чувствительным экраном (см. рис. 1.2 *г* в разделе 1.4). На экране появлялся рисунок из пятнышек, напоминавший картину, которая осталась бы от множества точечных попаданий огромного количества частиц, выпущенных из одного источника. Этот рисунок подтверждает фактическую корпускулярную природу квантовых объектов, летящих в экран. Тем не менее в общем рисунке попаданий на экран явно просматривается система параллельных полос, что четко свидетельствует об интерференции. Именно такой интерференционный узор образовался бы при наложении волноподобных объектов из двух когерентных источников, если бы волны одновременно проходили через две щели. Однако создается впечатление, что каждая из точек на экране — это результат попадания отдельного объекта.

Такой результат станет еще явственнее, если снизить интенсивность бомбардировки так, чтобы интервал между пусками отдельных частиц оказался более длительным, чем время подлета частицы к экрану. Значит, мы действительно имеем дело с частицами, попадающими на экран одна за другой, но каждый такой корпускулярно-волновой объект демонстрирует интерференцию, возникающую между различными возможными траекториями движущихся частиц.

Все это очень напоминает эксперимент 2 из раздела 2.3 (см. рис. 2.4 *б*), где каждая корпускулярно-волновая частица вылетает из светоделителя (в точке M) в виде двух пространственно-разделенных компонентов, которые впоследствии встречаются и интерферируют на втором светоделителе (в точке C). Мы вновь наблюдаем, что *одиночный* корпускулярно-волновой объект может состоять из двух отдельных частей, и при последующем «воссоединении» этих частей могут возникать интерференционные эффекты. Следовательно, каждая такая волна-частица совсем не обязана быть локализованным объектом. Тем не менее она обладает свойствами единого целого. Она продолжает действовать как одиночный квант независимо от того, насколько разнесены ее компоненты и сколько таких компонентов может быть.

Как описать подобный странный корпускулярно-волновой объект? Даже если его природа кажется нам чуждой, нам повезло — она поддается очень красивому математическому описанию, поэтому мы, пожалуй, можем удовлетвориться (как минимум временно, в соответствии с так называемой копенгагенской интерпретацией) тем, что нам под силу сформулировать точные математические законы, которым подчиняется такой объект. Основное математическое свойство заключается в том, что, как и в случае с электромагнитной волной, мы можем *складывать* два таких корпускулярно-волновых состояния (на что я намекал в разделе 1.4). Кроме того, эволюция такой суммы *равна* сумме эволюций — это свойство вкладывается в термин *линейный* и будет подробнее описано в разделе 2.7 (также см. раздел A.11). Как раз такое сложение и называется *квантовой суперпозицией*. Когда у нас есть корпускулярно-волновой объект, состоящий из двух частей, то вся волна-частица будет просто суперпозицией двух этих частей. Каждая часть будет сама обладать свойствами корпускулярно-волнового объекта, но *целостный* корпускулярно-волновой объект будет суммой двух частей.

Более того, существуют разные способы сложения двух таких корпускулярно-волновых частиц в зависимости от фазовых соотношений между ними. Какие это способы? Нам предстоит узнать, что они возникают на уровне математического формализма при использовании комплексных чисел (см. разделы A.9 и 1.4, а также заключительные замечания из раздела 2.3). Следовательно, если обозначить

через **α** состояние одного из этих корпускулярно-волновых объектов, а через **β** — состояние другого, то можно построить различные комбинации **α** и **β** в виде

$$w\boldsymbol{\alpha} + z\boldsymbol{\beta},$$

где отдельные весовые коэффициенты w и z — это комплексные числа (точно как в ситуации, рассмотренной в разделе 1.4), именуемые, пожалуй, немного неожиданно — комплексными *амплитудами*[1], присваиваемыми соответствующим альтернативам — **α** и **β**. Существует правило: если предположить, что обе амплитуды — w и z — являются ненулевыми, то такая комбинация дает нам иное возможное состояние нашей волны-частицы. На самом деле семейство квантовых состояний в любой квантовой ситуации должно образовывать *комплексное векторное пространство* в том смысле, в каком этот термин понимается в разделе А.3, и мы подробнее исследуем этот аспект квантовых состояний в разделе 2.8. Нам необходимо, чтобы эти весовые коэффициенты w и z были комплексными (а не, скажем, неотрицательными вещественными числами, которые получились бы, если бы нас интересовали лишь вероятностно-взвешенные альтернативы), поэтому можно выразить *фазовые соотношения* между двумя компонентами **α** и **β**, имеющие ключевое значение для *интерференционных* эффектов, которые могут возникать между этими компонентами.

Такие интерференционные эффекты, обусловленные фазовыми соотношениями между компонентами **α** и **β**, проявляются как в эксперименте с двумя щелями из раздела 1.4, так и при работе с интерферометром Маха — Цендера (см. раздел 2.3), поскольку по сути каждое состояние представляет собой временны́е *колебания* с определенной частотой, которые могут гасить или усиливать друг друга (соответственно при несовпадении или совпадении фаз) в различных обстоятельствах (см. рис. 2.5 в разделе 2.3). Все зависит от отношений между состояниями **α** и **β**, а также между присущими каждому из них амплитудами w и z. На самом деле все, что нам нужно знать об амплитудах w и z, — это их *отношение* w/z (при этом допускается нулевое значение у b — в последнем случае мы можем просто присвоить этому отношению «числовое» значение «∞»). Однако необходимо следить, чтобы a и b не оказались *одновременно* равны нулю!

В контексте (комплексной) плоскости Весселя (см. раздел А.10), показанной на рис. А.42, у данного отношения w/z есть *аргумент* (при этом предполагается,

---

[1] В литературе, на которую я ссылаюсь в разделе А.10, *аргументом* комплексного числа (то есть θ в полярном представлении $re^{i\theta}$) иногда называется его *амплитуда*. С другой стороны, *интенсивность* волны (в этом выражении ей соответствует буква r) также часто называется ее *амплитудой*! Здесь я избегаю подобных противоречивых вариантов, и моя терминология (*стандартная* для квантовой механики) включает *оба* этих термина!

что w и z не равны нулю), и этот аргумент — угол θ между двумя лучами, выходящими из начала координат (0) в направлении точек w и z соответственно. В разделе А.10 будет показано, что это угол θ в полярном представлении z/w:

$$z/w = re^{i\theta} = r\,(\cos\theta + i\sin\theta)$$

(см. рис. А.42), но z/w теперь заменяет z из той схемы. Именно θ управляет разностью фаз между состояниями **α** и **β**, а также тем, смогут ли они — вернее, *где* смогут — погасить или усилить друг друга. Обратите внимание, что угол θ на плоскости Весселя отсчитывается против часовой стрелки (рис. 2.9). Еще в выражении w/z остается информация об отношении расстояний от 0 до w и от 0 до z, то есть r в вышеупомянутом полярном представлении z/w, причем это отношение определяет относительную интенсивность двух компонентов **α** и **β** в суперпозиции. В разделе 2.8 мы убедимся, что такие отношения интенсивностей (точнее, их квадрат) играют важную роль при работе с вероятностями, когда измерение **R** производится в квантовой системе. Строго говоря, вероятностная интерпретация применима лишь в случае, когда состояния **α** и **β** ортогональны друг другу, но эту концепцию мы также обсудим в разделе 2.8.

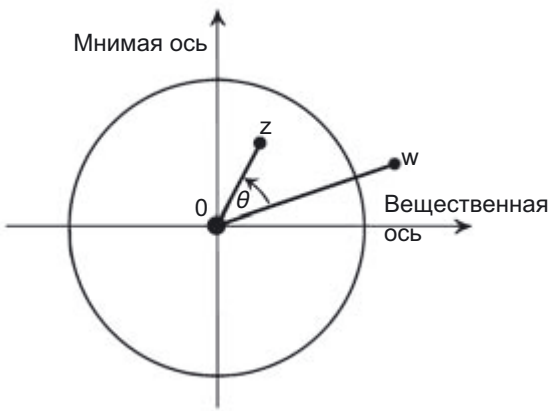


Рис. 2.9. На (комплексной) плоскости Весселя аргумент отношения z/w между двумя ненулевыми комплексными числами w и z — это угол θ между двумя лучами, выходящими из начала координат к этим точкам

Обратите внимание: обе интересующие нас величины из предыдущего абзаца, касающиеся весовых коэффициентов w и z (разность фаз и относительная интенсивность), связаны именно с их *отношением* w/z. Чтобы стало понятнее, почему такие отношения обладают особым значением, необходимо поговорить об

одной важной черте квантового формализма. Просто не все комбинации $w\boldsymbol{\alpha} + z\boldsymbol{\beta}$ можно считать различными с физической точки зрения; физическая сущность подобной комбинации в реальности зависит лишь от отношения амплитуд w/z. Дело в общем принципе, применимом к математическому описанию — *вектору состояния* (например, $\boldsymbol{\alpha}$) — любой квантовой системы, а не только корпускулярно-волновой частицы. Считается, что квантовое состояние системы остается *физически неизменным*, если умножить вектор состояния на любое ненулевое комплексное число, иными словами, вектор состояния $u\boldsymbol{\alpha}$ представляет собой *то же самое* корпускулярно-волновое состояние (или общее квантовое состояние), что и $\boldsymbol{\alpha}$, для любого ненулевого комплексного числа u. Соответственно, то же касается и нашей комбинации $w\boldsymbol{\alpha} + z\boldsymbol{\beta}$.

Любое произведение этого вектора состояния на ненулевое комплексное число u

$$u(w\boldsymbol{\alpha} + z\boldsymbol{\beta}) = uw\boldsymbol{\alpha} + uz\boldsymbol{\beta}$$

описывает *ту же самую* физическую сущность, что и $w\boldsymbol{\alpha} + z\boldsymbol{\beta}$. Отметим, что отношение $(uz)/(uw)$ равно z/w, именно поэтому нас на самом деле и интересует именно отношение z/w.

До сих пор мы рассматривали суперпозиции, которые могут образовываться всего двумя состояниями — $\boldsymbol{\alpha}$ и $\boldsymbol{\beta}$. Если состояний будет три — $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ и $\boldsymbol{\gamma}$, то из них можно образовать суперпозции

$$w\boldsymbol{\alpha} + z\boldsymbol{\beta} + v\boldsymbol{\gamma},$$

где амплитуды w, z и v являются комплексными числами (среди которых есть хотя бы одно ненулевое), причем мы считаем, что физическое состояние не изменится, если умножить все члены на ненулевое комплексное число u (чтобы получить $uw\boldsymbol{\alpha} + uz\boldsymbol{\beta} + uv\boldsymbol{\gamma}$). Вновь возникает ситуация, в которой лишь набор отношений $w/z/v$ физически отличает эти суперпозиции друг от друга. Такое множество дополняется до любого конечного числа состояний $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, ..., $\boldsymbol{\phi}$, что позволяет получить суперпозиции $v\boldsymbol{\alpha} + w\boldsymbol{\beta} + \ldots + z\boldsymbol{\phi}$, причем физически эти состояния будут различаться в зависимости от $v/w/\ldots/z$.

В действительности нужно быть готовым и к тому, что число таких суперпозиций может возрастать до *бесконечного множества* отдельных состояний. Кроме того, нужно обращать внимание и на такие вещи, как непрерывность, сходимость и т. д. (см. раздел А.10). С этими вопросами связаны серьезные математические сложности, которыми я не хочу напрасно утомлять читателя. Хотя некоторые специалисты по математической физике могут с полным правом счесть, что несомненные сложности, с которыми сталкивается квантовая теория (точнее, квантовая теория поля; см. разделы 1.4 и 1.6), требуют пристального внимания

именно к таким математическим тонкостям, здесь я предлагаю отнестись ним довольно небрежно. Нет, я не считаю такие нюансы маловажными — ничего подобного, математическая согласованность кажется мне приоритетным требованием, скорее дело в том, что кажущаяся несогласованность квантовой теории, о которой мы еще поговорим (в частности, в разделах 2.12 и 2.13), на мой взгляд, имеет более фундаментальный характер, слабо связанный с вопросами математической строгости как таковой.

Продолжая придерживаться такой «математически вольной» точки зрения, давайте попытаемся рассмотреть общее состояние одиночной точечной частицы, называемой *скалярной* частицей (не имеющей характеристик, связанных с направлением в пространстве, а значит, и со спином, не имеющим направления, то есть равным нулю). Простейшее базовое состояние — *координатное* представление — будет характеризоваться лишь некоторым пространственным расположением A, которое задается радиус-вектором **a** относительно некоторого известного начала координат O (см. разделы A.3 и A.4). Это очень идеализированное состояние, и принято считать, что оно задается (с точностью до постоянного коэффициента) выражением $\delta(\mathbf{x} - \mathbf{a})$, где δ — дельта-функция Дирака, о которой мы вскоре поговорим. Кроме того, это не самое «внятное» состояние для реальной физической частицы, поскольку шрёдингеровская эволюция приведет к тому, что это состояние немедленно начнет «размазываться» — этот эффект можно считать одним из следствий *принципа неопределенности Гейзенберга* (который мы коротко обсудим в конце раздела 2.13): если абсолютно точно установить положение частицы, то ничего не будет известно о ее импульсе, поэтому неопределенный импульс моментально «размазывает» состояние частицы. Однако на данном этапе я не стану останавливаться на том, как такое состояние может развиваться в будущем. Будет уместно рассуждать о ситуации всего лишь в контексте того, как квантовые состояния могут проявлять себя в конкретный момент времени, скажем $t = t_0$.

На самом деле дельта-функция — это не обычная функция, а предельный случай функции, которая стремится к нулю везде, кроме точки $x = 0$, а в точке $x = 0$ устремляется в бесконечность, при этом площадь под кривой на графике такой функции равна единице. Соответственно, $\delta(x - a)$ будет вести себя аналогично, с той только разницей, что вместо точки $x = 0$ она будет устремляться в бесконечность в точке $x = a$, а во всех прочих точках будет равна нулю. Составить впечатление об этом предельном случае поможет рис. 2.10. (Чтобы лучше понять математические аспекты, см., например, [Lighthill, 1958] и [Stein and Shakarchi, 2003][1].) Мне не хотелось бы углубляться в дебри математической строгости,

[1] Или прекрасный русскоязычный учебник Я. Б. Зельдовича «Высшая математика для начинающих и ее приложения к физике». — *Примеч. ред.*

здесь я намерен лишь донести до читателя основные идеи и показать, как всем этим пользоваться. (Действительно, технически такие состояния в любом случае не входят в состав стандартного квантово-механического формализма гильбертова пространства, о котором мы поговорим в разделе 2.8; идеальные пространственные состояния оказываются физически нереализуемыми. Тем не менее они очень удобны в плане дискурса.)

Исходя из вышесказанного, можно рассмотреть дельта-функцию *трехмерного вектора* **x**; при этом можно записать следующее: $\delta(\mathbf{x}) = \delta(x_1)\,\delta(x_2)\,\delta(x_3)$, где x_1, x_2, x_3 — это три декартовых компонента **x**. Тогда $\delta(\mathbf{x}) = 0$ для всех *ненулевых*

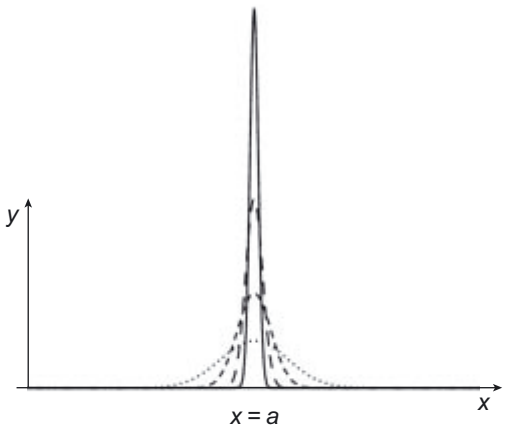


Рис. 2.10. Дельта-функция Дирака $\delta(x)$ проиллюстрирована здесь так: мы считаем $\delta(x - a)$ пределом последовательности гладких положительных функций, причем площадь под кривой каждой такой функции равна единице и большая часть этой площади приходится на область вблизи $x = a$

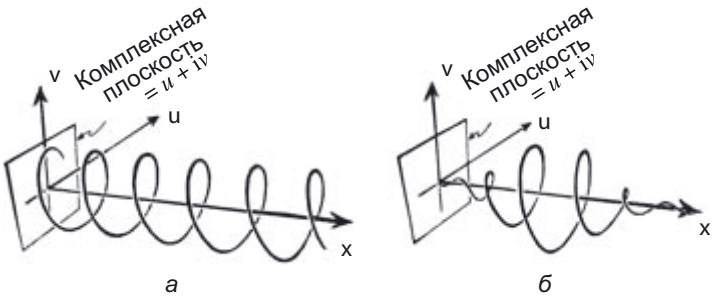


Рис. 2.11. Волновая функция импульсного состояния $e^{-i\mathbf{p}\cdot\mathbf{x}}$ для заданного трехмерного импульса **p** (*а*) и волновая функция волнового пакета (*б*)

значений трехмерного вектора **x**, но мы опять же должны считать $\delta(\mathbf{0})$ крайне большим, порождающим *единичный* 3-объем. Тогда можно представить $\delta(\mathbf{x}-\mathbf{a})$ в виде функции, дающей нулевую амплитуду для частицы, находящейся где угодно, *кроме* точки X, с радиус-вектором **x** (то есть $\delta(\mathbf{x}-\mathbf{a})$ не равна нулю, лишь если $\mathbf{x}=\mathbf{a}$) и присваивающей очень большую (бесконечную) амплитуду частице, находящейся *как раз* в точке A (то есть когда $\mathbf{x}=\mathbf{a}$). Далее можно рассмотреть *непрерывные* суперпозиции таких частных состояний, причем в такой суперпозиции комплексная амплитуда присваивается каждой пространственной точке X. Следовательно, эта комплексная амплитуда (а не просто обычное комплексное число) является некоторой комплексной функцией позиции 3-вектора **x**. Эта функция часто обозначается греческой буквой ψ (пси), а функция $\psi(\mathbf{x})$ называется (шрёдингеровской) *волновой функцией* частицы.

Следовательно, комплексное число $\psi(\mathbf{x})$, присваиваемое функцией ψ каждой отдельной пространственной точке X, — это амплитуда, которой должна обладать частица, чтобы она оказалась именно в точке X. При этом ситуация останется неизменной с физической точки зрения, если умножить амплитуду в каждой точке на одно и то же ненулевое комплексное число u. Таким образом, комплексная функция $w\psi(\mathbf{x})$ описывает ровно ту же физическую ситуацию, что и $\psi(\mathbf{x})$, если w — любое ненулевое (постоянное) комплексное число.

Важным примером волновой функции могла бы послужить плоская волна, имеющая определенные частоту и направление. Такое состояние, именуемое *импульсным* представлением, описывалось бы выражением $\psi e^{-i\Psi\mathbf{x}}$, где **p** — (постоянный) 3-вектор, описывающий импульс частицы; см. рис. 2.11 *а*, где вертикальная плоскость (u, v) означает плоскость Весселя $\psi = u + iv$. Что касается фотонов, их импульсные состояния будут нас интересовать в разделах 2.6 и 2.13. На рис. 2.11 *б* показан волновой пакет — о том, что это такое, рассказано в разделе 2.3.

На данном этапе было бы полезно отметить, что в квантовой механике принято нормировать описание квантового состояния в терминах «размера», который может быть присвоен волновой функции ψ. Эта величина — положительное вещественное число, именуемое *нормой* функции, так что[1] можно записать (см. раздел A.3):

$$\|\psi\|,$$

[1] В рассматриваемом здесь случае скалярной волновой функции это будет интеграл (см. раздел A.11) вида $\int \psi(\mathbf{x})\bar{\psi}(\mathbf{x})d^3\mathbf{x}$, охватывающий все трехмерное пространство. Следовательно, для нормированной волновой функции интерпретация $|\psi(\mathbf{x})|^2$ как плотности вероятностей соответствует тому, что полная вероятность найти частицу во всем пространстве равна единице.

где $\|\psi\| = 0$ тогда и только тогда, когда ψ является нулевой функцией, а не полноценной волновой функцией, и норма обладает следующим свойством калибровки:

$$\|w\psi\| = |w|^2\|\psi\|$$

для любого комплексного числа w ($|w|$ — его модуль; см. раздел A.10). Нормированная волновая функция имеет *единичную* норму:

$$\|\psi\| = 1,$$

а если ψ исходно не была нормирована, это всегда можно сделать, заменив ψ на $u\psi$, где $u = \|\psi\|^{-1/2}$. Нормировка ограничивает произвол при замене $\psi \mapsto \omega\psi$ и в случае со скалярной волновой функцией в этом заключается одно преимущество: в таком случае можно считать, что квадрат модуля волновой функции $\psi(\mathbf{x})$ дает *плотность вероятности* того, что мы сможем найти частицу в точке X.

Однако нормировка не убирает всю калибровочную свободу, поскольку любое умножение ψ на *чистую фазу*, то есть комплексное число с единичным модулем $e^{i\theta}$ (где θ — вещественное и постоянное)

$$\psi \mapsto e^{i\theta}\psi,$$

не влияет на нормировку (в принципе, это и есть та фазовая свобода, которую Вейль в итоге стал рассматривать в рамках своей теории электромагнетизма, упоминаемой в разделе 1.8). Хотя у подлинных квантовых состояний всегда есть норма и, следовательно, они поддаются перенормировке, это не так с определенными широко используемыми идеализированными квантовыми состояниями, например с рассмотренными выше координатными состояниями $\delta(\mathbf{x} - \mathbf{a})$ или с рассмотренными только что импульсными состояниями $e^{-i\mathbf{p}\cdot\mathbf{x}}$. Мы вернемся к разговору об импульсных состояниях в разделах 2.6 и 2.13, когда речь пойдет о фотонах. Пока я игнорирую эту проблему — в этом отчасти и заключается мое небрежное отношение к математике, в чем я признавался выше. В разделе 2.8 мы увидим, как такое представление о *норме* вписывается в более широкий аппарат квантовой механики.

2.6. Волновая функция фотона

Комплексная волновая функция ψ дает нам шрёдингеровскую картину единичного скалярного корпускулярно-волнового состояния. Но это до сих пор просто частица без структуры, то есть без какой-либо характеристики, связанной с направлением (спин 0). Правда, мы можем составить достаточно хорошее впечат-

ление о том, что происходит с волновой функцией отдельно взятого фотона. Это не скалярная частица, поскольку значение спина фотона равно $\hbar$, то есть спин равен единице в нормальных единицах Дирака (см. раздел 1.14). Далее, волновая функция имеет векторный характер, и оказывается, что ее можно представить как *электромагнитную волну*, причем если интенсивность волны чрезвычайно низкая, то эта волна соответствовала бы единичному фотону. Поскольку волновая функция — комплексная, ее можно описать формулой

$$\psi = \mathbf{E} + i\mathbf{B},$$

где $\mathbf{E}$ — 3-вектор электрического поля (ср. с разделами A.2 и A.3), а $\mathbf{B}$ — 3-вектор магнитного поля. (Строго говоря, чтобы получить настоящую волновую функцию свободного фотона, нужно взять так называемую *положительно-частотную часть* от $\mathbf{E} + i\mathbf{B}$, а эта задача уже связана с разложением в ряд Фурье, о чем говорится в разделе A.11, и сложить ее с положительно-частотной частью $\mathbf{E} - i\mathbf{B}$ [ПкР, раздел 24.3], но такие вопросы подробнее рассмотрены в Streater, Wightman [2000]. Однако подобные технические проблемы нас здесь мало интересуют, поскольку дальнейший материал с ними практически не связан.)

Ключевой аспект такого электромагнитно-волнового представления заключается в том, что мы должны учитывать поляризацию волны. Термин *поляризация* необходимо объяснить подробнее. Электромагнитная волна, движущаяся в определенном направлении в пространстве и имеющая конкретную частоту и интенсивность, то есть *монохроматическая* волна, может быть *плоскополяризованной*. В таком случае у нее будет так называемая *плоскость поляризации*, то есть плоскость, в которой располагается направление движения волны, причем в этой же плоскости происходят колебания электрического поля волны (рис. 2.12 *а*). Колеблющемуся электрическому полю сопутствует магнитное поле, которое также колеблется с той же частотой и в той же фазе, что и электрическое, но в плоскости, перпендикулярной плоскости поляризации, правда, тоже содержащей направление движения. (Можно считать рис. 2.12 *а* иллюстрацией поведения волны во времени, причем стрелка указывает на отрицательное направление времени.) В любой плоскости у нас могут быть монохроматические плоскополяризованные электромагнитные волны, причем направление движения волны будет лежать в ее плоскости поляризации. Более того, *любую* монохроматическую электромагнитную волну, для которой известно некоторое направление движения $\mathbf{k}$, можно разложить на сумму плоскополяризованных волн, чьи плоскости поляризации будут перпендикулярны друг другу. У солнечных очков Polaroid есть одна особенность: они пропускают компонент световой волны, где вектор $\mathbf{E}$ расположен в вертикальном направлении, а компонент, где вектор $\mathbf{E}$ расположен в горизонтальном направлении, поглощают. Свет от неба и свет, отраженный от поверхности воды, сильно поляризуется в горизон-

тальном направлении, поэтому если надеть именно такие очки, в глаза попадет значительно меньше отраженного от воды или от неба света.

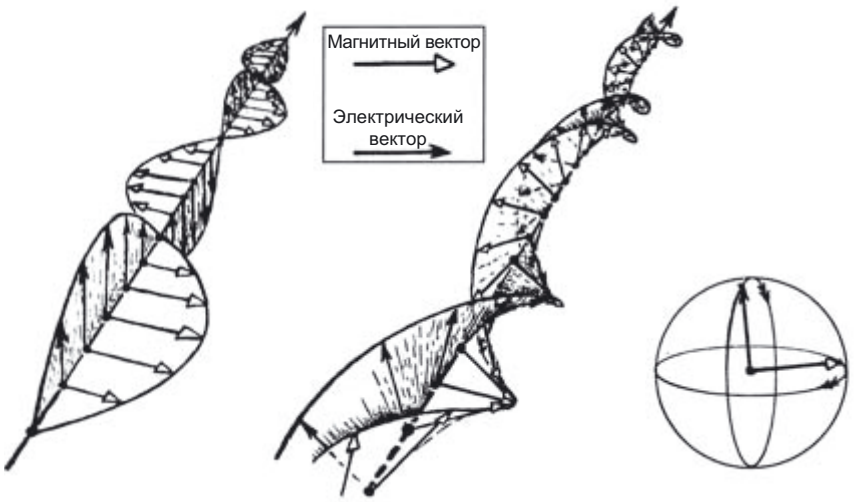


Рис. 2.12. Иллюстрация колебаний векторов электрического и магнитного поля в плоскополяризованной волне; движение можно рассматривать в пространстве или во времени (*а*); иллюстрация электромагнитной волны с круговой поляризацией (*б*); совместив линейную и круговую поляризацию, можно получить различную степень эллиптической поляризации (*в*)

Те поляризующие очки, которые сегодня часто используются для просмотра стереокино, устроены немного иначе. Чтобы понять, как именно, давайте обсудим еще один термин, а именно *круговую* поляризацию (см. рис. 2.12 *б*). При такой поляризации вектор электрического поля вращается по или против часовой стрелки (обычно говорят о правой или левой поляризации), и такая ориентация называется *спиральностью* волны с круговой поляризацией[1]. Такие очки хитроумно устроены: их линзы изготовлены из полупрозрачного материала, причем одна линза пропускает свет только с правой поляризацией, а другая — только с левой. Самое интересное, что в обоих случаях свет выходит из стекла очков в плоскополяризованном состоянии.

[1] По-видимому, в физике частиц и квантовой оптике действуют противоположные соглашения о знаках спиральности. Эта проблема рассмотрена в работе [Jackson, 1999, p. 206].

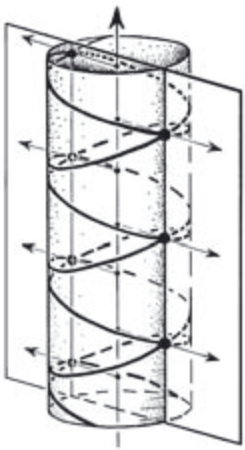


Рис. 2.13. Сложив равное число лево- и правополяризованных волн, можно получить плоскополяризованную волну, гребни и впадины которой обозначены горизонтальными стрелками. Меняя фазовые отношения между компонентами с круговой поляризацией, мы вращаем плоскость поляризации

Наряду с такими поляризованными состояниями бывает еще и эллиптическая поляризация (см. рис. 12.2 *в*), при которой круговая поляризация совмещается с плоскостной поляризацией в некотором направлении. Все эти поляризованные состояния могут возникать в результате комбинации всего *двух* базовых состояний. Речь может идти о двух волнах, одна из которых имеет левую поляризацию, а другая — правую, либо это могут быть волны, поляризованные в горизонтальной и вертикальной плоскостях, возможны и другие самые разные пары. Давайте рассмотрим, как все это работает.

Возможно, проще всего будет исследовать случай суперпозиции двух циркулярно-поляризованных волн разной ориентации, имеющих одинаковую интенсивность. Разнообразные фазовые отличия просто дадут всевозможные направления линейной поляризации. Чтобы разобраться в этом явлении, будем считать каждую волну *геликоидом*, то есть кривой, огибающей цилиндр и образующей с его осью некий фиксированный угол, но не 0° и не 90° (рис. 2.13). Эта спираль соответствует линии движения конца вращающегося электрического вектора при его перемещении вдоль оси (ось соответствует направлению волны). Каждая из двух наших волн представлена в виде левоориентированной или правоориентированной спирали (с равным шагом). Если начертить их обе на одном и том же цилиндре, то окажется, что все точки пересечения этих линий лежат в одной плоскости, которая и будет плоскостью поляризации суперпозиции двух волн, соответствующих этим линиям. Если поворачивать одну из спиралей вокруг

оси цилиндра, а другую оставить фиксированной, тем самым получая разные фазовые отличия, то можно получить все возможные плоскости поляризации суммарной волны. Если теперь мы увеличим вклад одной из волн в эту суперпозицию относительно вклада другой волны, то получим всевозможные состояния эллиптической поляризации.

Наше комплексное представление электрического и магнитного векторов в вышеприведенной форме (**E** + i**B**) позволяет непосредственно оценить, на что похожа волновая функция фотона. Давайте исходя из ситуации, описанной в прошлом абзаце, обозначим греческой буквой **α** волну с правой поляризацией, а буквой **β** — волну с левой поляризацией (интенсивности и частоты этих волн полагаются равными). Тогда можно будет представить различные состояния плоскостной поляризации, получаемые при помощи *равно*интенсивных суперпозиций двух волн (с точностью до коэффициента пропорциональности), в следующем виде:

$$\boldsymbol{\alpha} + z\boldsymbol{\beta}, \text{ где } z = e^{i\theta}.$$

При увеличении θ с 0 до 2π так, чтобы точка $z = e^{i\theta}$ (= cosθ +sinθ) сдвинулась вдоль единичной окружности на плоскости Весселя (см. раздел A.10 и 1.8, а также рис. 2.9), плоскость поляризации также поворачивается в том же направлении. В данном случае важно сравнить скорость вращения этой плоскости со скоростью движения *z*, поскольку положение *z* изменяется вдоль единичной окружности.

Вектору состояния $z\boldsymbol{\beta}$ соответствует левоориентированная спираль на рис. 2.13. После полного оборота *z* можно считать, что эта спираль непрерывно повернулась вокруг своей оси на 2π против часовой стрелки. Из рис. 2.13 видно, что такое вращение эквивалентно поступательному переносу левоориентированной спирали вверх, причем правоориентированная спираль остается на месте, пока левоориентированная спираль не совпадет с исходной конфигурацией. Мы видим, что при этом точки пересечения двух спиралей, изначально находившиеся на переднем плане, уходят на задний план, а плоскость поляризации возвращается в исходное положение. Соответственно, когда *z* один раз повернется вдоль единичной окружности на угол 2π (то есть на 360°), плоскость поляризации вернется в исходное положение, повернувшись всего на π, а не на 2π (то есть на 180°, а не на 360°). Если обозначить через φ угол, на который плоскость поляризации повернется за один любой шаг, то получим, что θ = 2φ, где, как и выше, θ — это угол на плоскости Весселя, который амплитуда *z* образует с действительной осью (в этом вопросе есть некоторые условные соглашения, но в данном случае я считаю, что мы рассматриваем нашу плоскость Весселя сверху; см. рис. 2.13).

В контексте точек на плоскости Весселя можно использовать комплексное число q (или число противоположного знака $-q$), чтобы показать угол, который эта плоскость образует с плоскостью поляризации, где $q = e^{i\varphi}$, так что тождество $\theta = 2\varphi$ приобретает вид:

$$z = q^2.$$

Далее мы увидим (рис. 2.20 в разделе 2.9), как обобщить q для описания состояния эллиптической поляризации.

Однако следует пояснить, что только что рассмотренные состояния фотонов — это очень частный случай так называемых *импульсных* состояний, когда фотоны переносят энергию в совершенно определенном направлении. Благодаря универсальности принципа суперпозиции понятно, что отдельные фотоны могут находиться и во множестве других состояний. Например, можно было бы рассмотреть просто суперпозицию двух таких импульсных состояний, где импульсы являются разнонаправленными. Опять же, в таких ситуациях мы получаем электромагнитные волны, являющиеся решениями максвелловских уравнений (поскольку в линейном виде максвелловские уравнения точно такие же, как и шрёдингеровские; см. разделы 2.4, A.11 и 2.7). Более того, если скомбинировать множество таких волн, направления движения которых различаются лишь незначительно, а частоты очень высокие и несильно отличающиеся друг от друга, то можно получить решения уравнений Максвелла, которые будут описывать волны, локализованные в ограниченной пространственной области, распространяющиеся со скоростью света в определенном направлении. Такие решения называются *волновыми пакетами* — они упоминались в разделе 2.3 как возможные кандидаты на роль корпускулярно-волновой сущности, необходимой для объяснения результата эксперимента 2, приведенного на рис. 2.4 *б*. Однако, как мы убедились в разделе 2.3, такое классическое описание отдельного фотона не объясняет результатов эксперимента 1, приведенного на рис. 2.4 *а*. Более того, такие волновые пакеты (классические решения уравнений Максвелла) не будут все время оставаться частицеподобными и очень быстро рассосутся. Эту ситуацию можно сравнить со свойствами отдельных фотонов, проделывающих очень большой путь, например прилетающих из далеких галактик.

Корпускулярные свойства фотонов связаны не с локализацией их волновых функций — нет, они связаны с тем, что производимое над фотонами измерение специально «заточено» под наблюдение корпускулярных свойств этих частиц, как, например, в эксперименте с фотопластинкой или (если речь идет о заряженных частицах) со счетчиком Гейгера. Наблюдаемая в результате *частицеподобная* природа такого квантового объекта — это свойство операции **R**, когда детектор

реагирует именно на частицы. Так, волновая функция фотона, прилетевшего из далекой галактики, распределилась бы в огромнейшей области пространства, и если зафиксировать ее в определенной точке на фотопластинке, то остается исчезающе малая вероятность того, что именно при данном конкретном измерении мы обнаружим ее в **R**-процессе. Мы вообще вряд ли увидели бы такой фотон, если бы наблюдаемая область этой далекой галактики не излучала фотоны в столь неимоверных количествах, компенсируя таким образом крайне малую вероятность засечь любой конкретный фотон!

Рассмотренный выше пример с поляризованным фотоном также иллюстрирует использование комплексных чисел для описания формирования линейных суперпозиций классических полей. На самом деле существует очень тесная связь между вышеупомянутой процедурой комплексной суперпозиции классических (электромагнитных) полей и квантовой суперпозицией состояний частиц, в данном случае фотонов. Фактически уравнение Шрёдингера для отдельного свободного фотона оказывается просто вариантом максвелловских уравнений свободного поля, только в этом случае напряженность электромагнитного поля *выражается комплексным числом*.

Здесь следует подчеркнуть одно отличие, а именно: если умножить описание состояния отдельного фотона на ненулевое комплексное число, то состояние от этого не изменится. При этом окажется, что в случае *классического* электромагнитного поля интенсивность (то есть плотность энергии) этого состояния будет пропорциональна квадрату напряженности поля. В то же время, чтобы увеличить энергосодержание *квантового* состояния, нам бы потребовалось увеличить *количество* фотонов, так как энергия каждого отдельного фотона регламентирована формулой Планка $E = h\nu$. Следовательно, в квантово-механической ситуации именно количество фотонов пропорционально квадрату модуля комплексного числа, если рассматривать волновую функцию как напряженность электромагнитного поля. В разделе 2.8 мы увидим, как это связано с законом вероятности для **R**, также называемым *правилом Борна* (см. также разделы 1.4 и 2.8).

2.7. Квантовая линейность

Однако прежде чем перейти к этому, нужно хотя бы вскользь рассмотреть ошеломительно необъятную тему квантовой линейности. Важная черта формализма квантовой механики заключается в *линейности* **U**. Как указано в разделе A.11, это особое свойство, упрощающее **U**, *не свойственно* большинству классических типов эволюции, хотя уравнения Максвелла на самом деле тоже линейны. Попытаемся понять, что означает эта линейность.

Как я отмечал ранее (в начале раздела 2.5), понятие *линейности* применительно к процессу эволюции во времени, например **U**, подразумевает аддитивность состояний системы или, точнее, возможность образовать из них *линейную комбинацию*. Если эволюция во времени сохраняет такое свойство, то говорят, что она линейна. В квантовой механике таков принцип *суперпозиции* квантовых состояний в следующем случае: если и $\boldsymbol{\alpha}$ и $\boldsymbol{\beta}$ являются допустимыми состояниями системы, то линейная комбинация

$$w\boldsymbol{\alpha} + z\boldsymbol{\beta},$$

где хотя бы одно из фиксированных комплексных чисел w и z не равно нулю, также является допустимым квантовым состоянием. Свойство *линейности*, присущее **U**, заключается всего лишь в следующем: если некоторое квантовое состояние $\boldsymbol{\alpha}_0$ развивается согласно U в состояние $\boldsymbol{\alpha}_t$ за конкретный промежуток времени t:

$$\boldsymbol{\alpha}_0 \rightsquigarrow \boldsymbol{\alpha}_t,$$

и если бы другое состояние $\boldsymbol{\beta}_0$ развилось в $\boldsymbol{\beta}_t$ за тот же промежуток времени t:

$$\boldsymbol{\beta}_0 \rightsquigarrow \boldsymbol{\beta}_t,$$

то любая суперпозиция $w\boldsymbol{\alpha}_0 + z\boldsymbol{\beta}_0$ спустя такой промежуток времени t развилась бы в $w\boldsymbol{\alpha}_t + z\boldsymbol{\beta}_t$:

$$w\boldsymbol{\alpha}_0 + z\boldsymbol{\beta}_0 \rightsquigarrow w\boldsymbol{\alpha}_t + z\boldsymbol{\beta}_t,$$

причем комплексные числа w и z при этом не изменялись бы. В сущности, это и есть линейность (см. также раздел А.11). Суть линейности можно кратко выразить в виде афоризма:

эволюция суммы есть сумма эволюций,

однако в данном случае «сумму» на самом деле нужно понимать как всеобъемлющую линейную комбинацию.

До сих пор, как понятно из разделов 2.5 и 2.6, мы рассматривали линейные суперпозиции состояний лишь для *одиночных* корпускулярно-волновых сущностей, но линейность Шрёдингера абсолютно универсально применяется к квантовым системам, независимо от того, сколько частиц может одновременно участвовать в процессе. Например, у нас может быть состояние, в котором участвуют две (разные) скалярные частицы, причем одна из них считается волной-частицей, локализованной на одном компактном участке около пространственной точки P, а другая локализована в совершенно другой пространственной точке Q. Наше

состояние **α** может соответствовать такой *паре* локализаций частиц. Во втором состоянии **β** первая частица может быть локализована в совершенно другой точке P′, а вторая — в другой точке Q′. Все четыре локализации никак не связаны друг с другом. Теперь допустим, что мы рассматриваем суперпозицию вида **α**+ **β**. Как ее интерпретировать? В первую очередь, необходимо уяснить, что интерпретация может нисколько не походить на поиск «среднего» между рассматриваемыми локализациями (то есть нельзя предположить, что первая частица окажется где-то на полпути от P до P′, а вторая — между Q и Q′). Подобное положение вещей весьма далеко от того, что говорит нам квантово-механическая линейность, — достаточно вспомнить, что даже к *отдельной* волне-частице такая интерпретация неприменима (то есть линейная суперпозиция одного состояния, когда частица находится в точке P, и другого состояния, когда она же расположена в точке P′, — это состояние, в котором частица одновременно находится сразу в двух этих точках, что определенно не эквивалентно местонахождению частицы в какой-то третьей точке между ними). Нет, суперпозиция **α** + **β** включает *все* четыре локализации (P, P′, Q, Q′) именно там, где они расположены, причем альтернативные пары точек (P, Q) и (P′, Q′) для двух частиц как бы *сосуществуют*! Как мы узнаем, существуют любопытные тонкости, связанные с возникающим здесь состоянием вида **α** + **β**, именуемые *запутанностью*, когда ни одна из частиц не обладает собственным состоянием, которое не зависело бы от состояния другой частицы.

Концепцию *запутанности* впервые предложил Шрёдингер в письме к Эйнштейну, где обозначил этот феномен немецким словом *Verschränkung*[1], а впоследствии сам перевел его на английский как *entanglement* [Schrödinger and Born, 1935]. Шрёдингер так прокомментировал квантовую странность и важность этого феномена:

> *Я бы назвал [запутанность] не просто отдельной, а самой характерной чертой квантовой механики, такой чертой, которая требует полностью отказаться от классического мышления.*

Далее, в разделе 2.10, мы рассмотрим некоторые очень странные и подлинно квантовые свойства, которые могут проявляться в таком состоянии — все в совокупности они называются эффектами Эйнштейна — Подольского — Розена (ЭПР) [см. Einstein et al., 1935]; именно благодаря им Шрёдингер и осознал концепцию запутанности. Сегодня считается, что запутанность обнаруживается на уровне таких явлений, как нарушение неравенств Белла, о чем пойдет речь в разделе 2.10.

---

[1] В переводе с немецкого — «ограничение». — *Примеч. пер.*

Для того чтобы разобраться в квантовой запутанности, было бы полезно вернуться к нотации дельта-функции, кратко упомянутой в разделе 2.5. Возьмем радиус-векторы $\mathbf{p}$, $\mathbf{q}$, $\mathbf{p}'$, $\mathbf{q}'$ для точек P, Q, P′, Q′ соответственно. Для простоты я не буду вдаваться ни в какие комплексные амплитуды, а также не стану заниматься нормировкой каких-либо наших состояний. Давайте для начала рассмотрим единственную частицу. Можно записать состояние этой частицы в точке P как $\delta(\mathbf{x} - \mathbf{p})$, а в Q как $\delta(\mathbf{x} - \mathbf{q})$. Сумма двух этих состояний будет равна:

$$\delta(\mathbf{x} - \mathbf{p}) + \delta(\mathbf{x} - \mathbf{q}),$$

что соответствует суперпозиции частицы, находящейся в двух этих местах одновременно (и принципиально отличается от ситуации $\delta(\mathbf{x} - \frac{1}{2}(\mathbf{p} + \mathbf{q}))$, которая бы описывала положение частицы на полпути между двумя этими точками (рис. 2.14 *a*). (Следует напомнить читателю о замечании из раздела 2.5, что подобная идеализированная волновая функция может иметь форму дельта-функции лишь в исходный момент, допустим в $t = t_0$. Шрёдингеровская эволюция требует, чтобы уже в следующий миг такие состояния размазались. Правда, в данном случае этот момент для нас неважен.) Когда требуется представить квантовое состояние двух частиц, одна из которых находится в точке P, а другая — в точке Q, то можно попытаться описать ситуацию следующим образом: $\delta(\mathbf{x} - \mathbf{p})\ \delta(\mathbf{x} - \mathbf{q})$, но это было бы неверно по целому ряду причин. Основное возражение таково: если просто взять произведения волновых функций, то есть взять две волновые функции частиц, $\psi(\mathbf{x})$ и $\phi(\mathbf{x})$, а затем перемножить их: $\psi(\mathbf{x})\ \phi(\mathbf{x})$, то это было бы принципиально неверно, так как нарушало бы линейность шрёдингеровской эволюции. Однако верный ответ не так уж сильно отличается от этого варианта (пока предположим, что речь идет о двух неоднотипных частицах, то есть нам не приходится учитывать проблемы, упомянутые в разделе 1.4 и связанные со свойствами фермионов и бозонов). Можно, как и раньше, описать положение первой частицы при помощи радиус-вектора $\mathbf{x}$, а для другой частицы воспользоваться вторым радиус-вектором $\mathbf{y}$. Тогда волновая функция $\psi(\mathbf{x})\ \phi(\mathbf{y})$ действительно описывала бы подлинное состояние — такое, где пространственное распределение амплитуды первой частицы задается $\psi(\mathbf{x})$, а второй — $\phi(\mathbf{y})$ и они полностью *независимы* друг от друга. Когда частицы *зависят* друг от друга, они называются *запутанными*, и тогда их состояние не соответствует простому произведению $\psi(\mathbf{x})\ \phi(\mathbf{y})$, а описывается более общей функцией $\Psi(\mathbf{x}, \mathbf{y})$ с двумя переменными радиус-вектора, $\mathbf{x}$ и $\mathbf{y}$. Этот принцип можно обобщить на случай многих частиц с радиус-векторами $\mathbf{x}$, $\mathbf{y}$, ..., $\mathbf{z}$, причем их общее (запутанное) квантовое состояние описывалось бы выражением $\Psi(\mathbf{x}, \mathbf{y}, ..., \mathbf{z})$, а полностью незапутанное состояние представляло бы собой частный случай: $\psi(\mathbf{x})\phi(\mathbf{y}) ... \chi(\mathbf{z})$.

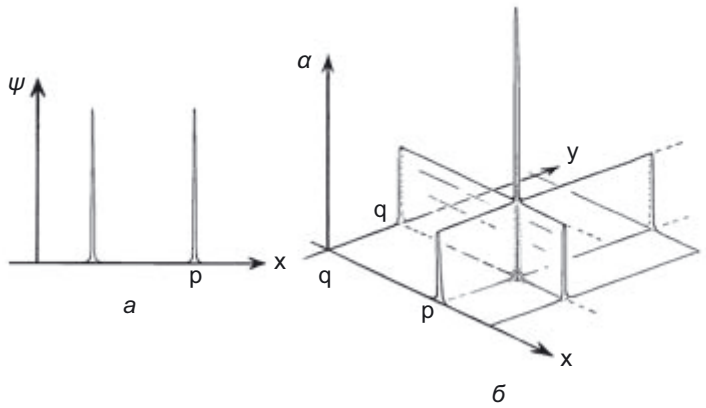


Рис. 2.14. *а*) Сумма дельта-функций $\psi(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{p}) + \delta(\mathbf{x} - \mathbf{q})$ дает волновую функцию скалярной частицы, находящейся в состоянии суперпозиции одновременно в точках P и Q; *б*) произведение двух дельта-функций $\alpha(\mathbf{x}, \mathbf{y}) = \delta(\mathbf{x} - \mathbf{p})\,\delta(\mathbf{y} - \mathbf{q})$ дает состояние, в котором участвуют две разные частицы, одна из которых находится в P, а другая — в Q. Обратите внимание: нужны две переменные (**x** и **y**), каждая из которых означает положение отдельной частицы

Вернемся к вышеприведенному примеру, где изначально рассматривалось такое состояние: первая частица находится в точке P, а вторая — в точке Q. В таком случае волновая функция будет незапутанной:

$$\boldsymbol{\alpha} = \delta(\mathbf{x} - \mathbf{p})\,\delta(\mathbf{y} - \mathbf{q})$$

(рис. 2.14 *б*). В нашем примере она будет в суперпозиции с $\boldsymbol{\beta} = \delta(\mathbf{x} - \mathbf{p}')\,\delta(\mathbf{y} - \mathbf{q}')$, где первая частица находится в P′, а вторая — в Q′. Тогда (если не учитывать амплитуду) получаем:

$$\alpha + \beta = \delta(\mathbf{x} - \mathbf{p})\delta(\mathbf{y} - \mathbf{q}) + \delta(\mathbf{x} - \mathbf{p}')\delta(\mathbf{y} - \mathbf{q}').$$

Вот простой пример запутанного состояния. Если бы мы измерили положение первой частицы и нашли ее в P, то вторая автоматически оказалась бы в Q, а если первая была бы в P′, то вторая — в Q′ (Проблема квантовых измерений подробнее рассмотрена в разделе 2.8, а вопрос квантовой запутанности полнее обсуждается в разделе 2.10.)

Такие запутанные состояния пар частиц, бесспорно, очень странны и чужды нам, но это только начало. Мы видели, что такая квантовая запутанность возникает

как один из аспектов квантового принципа линейной суперпозиции. Однако, вообще говоря, этот принцип применяется не только к парным частицам, а гораздо шире. В случае с тройками частиц мы получим тройную запутанность, а также в состоянии запутанности могут оказываться и группы по четыре частицы. Численность частиц в такой группе вообще может быть любой.

Мы убедились, что такой принцип квантовой суперпозиции состояний имеет фундаментальное значение для линейности квантовой эволюции, которая, в свою очередь, играет ключевую роль при **U**-эволюции квантового состояния во времени (уравнение Шрёдингера). Стандартная квантовая механика никак не ограничивает масштабы, в которых **U** может применяться к физической системе. Например, вспомним кота из раздела 1.4. Допустим, есть комната, в которую можно попасть извне через две разные двери, A и B. За пределами комнаты — голодный кот, а в самой комнате лежит какая-то вкуснятина. Изначально обе двери закрыты. Предположим, к каждой из двух дверей подключен детектор высокоэнергетических фотонов — соответственно в точках A и B, — автоматически отворяющий «свою» дверь, как только улавливает фотон от светоделителя, расположенного в точке M (и вероятность такого попадания 50 %). С высокоэнергетическим фотоном, вылетающим из лазера L в точку M, в итоге могут произойти два события (рис. 2.15). Ситуация точно такая же, как и в эксперименте 1 из раздела 2.3 (см. рис. 2.4 *a*). При любой реальной постановке этого эксперимента перед котом откроется либо одна дверь, либо другая, и, соответственно, он пройдет или через первую дверь, или через вторую (вероятность каждого из вариантов здесь составляет по 50 %). Однако если бы мы попробовали подробно проследить эволюцию системы, то есть действовали в соответствии с **U** и ее подразумеваемой линейностью, которая касалась бы всех релевантных элементов опыта — лазера, излученного фотона, вещества в светоделителе, детекторов, дверей, самого кота, воздуха в комнатах и т. д., то состояние суперпозиции (начинающееся так: фотон, вылетая из светоделителя, находится в суперпозиции; он одновременно проскочил через светоделитель и отразился) должно было бы развиться в суперпозицию двух состояний, в каждом из которых открыта всего одна дверь. Наконец, должна наступить кошачья суперпозиция, в которой кот одновременно проходит сразу через две двери!

Это всего лишь частный случай вышеупомянутого свойства линейности. Мы считаем, что эволюция $\alpha_0 \rightsquigarrow \alpha_1$ должна начаться с того, что фотон покидает светоделитель в точке M и летит в A, и при этом кот находится снаружи. Заканчивается эволюция так: кот что-то ест в комнате, в которую проник через дверь A. Эволюция $\beta_0 \rightsquigarrow \beta$ протекает сходным образом, только фотон, покидающий светоделитель в M, летит в детектор B, и, следовательно, кот добирается

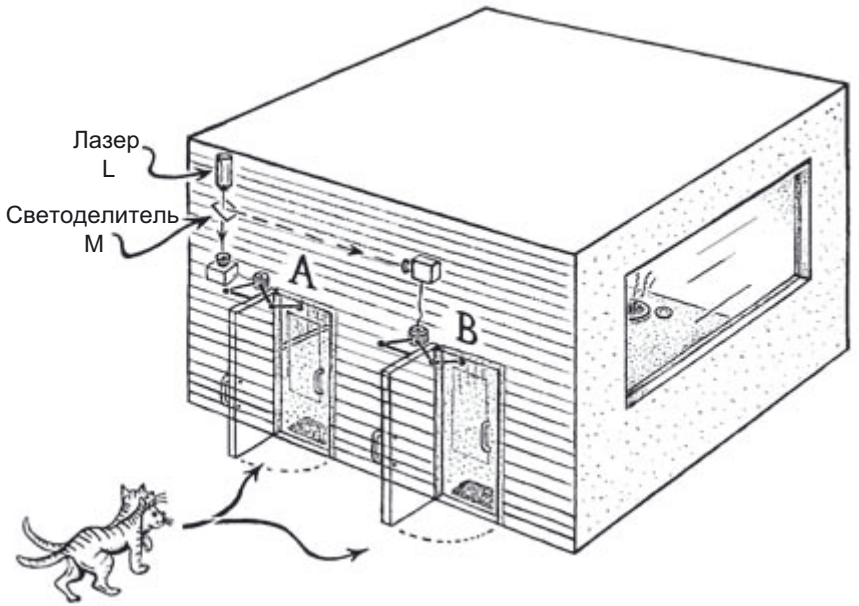


Рис. 2.15. Высокоэнергетический фотон вылетает из лазера L, направленного на светоделитель M. Два альтернативных пути (находящихся в суперпозиции) проходят через детекторы, причем, получив фотон, один из детекторов открывает дверь A, а другой — дверь B. Эти фотонные состояния находятся в квантовой линейной суперпозиции, и в такой же суперпозиции должно находиться открывание дверей, что следует из линейности квантового (**U**)-формализма. Более того, согласно **U**, движения кота также должны быть в суперпозиции, чтобы кот, таким образом, бежал к пище в комнате сразу по двум дорожкам!

до своей пищи через дверь B. Однако *общее* состояние, когда фотон вылетает из светоделителя, должно быть суперпозицией, скажем, $\boldsymbol{\alpha}_0 + \boldsymbol{\beta}_0$, которая будет развиваться согласно **U** как $\alpha_0 + \beta_0 \rightsquigarrow \alpha_1 + \beta_1$, поэтому и движение кота между двумя комнатами будет представлять суперпозицию, в которой кот проходит через обе двери одновременно, — определенно, ничего подобного не наблюдается! Это вариант парадокса под названием «кот Шрёдингера», и нам придется вернуться к нему в разделе 2.13. В стандартной квантовой механике такие вопросы решаются в соответствии с копенгагенской интерпретацией, предполагающей, что квантовое состояние *не описывает* физическую реальность, а всего лишь дает возможность вычислить различные вероятности тех или иных исходов, которые могут наступить в наблюдаемой системе. Это действие в процессе **R**, который упоминался в разделе 2.4 и к обсуждению которого мы как раз подходим.

2.8. Квантовое измерение

Для того чтобы понять, как квантовая механика управляется с такими, казалось бы, вопиющими несоответствиями между объективной реальностью и процессом **U**-эволюции, нужно сначала разобраться, как именно в квантовой теории устроена процедура **R**. Это проблема *квантового измерения*. Согласно квантовой теории, из любого квантового состояния системы можно извлечь лишь ограниченное количество информации и путем измерения удостовериться, каково *именно* данное квантовое состояние, а не каким оно могло бы быть. Действительно, любой измерительный прибор позволяет различать лишь некоторое ограниченное множество альтернатив для конкретного состояния. Если окажется, что состояние до измерения *не являлось* ни одной из этих альтернатив, то, согласно странной процедуре, которую мы должны принять в соответствии с **R**, состояние мгновенно *перескакивает* к одной из этих допустимых альтернатив, причем сама теория определяет вероятность такого скачка (на самом деле вероятность вычисляется по правилу Борна, о котором я вскользь упоминал выше, в разделе 2.6, а также в разделе 1.4, а подробно описываю далее).

Такие квантовые скачки — одно из наиболее странных свойств квантовой механики, и многие теоретики глубоко сомневаются в том, насколько вообще реальна эта процедура с физической точки зрения. По словам Вернера Гейзенберга [1971, p. 73–76], даже сам Эрвин Шрёдингер говорил: «Если эти проклятые квантовые скачки сохранятся в физике, я простить себе не смогу, что вообще связался с квантовой теорией!» В ответ на отчаянное замечание Шрёдингера Бор сказал: «Но зато все мы чрезвычайно благодарны вам... Ваша волновая механика явилась гигантским шагом вперед по сравнению со всеми прежними формами квантовой механики». Тем не менее, если принять эту процедуру, она дает нам квантово-механические результаты, которые полностью согласуются с наблюдениями!

Здесь потребуется гораздо более конкретно рассказать о процедуре **R**. В общепризнанных учебниках проблема квантовых измерений рассматривается в контексте свойств *линейных операторов* определенных типов (которые упоминались в разделе 1.14). Правда, хотя в конце данного раздела я ненадолго и вернусь к обсуждению этой связи с операторами, будет проще описать работу **R** более «прямолинейным» образом.

Прежде всего необходимо отметить факт (уже упоминавшийся в разделе 2.5), что семейство векторов состояния для некоторой квантовой системы всегда образует *комплексное векторное пространство* в том смысле, как оно объясняется в разделе А.3. В такой ситуации мы также должны включить в систему особый элемент **0** — *нулевой вектор*, не соответствующий какому-либо физическому

состоянию. Я обозначу это векторное пространство $\mathcal{H}$ (от названия *гильбертово пространство* — о том, что это такое, мы подробнее поговорим далее). Сначала я расскажу об общих, так называемых невырожденных квантовых измерениях, но существуют еще и *вырожденные измерения*, при которых невозможно отличить одну альтернативу от другой. Они будут кратко рассмотрены в конце этого раздела, а также в разделе 2.12.

В случае невырожденного измерения можно описать вышеупомянутый ограниченный набор альтернатив (возможных результатов измерения) как *ортогональный базис* для $\mathcal{H}$ в том смысле, который изложен в разделе А.4. Соответственно, все базисные элементы $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$, $\boldsymbol{\varepsilon}_3$, ... ортогональны друг другу таким образом, что образуют базис, простирающийся во всем пространстве $\mathcal{H}$. Последнее условие (подробнее описанное в разделе А.4) означает, что любой элемент $\mathcal{H}$ можно выразить как суперпозицию $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$, $\boldsymbol{\varepsilon}_3$, ... Кроме того, выражение конкретного состояния $\mathcal{H}$ через $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$, $\boldsymbol{\varepsilon}_3$, ... *единственно*, и фактически $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$, $\boldsymbol{\varepsilon}_3$, ... формирует *базис* для семейства. Термином *ортогональный* обозначаются пары таких состояний, которые в определенном смысле *независимы* друг от друга.

Смысл *независимости* квантовых ортогональных состояний нелегко понять в контексте классической физики. Возможно, самым близким классическим аналогом будет явление, связанное с модами колебаний, возникающих, например, при ударе в барабан или колокол, при которых в колоколе возбуждается целый набор колебаний, каждое из которых имеет свою частоту. Можно считать, что различные «чистые» моды колебаний независимы друг от друга, или *ортогональны* друг другу (в разделе А.11 обсуждается другой пример с колебаниями скрипичных струн), но с такой аналогией мы далеко не уйдем. В квантовой теории требуется нечто гораздо более конкретное (и тонкое), поэтому может быть полезно привести еще несколько примеров. Два корпускулярно-волновых состояния, которые совершенно не пересекаются друг с другом (например, траектории МАС и МВС в интерферометре Маха — Цендера из раздела 2.3; см. рис. 2.4 *б*), были бы ортогональными, но это условие не является обязательным. Существует множество других вариантов возникновения ортогональности между корпускулярно-волновыми состояниями. Например, два бесконечных волновых цуга с разными частотами также можно считать ортогональными. Более того, фотоны (с равной частотой и одинаковым направлением) с плоскостями поляризации, расположенными под прямым углом одна относительно другой, являются ортогональными. Вспомните классическое описание плоских волн электромагнитного поля из раздела 2.6, особенно тот факт, что вполне можно считать такие классические описания хорошими моделями состояний отдельных фотонов. Состояние фотона с круговой поляризацией не будет ортогонально никакому (в остальном идентичному) состоянию плоскополяризованного фотона, но со-

стояния левосторонней и правосторонней круговой поляризации ортогональны друг другу. Необходимо помнить, однако, что направление поляризации — это лишь малая часть информации об общем состоянии фотона, и вне зависимости от состояний поляризации состояния импульса двух фотонов (см. разделы 2.6 и 2.13) с разной частотой или направлением были бы ортогональными.

В геометрии термин *ортогональный* трактуется как *перпендикулярный* или *расположенный под прямым углом*, и хотя значение этого термина в квантовой теории обычно не имеет явного отношения к обычной пространственной геометрии, такое представление о перпендикулярности действительно актуально в геометрии комплексного векторного пространства $\mathcal{H}$ квантовых состояний (вместе с $\mathbf{0}$). Такое векторное пространство называется *гильбертовым пространством* в честь выдающегося математика XX века Давида Гильберта, предложившего такую концепцию в начале XX века[1] в ином контексте. Важная в данном случае концепция *под прямым углом* действительно относится к геометрии гильбертова пространства.

Что такое гильбертово пространство? С математической точки зрения это векторное пространство (см. раздел A.3), в котором может быть либо конечное, либо бесконечное число измерений. Скаляры представляются комплексными числами (элементами $\mathbb{C}$; см. раздел A.9), над элементами гильбертова пространства определено *скалярное произведение* $\langle ... | ... \rangle$ (см. раздел A.3), обладающее свойством *эрмитовости*:

$$\langle \beta \,|\, \alpha \rangle = \overline{\langle \alpha \,|\, \beta \rangle}$$

(линия сверху означает комплексное сопряжение; см. раздел A.10), и *неотрицательности*:

$$\langle \alpha \,|\, \alpha \rangle \geq 0,$$

причем

$$\langle \alpha \,|\, \alpha \rangle = 0, \text{ лишь если } \alpha = 0.$$

Вышеописанное понятие ортогональности очень просто определить через такое скалярное произведение: два вектора состояния α и β (ненулевые элементы

[1] Основную часть своей работы Гильберт опубликовал между 1904 и 1906 годами (это шесть статей, объединенных в издании [Hilbert, 1912]). На самом деле первую важную статью на эту тему написал Эрик Ивар Фредгольм [1903], чье исследование общей концепции «гильбертова пространства» предшествовало работе Гильберта. Вышеуказанные замечания адаптированы по [Dieudonné, 1981].

гильбертова пространства) ортогональны, если их скалярное произведение равно нулю:

$$\alpha \perp \beta, \text{ если } \langle \alpha \,|\, \beta \rangle = 0.$$

Также существует требование полноты, приобретающее важность тогда, когда пространство Гильберта содержит бесконечное число измерений; еще существует условие сепарабельности, ограничивающее бесконечный «размер» такого гильбертова пространства, но здесь мне не требуется отвлекаться на эти вопросы.

В квантовой механике комплексные скаляры $a, b, c, \ldots$ для такого векторного пространства — это комплексные амплитуды, подчиняющиеся принципу квантово-механической суперпозиции, а сам этот принцип суперпозиции обеспечивает операцию сложения в векторном гильбертовом пространстве. Размерность гильбертова пространства может быть конечной или бесконечной. В интересующем нас здесь контексте будет уместно предположить, что размерность выражается некоторым конечным числом n, хотя это число вполне может быть очень велико. Для этого случая я воспользуюсь нотацией

$$\mathcal{H}^n,$$

обозначив таким образом это гильбертово n-пространство (в сущности, уникальное для каждого n). Нотация $\mathcal{H}^\infty$ будет означать гильбертово пространство с бесконечным числом измерений. Для простоты я в основном ограничусь рассмотрением случая с конечным числом измерений. Поскольку скаляры являются комплексными числами, то и измерения являются комплексными (в таком смысле, как в разделе А.10), поэтому вещественное (евклидово) многообразие $\mathcal{H}^n$ было бы $2n$-мерным. Норма вектора состояния α, описанного в разделе 2.5, определяется как

$$\| \alpha \| = \langle \alpha \,|\, \alpha \rangle .$$

На самом деле это квадрат длины вектора α в обычном вещественном евклидовом понимании, если считать $\mathcal{H}^n$ $2n$-мерным евклидовым пространством. Понятие *ортогонального базиса*, упоминавшегося ранее, в таком случае соответствует описанию в разделе А.4 (для случая с конечным числом измерений) и представляет собой множество из n ненулевых векторов состояния $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$, $\boldsymbol{\varepsilon}_3$, ..., $\boldsymbol{\varepsilon}_n$, нормированных на единицу:

$$\| \boldsymbol{\varepsilon}_1 \| = \| \boldsymbol{\varepsilon}_2 \| = \| \boldsymbol{\varepsilon}_3 \| = \ldots = \| \boldsymbol{\varepsilon}_n \| = 1 ,$$

и взаимно ортогональных

$$\boldsymbol{\varepsilon}_j \perp \boldsymbol{\varepsilon}_k \text{ всегда, когда } j \neq k \ (j, k = 1, 2, 3, ..., n).$$

Любой вектор $\mathbf{z}$ в $\mathcal{H}^n$ (то есть вектор квантового состояния) можно (единственным способом) выразить как линейную комбинацию базисных элементов:

$$z = z_1\boldsymbol{\varepsilon}_1 + z_2\boldsymbol{\varepsilon}_2 + ... + z_n\boldsymbol{\varepsilon}_n,$$

где комплексные числа $z_1, z_2, ..., z_n$ (амплитуды) являются *компонентами* $\mathbf{z}$ в базисе $\{\boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_n\}$.

Следовательно, векторное пространство возможных состояний конечномерной квантовой системы является неким гильбертовым пространством $\mathcal{H}^n$ с n комплексными измерениями. Часто при обсуждении квантово-механических вопросов используются гильбертовы пространства с бесконечным числом измерений, особенно в квантовой теории поля (КТП; см. раздел 1.4). Однако могут возникать некоторые нетривиальные математические задачи, в частности, когда такая бесконечность является «несчетной» (то есть мощность множества больше канторовского $\aleph_0$; см. раздел А.2), в которых гильбертово пространство не удовлетворяет аксиоме сепарабельности (кратко упомянутой ранее), часто выдвигаемой в качестве обязательной [Streater and Wightman, 2000]. Эти задачи не играют особой роли при обсуждении тех проблем, о которых я хочу здесь рассказать, поэтому, приступая к обсуждению бесконечномерного гильбертова пространства, я имею в виду сепарабельное гильбертово пространство $\mathcal{H}^\infty$, для которого существует счетный ортонормированный базис $\{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \boldsymbol{\varepsilon}_3, \boldsymbol{\varepsilon}_4, ...\}$ (на самом деле много таких базисов). Когда базис бесконечен в таком смысле, при работе с ним придется учитывать проблемы сходимости (см. раздел А.10) следующим образом: чтобы у элемента $\mathbf{z}(= z_1\boldsymbol{\varepsilon}_1 + z_2\boldsymbol{\varepsilon}_2 + z_3\boldsymbol{\varepsilon}_3 + ...)$ была конечная норма $\|\mathbf{z}\|$, требуется, чтобы ряд $|z_1|^2 + |z_2|^2 + |z_3|^2 + ...$ сходился к конечному значению.

Следует напомнить, что в разделе 2.5 я подчеркивал разницу между физическим состоянием квантовой системы и ее математическим описанием, причем в последнем используются векторы состояний, скажем $\boldsymbol{\alpha}$. Векторы $\boldsymbol{\alpha}$ и $q\boldsymbol{\alpha}$, где q — ненулевое комплексное число, должны соответствовать одному и тому же физическому квантовому состоянию. Следовательно, различные физические состояния сами по себе будут представлены различными одномерными подпространствами $\mathcal{H}$ (комплексными линиями, проходящими через начало координат), или *лучами*. Каждый такой луч дает полное семейство комплексных величин, кратных $\boldsymbol{\alpha}$; в терминах вещественных чисел такой луч является копией плоскости Весселя (чей нуль 0 находится в начале координат $\mathbf{0}$ в $\mathcal{H}$). Можно считать, что нормированные векторы состояний соответствуют единичным ме-

рам длины на этой плоскости Весселя, то есть представляют точки на единичной окружности на этой плоскости. Все векторы состояний, задаваемые точками на этой единичной окружности, представляют одно и то же физическое состояние, поэтому среди единичных векторов состояния по-прежнему сохраняется фазовая свобода умножения на комплексное число единичного модуля ($e^{i\theta}$, где θ — вещественное число) без изменения физического состояния (см. в конце раздела 2.5). Однако следует отметить, что это касается лишь полного состояния системы. При умножении различных элементов состояния на разные фазы вполне можно изменить физику состояния.

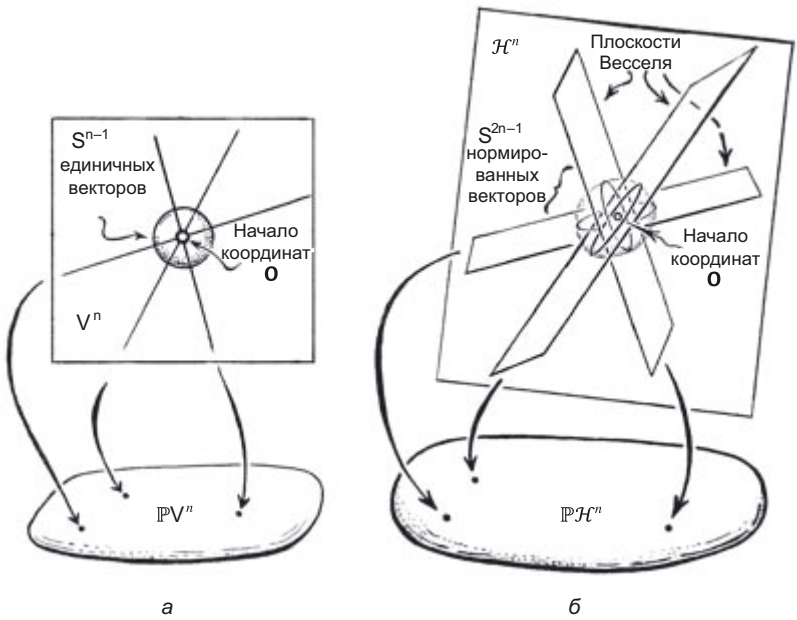


Рис. 2.16. *Проективное* пространство для n-мерного векторного пространства V^n — это $(n-1)$-мерное компактное пространство $\mathbb{P}V^n$ из лучей (одномерных подпространств) V^n, тогда как V^n (без начала координат) — это расслоение поверх $\mathbb{P}V^n$. Это проиллюстрировано в вещественном случае (*а*) и в комплексном случае (*б*), где лучи копируют плоскости Весселя; здесь же показаны их единичные окружности. На рис. *б* представлена квантово-механическая ситуация, где $\mathcal{H}^n$ — пространство векторов квантовых состояний, $\mathbb{P}\mathcal{H}^n$ — пространство физически отличающихся квантовых состояний, причем нормированные состояния образуют круговое расслоение на $\mathbb{P}\mathcal{H}^n$

$(n-1)$ -комплексномерное пространство, каждая из отдельных точек которого соответствует одному из этих лучей, называется проективным гильбертовым пространством $\mathbb{P}\mathcal{H}^n$. На рис. 2.16 проиллюстрирована идея вещественного проек-

тивного пространства $\mathbb{P}V^n$, производного от *вещественного* векторного n-мерного пространства V^n (см. рис. 2.16 *а*). В комплексном случае лучи являются копиями плоскости Весселя (см. рис. 2.16 *б* и рис. A.34 в разделе A.10). На рис. 2.16 *б* также показаны подпространства *нормированных* (то есть единичных) векторов, образующие сферу S^{n-1} в вещественном случае и S^{2n-1} в комплексном случае. Каждое физически отличное квантовое состояние системы будет представлено неким комплексным лучом в гильбертовом пространстве $\mathcal{H}^n$ и, следовательно, некоей отдельной точкой в проективном пространстве $\mathbb{P}\mathcal{H}^n$. Геометрия пространства физически отличающихся квантовых возможностей некоторой конечной физической системы может трактоваться как комплексная проективная геометрия некоторого проективного пространства Гильберта $\mathbb{P}\mathcal{H}^n$.

Теперь мы уже можем понять, как правило Борна действует при общих, то есть невырожденных, измерениях (см. также раздел 1.4). Квантовая механика говорит, что для любого такого измерения будет ортонормированный базис $(\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \boldsymbol{\varepsilon}_3, \ldots)$ возможных результатов и физический исход измерения всегда должен соответствовать одному из них. Предположим, что квантовое состояние до измерения описывается вектором $\boldsymbol{\psi}$, который мы сейчас будем считать нормированным ($\|\boldsymbol{\psi}\| = 1$). В терминах базиса эту ситуацию можно выразить следующим образом:

$$\boldsymbol{\psi} = \psi_1\boldsymbol{\varepsilon}_1 + \psi_2\boldsymbol{\varepsilon}_2 + \psi_3\boldsymbol{\varepsilon}_3 + \ldots,$$

причем комплексные числа $\psi_1, \psi_2, \psi_3, \ldots$ (компоненты $\boldsymbol{\psi}$ в базисе) — это те самые амплитуды, что упоминались в разделах 1.4 и 1.5. Квантово-механическая процедура **R** не сообщает нам, к какому из состояний $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \boldsymbol{\varepsilon}_3, \ldots$ перепрыгнет состояние ψ непосредственно после измерения, однако дает нам вероятность p_j для каждого возможного результата, которая задается правилом Борна, которое гласит: вероятность того, что в результате измерения мы обнаружим состояние $\boldsymbol{\varepsilon}_j$ (то есть того, что произошел «скачок» к $\boldsymbol{\varepsilon}_j$), равна *квадрату модуля* соответствующей амплитуды ψ_j, а именно:

$$p_j = |\psi_j|^2 = \psi_j\overline{\psi}_j.$$

Необходимо отметить следующее. Математическое требование нормировки вектора состояния ψ и всех базисных векторов ε_j имеет следствие: необходимое свойство вероятностей таково, что все разнообразные возможности должны в сумме давать единицу. Этот поразительный факт проистекает просто из выражения, описывающего условие нормировки:

$$\|\boldsymbol{\psi}\| = 1$$

в контексте ортонормированного базиса $\{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \boldsymbol{\varepsilon}_3, ...\}$, что благодаря $\langle \boldsymbol{\varepsilon}_i \,|\, \boldsymbol{\varepsilon}_j \rangle = \delta_{ij}$ (см. раздел А.10) дает:

$$\begin{aligned} \| \boldsymbol{\psi} \| &= \langle \boldsymbol{\psi} \,|\, \boldsymbol{\psi} \rangle = \langle \psi_1 \boldsymbol{\varepsilon}_1 + \psi_2 \boldsymbol{\varepsilon}_2 + \psi_3 \boldsymbol{\varepsilon}_3 + \ ... \ | \ \psi_1 \boldsymbol{\varepsilon}_1 + \psi_2 \boldsymbol{\varepsilon}_2 + \psi_3 \boldsymbol{\varepsilon}_3 + \ ... \rangle = \\ &= |\psi_1|^2 + |\psi_2|^2 + |\psi_3|^2 + \ ... = \\ &= p_1 + p_2 + p_3 + \ ... = 1. \end{aligned}$$

Так проявляется примечательнейшее соответствие между общим математическим аппаратом квантовой механики и согласованностью требований вероятностного квантового поведения!

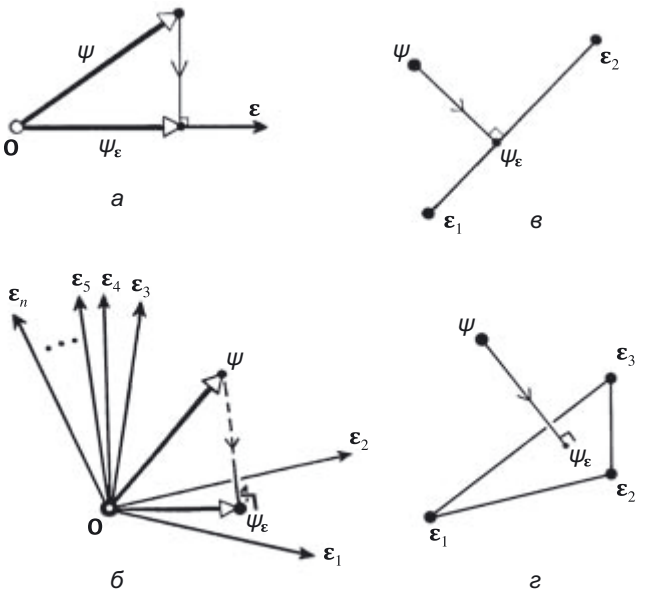


Рис. 2.17. Правило Борна в геометрическом выражении (с ненормированными векторами состояния), где вероятность того, что $\boldsymbol{\psi}$ перепрыгнет к ортогонально спроецированному $\boldsymbol{\psi}_\varepsilon$ задается выражением $\| \boldsymbol{\psi}_\varepsilon \| \div \| \boldsymbol{\psi} \|$, при невырожденном измерении (*а*), при вырожденном измерении (*б*), которое не позволяет отличить $\boldsymbol{\varepsilon}_1$ от $\boldsymbol{\varepsilon}_2$ (проекционный постулат), а также сущность геометрии последнего, показанная в проективном гильбертовом пространстве (*в*), и проективная картина, когда вырождение существует между тремя состояниями: $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$ и $\boldsymbol{\varepsilon}_3$ (*г*)

Тем не менее можно переформулировать правило Борна так, чтобы не требовалось нормировать ни $\boldsymbol{\psi}$, ни базисные векторы. Для этого обратимся к евклидову понятию *ортогональной проекции*. Допустим, что измеряется состояние, которое непосредственно перед актом измерения обладает вектором состояния $\boldsymbol{\psi}$, а изме-

рение удостоверяет, что вектор состояния после измерения оказывается («перескакивает») кратным $\boldsymbol{\varepsilon}$. Тогда вероятность p такого результата представляет собой отношение нормы ортогональной проекции $\boldsymbol{\psi}_\varepsilon$ вектора $\boldsymbol{\psi}$ на (комплексное направление) $\boldsymbol{\varepsilon}$ к норме самого вектора $\boldsymbol{\psi}$, то есть это величина

$$p = \frac{\|\psi_\varepsilon\|}{\|\psi\|}.$$

Данная проекция проиллюстрирована на рис. 2.17 *а*. Следует отметить, что вектор $\boldsymbol{\psi}_\varepsilon$ является уникальным скаляром, кратным $\boldsymbol{\varepsilon}$, таким, что $(\psi - \psi_\varepsilon) \perp \varepsilon$. Именно так *ортогональная проекция* и трактуется в этом контексте.

Одно из преимуществ данной интерпретации правила Борна заключается в том, что она естественным образом распространяется и на более общую ситуацию *вырожденного* измерения, для которого должно учитываться и другое правило, известное под названием *проекционный постулат*. Некоторые физики пытались доказать, что этот постулат (даже в простой форме квантового скачка, когда вырождения измерения не происходит) не нужен в стандартной квантовой механике, поскольку при нормальном измерении результирующее состояние измеряемого объекта вряд ли окажется независимым — оно запутается с измерительным прибором вследствие взаимодействия с ним. Однако этот постулат действительно необходим, особенно в ситуациях, именуемых *нулевыми измерениями* (см., например, ПкР, раздел 22.7), когда квантовый скачок состояния неизбежен, даже если он не затрагивает измерительный прибор.

Для вырожденного измерения характерен следующий факт: оно не позволяет различить результаты, которые определенно неодинаковы с физической точки зрения. В такой ситуации у нас нет по существу единственного базиса различимых результатов ($\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$, $\boldsymbol{\varepsilon}_3$, ...), но некоторые $\boldsymbol{\varepsilon}_j$ дают одинаковые результаты при измерении. Допустим, $\boldsymbol{\varepsilon}_1$ и $\boldsymbol{\varepsilon}_2$ — как раз такие варианты. Тогда все линейное пространство состояний, задаваемое $\boldsymbol{\varepsilon}_1$ и $\boldsymbol{\varepsilon}_2$, будет давать один и тот же результат измерения. Эта (комплексная) плоскость, проходящая через начало координат в гильбертовом пространстве $\mathcal{H}^n$ (задаваемом $\boldsymbol{\varepsilon}_1$ и $\boldsymbol{\varepsilon}_2$), — не просто отдельный луч, который мы получаем для отдельного физического квантового состояния (рис. 2.17 *б*). (Здесь возможна путаница с терминологией, поскольку иногда *комплексной* называют такую плоскость, которую я именую *плоскостью Весселя* (см. раздел А.10). В плоскости такого типа, о которой идет речь здесь, два комплексных измерения и, следовательно, четыре вещественных.) В контексте проективного гильбертова пространства $\mathbb{P}\mathcal{H}^n$ физических квантовых состояний у нас теперь целый (комплексный) ряд возможных результатов измерения, а не просто точка (см. рис. 2.17 *в*). Если такое измерение на $\boldsymbol{\psi}$ теперь даст результат,

который, как выяснится, лежит на этой плоскости в $\mathcal{H}^n$ (то есть это будет линия на $\mathbb{P}\mathcal{H}^n$), то конкретное состояние, получаемое в результате измерения, должно быть пропорционально проекции $\boldsymbol{\psi}$ на этой плоскости (см. рис. 2.17 *б* и *в*). Подобная ортогональная проекция применяется и в случае, когда вырождению подвергаются три или более состояния (см. рис. 2.17 *г*).

В некоторых крайних случаях вырожденных измерений все это может применяться сразу к нескольким различным наборам состояний. Соответственно, вместо измерения, которое позволило бы определить базис для альтернативных возможных исходов, такой процесс позволяет определить семейство линейных подпространств разной размерности, и каждое такое пространство будет ортогонально всем прочим, причем именно подпространства можно будет различать в результате такой операции. Любое состояние $\boldsymbol{\psi}$ можно будет уникально выразить как сумму различных проекций, определяемых путем измерения, причем эти проекции будут альтернативами, в направлении которых при измерении может произойти скачок $\boldsymbol{\psi}$. Соответствующие вероятности опять же даются отношениями норм проекций $\|\boldsymbol{\psi}\|$ к норме $\boldsymbol{\psi}$.

Хотя в вышеизложенном описании мне и удалось постулировать все, что нам нужно от квантового измерения $\mathbf{R}$, даже не упоминая о квантовых операторах, было бы правильно познакомиться также и с этим, более традиционным, способом описания явлений. Интересующие нас операторы называются *эрмитовыми*, или *самосопряженными*, операторами (между двумя этими понятиями есть тонкое различие, актуальное лишь в случае с бесконечным числом измерений и не важное для нас), причем оператор $\mathbf{Q}$ удовлетворяет равенству

$$\langle \varphi \,|\, \mathbf{Q}\, \psi \rangle = \langle \mathbf{Q}\, \varphi \,|\, \psi \rangle$$

для любой пары состояний $\boldsymbol{\varphi}$, $\boldsymbol{\psi}$ в $\mathcal{H}^n$. (Читатель, знакомый с понятием *эрмитовой матрицы*, увидит, что это именно то, что утверждается выше относительно $\mathbf{Q}$.) В таком случае базис $\{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \boldsymbol{\varepsilon}_3, \ldots\}$ состоит из так называемых *собственных векторов* $\mathbf{Q}$, причем собственный вектор — это элемент μ в $\mathcal{H}^n$, для которого

$$\mathbf{Q}\, \mu = \lambda\, \mu,$$

где число λ называется *собственным значением*, соответствующим μ. В квантовом измерении собственное значение λ_j представляло бы собой определяемое при измерении *числовое значение*, которое на самом деле получает $\mathbf{Q}$, когда после измерения происходит квантовый скачок состояния к собственном вектору $\boldsymbol{\varepsilon}_j$. (Фактически λ_j должно быть *вещественным* числом для эрмитова оператора, что действительно согласуется со следующим фактом: проводя измерение (в обычном смысле этого слова), мы получаем в результате *вещественные* числа.) В ка-

честве комментария хотел бы пояснить, что собственное значение λ_j не имеет ничего общего с амплитудой ψ_j. Квадрат модуля ψ_j сообщает *вероятность* того, что результат эксперимента действительно будет равен числовому значению λ_j.

Все это очень формально, и абстрактная комплексная геометрия гильбертова пространства кажется весьма отвлеченной от непосредственно наблюдаемой нами геометрии обычного пространства. Тем не менее общему математическому аппарату квантовой механики присуща очевидная геометрическая красота; такая красота есть и в отношении квантовой механики к **U** и **R**. Поскольку размерности квантовых гильбертовых пространств представляются довольно высокими, не говоря уж о том, что речь идет о геометрии в комплексных числах, а не о привычной нам геометрии вещественных чисел, непосредственно визуализировать такую геометрию нелегко. Однако в следующем разделе мы увидим, что при обсуждении сугубо квантово-механического феномена под названием *спин* геометрию можно понять, непосредственно связав ее с геометрией обычного трехмерного пространства, и это, возможно, поможет уяснить всю суть квантовой механики.

2.9. Геометрия квантового спина

Наиболее четкая взаимосвязь между геометрией гильбертова пространства и геометрией обычного трехмерного пространства действительно прослеживается в случае с состояниями спина. Это особенно характерно для частицы, которая обладает массой, со спином $\frac{1}{2}$ — таковы, например, электрон, протон или нейтрон либо определенные атомные ядра или атомы. Изучая такие состояния спина, мы можем лучше представить себе, как именно осуществляется процесс измерения в квантовой механике.

Частица со спином $\frac{1}{2}$ всегда имеет определенный момент количества движения, а именно $\frac{1}{2}\hbar$ (см. раздел 1.14), но *направление* спина всегда проявляется тонким квантово-механическим образом. Если ненадолго представить эту ситуацию как классическую, то направление спина можно определить по правилу буравчика: если мы завинчиваем винт, то есть вектор момента импульса направлен от нас, то мы должны вращать этот винт по часовой стрелке. Такое положение спина является *правоориентированным*. Частица с любым спином могла бы вращаться вокруг своей оси и в противоположную сторону, то есть была бы левоориентированной, но такое вращение принято также описывать как правоориентированное, правда, спин тогда будет направлен в *противоположное* направление.

Квантовые состояния спина у частицы со спином $\frac{1}{2}$ в точности соответствуют этим классическим состояниям, хотя и подчиняются странным квантово-механическим законам. Следовательно, для любого пространственного направления возможно такое состояние, в котором частица вращается относительно этого направления по часовой стрелке, причем величина ее момента количества движения составляет $\frac{1}{2}\hbar$. Однако, согласно квантовой механике, все эти возможные состояния можно выразить как линейные суперпозиции любых двух таких различных состояний, которые полностью перекрывают все пространство возможных состояний спина. Если взять два этих состояния и условиться, что у них разнонаправленные спины, то эти состояния будут *ортогональными*. Итак, у нас есть 2-комплексномерное гильбертово пространство $\mathcal{H}^n$, и *ортогональный базис* для состояний спина всегда будет (речь о правоориентированном спине) такой парой противоположных направлений. Вскоре мы увидим, как любое другое направление спина такой частицы действительно можно выразить в виде квантовой линейной суперпозиции двух этих противоположных спиновых состояний.

В литературе часто используется базис, в котором два этих направления представляются *верхним* и *нижним* и записываются следующим образом:

$$\left|\uparrow\right\rangle \text{ и } \left|\downarrow\right\rangle$$

соответственно. Далее я начинаю пользоваться предложенной Дираком кет-нотацией для описания векторов квантовых состояний. В такой нотации те или иные описательные символы или буквы записываются между символами «$|...\rangle$». Здесь я не буду подробно останавливаться на смысле этой нотации, однако можно отметить, что в данной форме векторы состояния называются *кет*-векторами, а *дуальные* им (см. раздел 1.4) именуются «*бра*-векторами» и записываются в виде «$\langle ...|$», так что бра и кет образуют цельную скобку «$\langle ...|...\rangle$», когда вычисляется внутреннее произведение [Dirac, 1947]. Любое другое возможное состояние спина $\left|\nearrow\right\rangle$ нашей частицы должно представлять линейную комбинацию двух базисных состояний:

$$\left|\nearrow\right\rangle = w\left|\uparrow\right\rangle + z\left|\downarrow\right\rangle,$$

причем из раздела 2.5 мы помним, что лишь отношение $z:w$ позволяет различать физически разные квантовые состояния. Это комплексное отношение по сути представляет собой просто частное:

$$u = \frac{z}{w},$$

но мы при этом должны допускать и такую возможность, что w окажется равным нулю, например, для того чтобы отразить состояние $|\downarrow\rangle$. Эта задача решается, если просто допустить для такого отношения (формальную) запись $u = \infty$, когда $w = 0$. С геометрической точки зрения, включая в плоскость Весселя точку со значением ∞, мы просто сворачиваем эту плоскость в очень большую сферу (примерно как когда мы стоим на поверхности Земли, она представляется нам плоской, на самом же деле Земля является очень большим шаром), замыкая ее в точке ∞. В результате получаем простейшую риманову поверхность (см. разделы A.10 и 1.6), именуемую *римановой сферой* (хотя в данном контексте она иногда может называться *сферой Блоха* или *сферой Пуанкаре*).

Обычно эту геометрию представляют следующим образом. Пусть плоскость Весселя горизонтально расположена в евклидовом трехмерном пространстве, где риманова сфера — это сфера с единичным диаметром с центром в точке O, причем эта точка совпадает с началом координат 0 в плоскости Весселя. Единичная окружность в плоскости Весселя принимается за *экватор* римановой сферы. Теперь рассмотрим точку S, расположенную на *южном полюсе* сферы, и выполним проекцию поверхности сферы на плоскость Весселя таким образом, чтобы точка Z на плоскости Весселя соответствовала точке Z′ на римановой сфере, если все три точки — S, Z и Z′ — расположены на одной прямой (это стереографическая проекция; см. рис. 2.18). В контексте стандартных декартовых координат (x, y, z) для трехмерного пространства плоскость Весселя представляется уравнением $z = 0$, а риманова сфера — уравнением $x^2+y^2+z^2 = 1$. Тогда комплексное число $u = x + \mathrm{i}y$ в плоскости Весселя с декартовыми координатами $(x, y, 0)$ соответствует Z′ на римановой сфере с декартовыми координатами $(2\lambda x, 2\lambda y, \lambda(1 - x^2 - y^2))$, где $\lambda = (1 + x^2 + y^2)^{-1}$. *Северный полюс* N соответствует началу координат O в плоскости Весселя и представляет комплексное число 0. Все точки на единичной окружности в плоскости Весселя ($\mathrm{e}^{\mathrm{i}\theta}$, где θ является вещественным; см. раздел A.10), включая 1, i, –1, –i, совпадают с соответствующими точкам на экваторе римановой сферы. Южный полюс S римановой сферы соответствует дополнительной точке ∞ на сфере, и на плоскости Весселя эта точка отображается на бесконечность.

Теперь давайте рассмотрим, как это математическое представление комплексных соотношений связано с состояниями спина нашей частицы (со спином $\frac{1}{2}$). Эти состояния описываются выражением $\left|\nearrow\right\rangle = w\left|\uparrow\right\rangle + z\left|\downarrow\right\rangle$, где $u = z/w$. Пока $w \neq 0$, можно считать $u = z/w$ обычным комплексным числом, и для удобства выберем такой масштаб, при котором $w = 1$ (опустив при этом какие-либо требования о нормировании $\left|\nearrow\right\rangle$), тогда $u = z$, а наш вектор состояния будет равен

$$\left|\nearrow\right\rangle = \left|\uparrow\right\rangle + z\left|\downarrow\right\rangle.$$

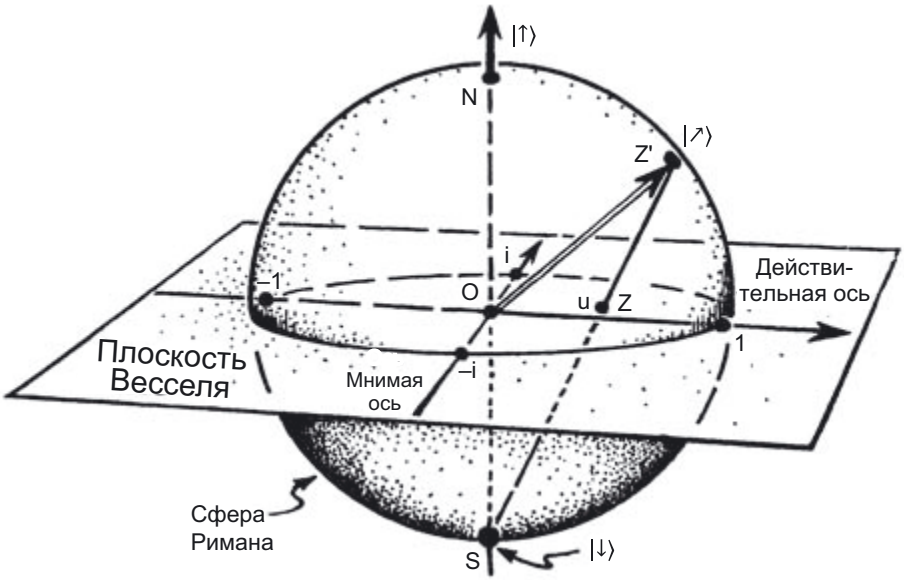


Рис. 2.18. Сфера Римана соотнесена при помощи стереографической проекции с плоскостью Весселя, проходящей через ее экватор, причем точка Z на плоскости проецируется от южного полюса на точку сферы Z′ (N — северный полюс, а O — центр сферы). Так мы получаем геометрическое представление направления спина $|\nearrow\rangle = |\uparrow\rangle + u\,|\downarrow\rangle$ для частицы, обладающей массой, со спином $\frac{1}{2}$, где Z и Z′ соответствуют комплексному числу u на плоскости Весселя и на сфере Римана соответственно

Можно считать, что значению z будет соответствовать точка Z на нашей плоскости Весселя, и эта точка соответствует точке Z′ на римановой сфере. Если правильно подобрать фазы для $|\uparrow\rangle$ и $|\downarrow\rangle$, то, к счастью, можно убедиться в том, что направление $\overrightarrow{OZ'}$ есть направление спина $|\nearrow\rangle$. Состояние $|\downarrow\rangle$ соответствует южному полюсу на сфере Римана и при этом соответствует $z = \infty$, но для получения последнего равенства нам потребовалось бы нормировать состояние $|\nearrow\rangle$ по-другому, например так: $|\nearrow\rangle = z^{-1}|\uparrow\rangle + |\downarrow\rangle$ (и считать $\infty^{-1} = 0$).

Риманова сфера, будучи просто пространством отношений $w : z$ для пары комплексных чисел w и z, из которых хотя бы одно не равно нулю, на самом деле является проективным гильбертовым пространством $\mathbb{P}\mathcal{H}^2$, описывающим массив возможных физически различных квантовых состояний, возникающих в результате суперпозиций любых двух различных независимых квантовых состояний какого-либо рода. Но вот что особенно поразительно в случае частиц (обладающих массой) со спином $\frac{1}{2}$: эта риманова сфера в точности соответствует

направлениям в окрестностях точки, расположенной в обычном трехмерном физическом пространстве. Если бы в мире насчитывалось иное число пространственных измерений, чего, по-видимому, требует современная теория струн (см. раздел 1.6), то не возникало бы такого простого и красивого отношения между пространственной геометрией и квантовой комплексной суперпозицией. Тем не менее, даже если не требовать такой прямой геометрической интерполяции, представление $\mathbb{P}\mathcal{H}^2$, в качестве римановой сферы остается полезным. Любой ортогональный базис для двумерного гильбертова пространства $\mathcal{H}^2$ по-прежнему будет соответствовать паре диаметрально противоположных точек на (абстрактной) римановой сфере, причем оказывается, что, прибегнув к небольшому упрощению геометрии, всегда можно интерпретировать правило Борна по следующему геометрическому принципу. Допустим, C — точка на сфере, соответствующая исходному (скажем, спиновому) состоянию. Затем делается измерение, после которого мы выбираем точку A или B, а после этого проводим перпендикуляр от C к диаметру AB и находим на этом диаметре точку D. Далее правило Борна можно интерпретировать следующим геометрическим образом (рис. 2.19):

$$\text{вероятность скачка от C к A} = \frac{\text{DB}}{\text{AB}},$$

$$\text{вероятность скачка от C к B} = \frac{\text{AD}}{\text{AB}}.$$

Иными словами, если принять за единицу диаметр (а не радиус) сферы, то длины отрезков DB и AD непосредственно дают нам вероятности перехода из текущего состояния в A или в B.

Поскольку это применимо к любой ситуации, где измерения делаются в системе из двух состояний, а не только для обладающих массой частиц со спином $\frac{1}{2}$, данная закономерность в простейшей форме применима для ситуаций, рассмотренных в разделе 2.3, где в первом эксперименте (см. рис. 2.4 *a*) светоделитель помещает фотон в суперпозицию двух возможных путей, причем каждый из детекторов фотонов будет представлен состоянием, представляющим собой суперпозицию наличия и отсутствия фотона. Давайте (формально) сравним эту ситуацию с частицей, имеющей спин $\frac{1}{2}$, причем исходно будем считать, что спин находится в состоянии $|\downarrow\rangle$. На рис. 2.4 *a* это состояние соответствовало бы импульсному состоянию фотона, вылетающего из лазерного источника и летящего вправо (направление MA). После того как этот фотон попадает в светоделитель, его исходное импульсное состояние становится суперпозицией двух состояний:

состояния «движение вправо» (формально по-прежнему соответствующего $|\downarrow\rangle$), которое можно уподобить точке A на рис. 2.19, и состояния «движение вверх» (отрезок MB на рис. 2.4 *а*), которое теперь соответствует $|\uparrow\rangle$ и представлено точкой B на рис. 2.19. Теперь импульс фотона представляет собой суперпозицию двух этих состояний, и мы можем уподобить его точке F на рис. 2.19, приписав каждому из альтернативных вариантов равную вероятность — по 50 %. Что касается второго эксперимента из раздела 2.3 (с прибором Маха — Цендера), показанного на рис. 2.4 *б*, с детекторами в точках D и E, то зеркала и светоделитель фактически нужны для того, чтобы вернуть импульс фотона в исходное состояние (соответствующее $|\downarrow\rangle$ и представленное точкой A на рис. 2.19), на регистрацию которого специально настроен детектор D (см. рис. 2.4 *б*). Так возникает 100 %-я вероятность поймать фотон в D и 0 %-я вероятность того, что фотон обнаружится в точке E.

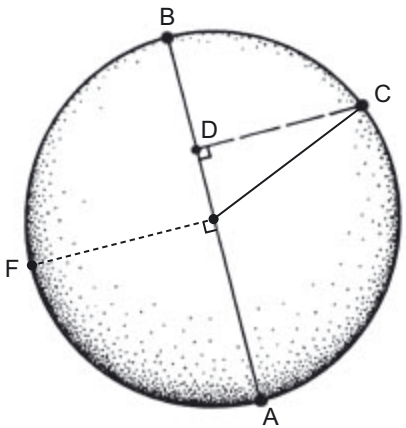


Рис. 2.19. Измерение квантовой системы, допускающей два состояния. Такое измерение позволяет различать ортогональные пары состояний A, B, которые, соответственно, представлены на римановой сфере $\mathbb{P}\mathcal{H}^2$ диаметрально противоположными точками A и B. Измерительный прибор находится в состоянии C, представленном точкой C на $\mathbb{P}\mathcal{H}^2$. По правилу Борна имеем, что с вероятностью DB/AB можно зарегистрировать на приборе скачок из C в A и с вероятностью AD/AB — скачок из C в B, где D — ортогональная проекция C на диаметр AB

Этот пример весьма ограничен по числу представленных в нем суперпозиций, но не составляет труда расширить его так, чтобы в нем проявлялся весь спектр комплексно-взвешенных альтернатив. Во многих реальных экспериментах это было проделано в контексте состояний поляризации фотонов, а не в контексте различных импульсных состояний. Поляризация фотона — еще один пример

квантово-механического спина, но оказывается, что спин будет либо полностью правоориентирован, либо полностью левоориентирован в направлении движения в соответствии с двумя состояниями круговой поляризации, рассмотренными в разделе 2.6. Опять же, имеем систему из двух состояний и проективное гильбертово пространство $\mathbb{P}\mathcal{H}^2$.[1] Соответственно, общее состояние можно уподобить точке на римановой сфере, но в таком случае у нас получится немного иная геометрия.

Для того чтобы исследовать эту проблему, давайте расположим сферу так, чтобы ее северный полюс N находился в направлении движения фотона, чтобы N соответствовало правоориентированному состоянию спина $|\circlearrowleft\rangle$. Таким образом, южный полюс S — это левоориентированное состояние спина $|\circlearrowright\rangle$. Общее состояние

$$|\psi\rangle = |\circlearrowleft\rangle + w|\circlearrowright\rangle$$

на римановой сфере можно уподобить точке Z', соответствующей z/w, как в вышеописанном случае с частицей, обладающей массой и имеющей спин $\frac{1}{2}$ (см. рис. 2.18). Однако с геометрической точки зрения было бы гораздо уместнее представить это состояние точкой Q на римановой сфере — такой точкой, которая соответствовала бы квадратному корню из z, то есть комплексному числу q (в принципе, это то самое q, о котором шла речь в разделе 2.6), чтобы было справедливо выражение:

$$q^2 = z/w.$$

Показатель степени 2 появляется из-за того, что спин фотона равен 1, то есть *вдвое больше базовой единицы спина*, равной $\frac{1}{2}\hbar$ (ею обладает электрон). Если бы речь шла о безмассовой частице со спином $\frac{1}{2}n$, то есть в n раз больше базовой единицы спина, то мы бы имели дело со значениями q, удовлетворяющими равенству $q^n = z$. У фотона $n = 2$, так что $q = \pm\sqrt{z}$. Чтобы найти соотношение Q с эллипсом поляризации фотона (см. раздел 2.5), сначала найдем большой круг, представляющий собой пересечение римановой сферы и плоскости, проходящей через O перпендикулярно линии OQ (рис. 2.20); затем спроецируем полученную фигуру на горизонтальную плоскость Весселя и получим эллипс. Оказывается, этот контур и является эллипсом поляризации фотона и наследует правую

[1] В последних экспериментах при помощи отдельных фотонов удалось экспериментально создать гильбертовы пространства высокой размерности; для этого были использованы степени свободы, связанные с орбитальным моментом импульса состояния фотона [Fickler et al., 2012].

ориентацию вектора OQ большого круга сферы. Вектор OQ связан с так называемым *вектором Стокса*, но более непосредственно — с *вектором Джонса* (техническое описание этих векторов дается в главе 3 книги Hodgkinson and Wu [1998]; см. также ПкР, раздел 22.9). Можно отметить, что q и $-q$ дают одинаковый эллипс и одинаковую ориентацию.

Довольно интересно рассмотреть, как *массивные состояния высших спинов* также могут быть представлены в контексте римановой сферы при помощи так называемого майоранова описания [Majorana, 1932; ПкР, раздел 22.10]. Для частицы, обладающей массой (скажем, для атома), со спином $\frac{1}{2}n$, где n — неотрицательное целое число (такое, что пространство физически различных возможностей может быть представлено как $\mathbb{P}\mathcal{H}^{n+1}$), а любое физическое спиновое состояние задается неупорядоченным множеством из n точек на римановой сфере (совпадения допускаются), можно считать, что каждая из этих точек соответствует определенному вкладу спина $\frac{1}{2}$ в направление от центра к этой точке (рис. 2.21). Каждое из этих направлений я буду называть *майорановым направлением*.

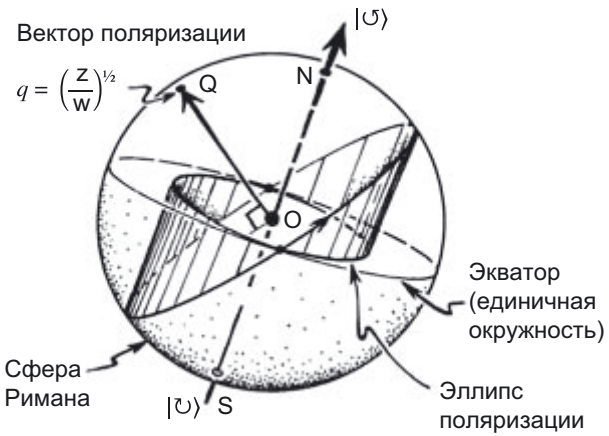


Рис. 2.20. Общее состояние поляризации фотона $w|\circlearrowleft\rangle + z|\circlearrowright\rangle$, где направление движения указывается стрелкой вправо вверх; два (нормированных) состояния поляризации фотона обозначаются как $|\circlearrowleft\rangle$ (правое) и $|\circlearrowright\rangle$ (левое). С геометрической точки зрения эту ситуацию можно представить на сфере Римана комплексным числом q (обозначенным буквой Q), где $q^2 = z/w$ (причем в левоориентированном случае допускается $q = \infty$, когда Q находится на южном полюсе S, и $q = 0$ в правоориентированном случае, когда Q находится на северном полюсе N. При этом ON — направление движения фотона, а O — центр сферы). Эллипс поляризации фотона является экваториальной проекцией большого круга, лежащего в плоскости, перпендикулярной направлению OQ

Такие общие состояния для спина $> \frac{1}{2}$ нечасто рассматриваются физиками, однако они склонны рассуждать о состояниях высших спинов в терминах, описывающих тип измерения, используя *аппарат Штерна — Герлаха* (рис. 2.22). В опыте Штерна — Герлаха используется сильно неоднородное магнитное поле, позволяющее измерить магнитный момент частицы (как правило, ориентированный точно так же, как и спин), и это поле отклоняет частицы[1] на угол, зависящий от величины проекции спина (строго говоря, магнитного момента) на направление магнитного поля. Для частицы со спином $\frac{1}{2}n$ найдется $n + 1$ различных возможностей, которые будут для вертикально ориентированного поля майорановыми состояниями:

$$\left|\uparrow\uparrow\uparrow \ldots \uparrow\right\rangle,\ \left|\downarrow\uparrow\uparrow \ldots \uparrow\right\rangle,\ \left|\downarrow\downarrow\uparrow \ldots \uparrow\right\rangle,\ \ldots,\ \left|\downarrow\downarrow\downarrow \ldots \downarrow\right\rangle.$$

Каждое майораново направление соответствует стрелке вверх или стрелке вниз, но кратность этих направлений различается. Согласно стандартной терминологии, каждое из этих состояний обладает собственным m-значением, которое вычисляется так: половина количества стрелок, направленных вверх, минус половина числа стрелок, направленных вниз. В принципе, это и есть то самое m-значение, которое появляется при гармоническом анализе сферы, о котором пойдет речь в разделе А.11. Все эти $n + 1$ состояний ортогональны друг другу.

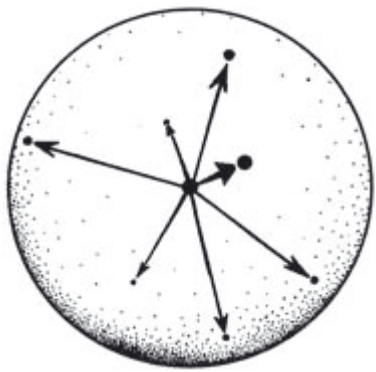


Рис. 2.21. Майораново описание для общего состояния спина частицы, обладающей массой, со спином $\frac{1}{2}n$ представляет это состояние как неупорядоченное множество из n точек на сфере Римана, причем каждую из этих точек можно считать направлением компонента «спин-$\frac{1}{2}$», вносящего вклад в общее состояние «спин-n»

[1] На самом деле по причинам технического характера такая процедура неприменима непосредственно к электронам [Mott and Massey, 1965], но успешно работает с различными атомами.

Рис. 2.22. В опыте Штерна — Герлаха используется сильно неоднородное магнитное поле, позволяющее измерять спин (точнее, магнитный момент) атома со спином $\frac{1}{2}n$ в выбранном направлении. Различные возможные результаты дают разные значения для компонентов спина в этом направлении, вида $-\frac{1}{2}n,\ -\frac{1}{2}n+1,\ -\frac{1}{2}n+2, ...,\ \frac{1}{2}n-2,\ \frac{1}{2}n-1,\ \frac{1}{2}n$ в единицах спина $\hbar$, соответствующих майорановым состояниям $|\downarrow\downarrow\downarrow ... \downarrow\rangle,\ |\uparrow\downarrow\downarrow ... \downarrow\rangle,\ |\uparrow\uparrow\downarrow ... \downarrow\rangle,\ ...,\ |\uparrow\uparrow ... \uparrow\downarrow\downarrow\rangle,\ |\uparrow\uparrow\uparrow ... \uparrow\downarrow\rangle,\ |\uparrow\uparrow\uparrow ... \uparrow\uparrow\rangle$. Очевидно, что аппарат можно вращать вокруг направления луча так, чтобы измерять различные выбранные нами направления спина, ортогональные направлению луча

Общее спиновое состояние $\frac{1}{2}n$, однако, не налагает никаких ограничений на соответствующие майорановы направления. Тем не менее сама идея Штерна и Герлаха позволяет определить фактические майорановы направления. Любое майораново направление $\nwarrow$ определяется следующим фактом: если измерение по Штерну — Герлаху выполняется в магнитном поле, ориентированном в направлении $\nwarrow$, то вероятность обнаружения состояния, в котором все стрелки ориентированы в противоположном направлении $|\searrow\searrow\searrow\searrow ... \searrow\rangle$, равна нулю [Zimba and Penrose, 1993].

2.10. Квантовая запутанность и ЭПР-эффекты

В квантовой механике есть очень примечательное семейство следствий, именуемых явлениями *Эйнштейна — Подольского — Розена* (ЭПР). Эти явления дают нам одни из самых загадочных и масштабных подтверждений стандартной квантовой теории. ЭПР-эффекты были сформулированы Альбертом Эйнштейном, пытавшимся показать, что аппарат квантовой механики содержит принципиальные ошибки или по меньшей мере неполон. Эйнштейн сотрудничал с двумя коллегами — Борисом Подольским и Натаном Розеном, вместе с которыми опубликовал ныне знаменитую статью «Можно ли считать квантово-механическое описание физической реальности полным?» [*Einstein* et al., 1935][1]. Фактически авторы указали, что из квантовой механики проис-

[1] Есть русский перевод: *Эйнштейн А.* Собрание сочинений. М.: Наука, 1966. Т. 3. С. 604–610. — *Примеч. пер.*

текает следствие, которое показалось им, как и многим другим людям (причем даже сегодня), неприемлемым. Следствие таково: пара сколь угодно удаленных друг от друга частиц все равно должна рассматриваться как цельная, связная система! Измерение, которому подвергнется любая из этих частиц, *немедленно* затронет и другую частицу, которая при этом окажется в квантовом состоянии, зависящем не только от результата измерения, проделанного с первой частицей, но и — что самое поразительное — от того, какое *именно* измерение мы выберем.

Для того чтобы как следует прочувствовать поразительную подоплеку такой ситуации, давайте внимательно рассмотрим спиновые состояния двух частиц, каждая из которых имеет спин $\frac{1}{2}$. Простейший пример ЭПР-парадокса относится именно к такому типу частиц и был предложен Дэвидом Бомом в книге по квантовой механике, которую он написал в 1951 году (по-видимому, как раз для того, чтобы убедить самого себя — оказалось, что безуспешно, — в полной справедливости квантового формализма [*Bohm*, 1951]). В примере Бома исходная частица со спином 0 расщепляется на две другие частицы: P_L и P_R; при этом обе они имеют спин $\frac{1}{2}$ и движутся в противоположных направлениях от исходной точки O до тех пор, пока, наконец, не попадают в детекторы, расположенные в точках L (слева) и R (справа), сильно удаленных друг от друга. При этом мы предполагаем, что каждый из двух детекторов может свободно и независимо вращаться и что выбор направления, в котором мы будем измерять спин на каждом детекторе, совершается лишь после того, как обе частицы окажутся в полете (рис. 2.23).

Рис. 2.23. ЭПР-эксперимент по измерению спина, предложенный Бомом. Две обладающие массой частицы (например, атомы) P_L и P_R, имеющие спин $\frac{1}{2}$, движутся в противоположных направлениях, начиная движение в исходном состоянии, характеризующемся суммарным спином 0, и разлетаются на существенное расстояние, где их спины независимо измеряются в приборах Штерна — Герлаха, установленных в точках L и R соответственно. Эти приборы могут вращаться независимо друг от друга в разных направлениях

Теперь оказывается, что если выбрать для обоих детекторов *одно и то же* направление, то результат измерения спина левоориентированной частицы P_L

в этом направлении будет противоположен результату соответствующего измерения спина правоориентированной частицы P_R. Это следствие сохранения момента импульса (см. раздел 1.14), поскольку исходное состояние обладает нулевым моментом импульса. Итак, если выбрано направление «вверх» ↑, то мы, обнаружив правоориентированную частицу в состоянии «вверх» $|\uparrow\rangle$ в точке R, предположили бы, что при подобном «вертикальном» измерении левоориентированной частицы в точке L непременно должны получить состояние «вниз» $|\downarrow\rangle$ в L; аналогично, обнаружив состояние «вниз» $|\downarrow\rangle$ в точке R, мы могли бы с уверенностью полагать, что в L нашли бы состояние «вверх» $|\uparrow\rangle$. Это было бы в равной степени справедливо для любого другого направления, скажем ↙, так что при измерении спина L-частицы в этом направлении мы бы обязательно получили ответ НЕТ, то есть состояние с противоположным направлением $|\nearrow\rangle$, при условии, что измерение R-частицы дало бы нам ответ ДА, то есть состояние $|\swarrow\rangle$; точно так же измерение L-частицы дает результат ДА, то есть состояние $|\swarrow\rangle$, если при измерении R-частицы получаем НЕТ, то есть $|\nearrow\rangle$.

В описанных явлениях пока нет ничего по-настоящему нелокального, хотя и кажется, что та картина, которая складывается в рамках этого стандартного квантового формализма, как представляется, противоречит нормальным ожиданиям локальной причинности. Можно задаться вопросом: каким образом эта картина явно противоречит локальности? Хорошо, давайте предположим, что измерение R делается всего на какой-то миг ранее измерения L. Если измерение R выявляет $|\swarrow\rangle$, то состояние L-частицы немедленно должно дать $|\nearrow\rangle$; если измерение R дает $|\nearrow\rangle$, то L-частица немедленно оказывается в состоянии $|\swarrow\rangle$. Можно было бы установить между двумя этими актами измерения столь краткий временной интервал, что за этот период даже световой сигнал, идущий между двумя точками, не успел бы передать информацию о состоянии L-частицы. Квантовая информация о том, в каком состоянии находится L-частица, нарушает стандартные требования теории относительности (рис. 2.23). Почему же в этом явлении нет «ничего подлинно нелокального»? *В сущности*, о нелокальности здесь речь не идет, так как мы можем с легкостью построить примитивную классическую модель, проявляющую именно такие свойства. Можно предположить, что каждая частица, покидая точку O, уже обладает информацией о том, как следует реагировать на любое измерение спина, которое могло бы быть над ней произведено. Чтобы ситуация согласовывалась с требованиями из предыдущего абзаца, нужно лишь допустить, что две частицы покидают точку O, имея *противоположные* инструкции для любого возможного направления, которое могло бы подвергнуться измерению. Для этого достаточно вообразить, что исходная частица со спином 0 содержит крошечную сферу, которая произвольным образом раскалывается на два полушария в тот самый момент, когда

эта частица делится на две другие частицы со спином $\frac{1}{2}$, и каждая из этих частиц уносит одно из полушарий по принципу поступательного движения (то есть ни одно из полушарий абсолютно не вращается). Полушарие, прикрепленное к каждой частице, ориентировано по отношению к направлению от центра так, чтобы, измеряя спин в соответствующем направлении, мы всегда получали ответ ДА. Легко убедиться в том, что такая модель всегда дает противоположные результаты при выборе любого направления и измерении спина одновременно у двух частиц — именно так и должно быть в соответствии с квантовыми соображениями, изложенными в предыдущем абзаце.

Этот пример выдающийся специалист по квантовой физике Джон Стюарт Белл сравнивал с аналогией, которую называл *носки Бертлманна* [Bell, 1981, 2004]. Рейнхольд Бертлманн, в настоящее время занимающий пост профессора физики в Венском университете, был коллегой Джона Белла в ЦЕРНе, и Белл заметил, что Бертлманн неизменно носит носки разного цвета. Не всегда получалось заприметить, какого именно цвета носок Бертлманна, но если сделать это все-таки удавалось и вы видели, что один носок зеленый, то вы — *мгновенно* — убеждались в том, что второй носок будет какого угодно цвета, но только не зеленого. Можно ли предположить, что какой-то таинственный флюид со сверхсветовой скоростью течет от одной ноги Бертлманна к другой, как только наблюдатель получает информацию о цвете одного из носков Бертлманна? Разумеется, нет. Все дело в том, что Бертлманн приучил нас к тому, что носки у него всегда разноцветные.

Однако в случае с парой частиц, каждая из которых имеет спин $\frac{1}{2}$ (пример Бома), ситуация радикально изменится, если мы допустим, что направления измерения спина на детекторах L и R могут варьироваться *независимо* друг от друга. В 1964 году Белл получил удивительный и фундаментальный результат, предполагавший, что никакая модель, подобная носкам Бертлманна, не позволяет объяснить совместные вероятности, существующие в квантовом формализме (включая стандартное правило Борна) для независимых измерений спина в L и R, для пар частиц со спином $\frac{1}{2}$, получаемых из общего квантового источника [Bell, 1964]. Белл фактически продемонстрировал, что существуют определенные соотношения (неравенства) между совместными результатами измерения спина при выборе разных направлений на L и R (ныне именуемые *неравенствами Белла*), которые непременно удовлетворяются в любой классической локальной модели, но *нарушаются* совместными квантово-механическими вероятностями по правилу Борна. Впоследствии ставились различные эксперименты [Aspect

et al., 1982; Rowe et al., 2001; Ma, 2009], и в настоящее время убедительно подтверждено, что они соответствуют ожиданиям квантовой механики, а нарушения неравенств Белла также детально подтверждены. На самом деле в этих экспериментах обычно работают с поляризацией фотона [Zeilinger, 2010], а не с состояниями спина для частиц со спином $\frac{1}{2}$, но, как мы видели в разделе 2.9, две ситуации формально идентичны.

Предлагалось множество теоретических примеров с ЭПР-экспериментами бомовского типа, и некоторые из них довольно просты. В этих экспериментах четко прослеживаются несоответствия между ожиданиями квантовой механики и ожиданиями классических локально реалистичных моделей (например, как в случае с носками Бертлманна) [Kochen and Specker, 1967; Greenberger et al., 1989; Mermin, 1990; Peres, 1991; Stapp, 1979; Conway and Kochen, 2002; Zimba and Penrose, 1993]. Однако я лучше не буду вдаваться в детали различных подобных примеров, а опишу один особенно поразительный эксперимент ЭПР-типа, поставленный Люсьеном Харди [1993], который несводим к ситуации Бома, но немного похож на нее. Эксперимент Харди обладает примечательным свойством: любые вероятности в нем могут принимать значения: 0 или 1 (то есть «не может произойти» или «определенно произойдет»), а также еще одно, о котором нам достаточно знать, что оно ненулевое (то есть «иногда происходит»). Как и в эксперименте Бома, две частицы со спином $\frac{1}{2}$ вылетают в противоположных направлениях из источника в точке O и летят к детекторам, которые находятся в достаточно удаленных точках L и R. Однако в примере Харди есть особенность: теперь в исходном состоянии (точка O) спин частицы равен не 0, а 1.

В конкретной версии эксперимента Харди, которую я описываю здесь (см. ПкР, раздел 23.5), исходному состоянию соответствуют два направления: ← (запад) и ↗ (северо-восток). Конкретные значения наклонов двух этих направлений понятны из рис. 2.24: направление ← в данном случае является горизонтальным (с отрицательной ориентацией), а у направления ↗ восходящий наклон 4/3 (с положительной ориентацией). Изюминка именно такого исходного состояния $\left|\leftarrow\nearrow\right\rangle$ заключается в том, что стоит нам только обнаружить, что

$$\left|\leftarrow\nearrow\right\rangle \textit{ не ортогонально паре } \left|\downarrow\right\rangle\left|\downarrow\right\rangle$$

(где ↓ — это юг, а → — восток), как мы также обнаруживаем, что:

$$\left|\leftarrow\nearrow\right\rangle \textit{ ортогонально} \text{ каждой из пар } \left|\downarrow\right\rangle\left|\leftarrow\right\rangle,\ \left|\leftarrow\right\rangle\left|\downarrow\right\rangle \text{ и } \left|\rightarrow\right\rangle\left|\rightarrow\right\rangle.$$

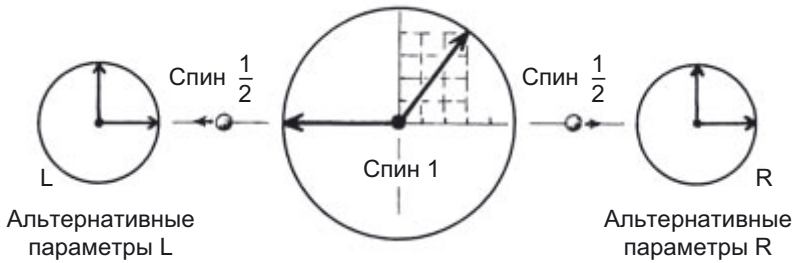


Рис. 2.24. Пример Харди: как и на рис. 2.23, но в исходном состоянии частица имеет спин 1, причем майорановы направления находятся под углом arctg(−4/3) друг к другу. Все рассматриваемые в данном случае вероятности могут быть равны 1 или 0, кроме одной, которая должна быть ненулевой (фактически здесь она равна 1/12)

Здесь под парой $|\alpha\rangle|\beta\rangle$ имеется в виду состояние, в котором $|\alpha\rangle$ находится в точке L, а $|\beta\rangle$ — в точке R. Благодаря сохранению момента импульса общий спин пары испущенных частиц целиком остается все в том же состоянии $|\leftarrow\nearrow\rangle$, пока не будет измерен спин у одной из частиц, поэтому соотношения ортогональности, характерные для исходного состояния, сохраняются и при последующих измерениях. (Ремарка для тех читателей, которых может смущать, что изображенные на рис. 2.23 устройства для измерения спина словно допускают только лишь вращение вокруг оси траектории полета частицы: отмечу, что все пространственные направления, критичные в данном примере, расположены в одной плоскости, поэтому эта плоскость может быть выбрана ортогональной траекториям частиц.)

Утверждение о неортогональности в первой из двух вышеприведенных посылок сообщает, что (i) если оба детектора для измерения спина, расположенные в точках L и R, будут настроены на ↓,то *иногда* (фактически — с вероятностью 1/12) действительно будут возникать ситуации, в которых мы получим на обоих детекторах ↓ (то есть ДА и ДА). Утверждение об ортогональности во второй из вышеприведенных посылок сообщает нам, что (ii) если один детектор настроен на регистрацию ↓, а другой — на регистрацию ←, то они не могут оба получить эти результаты (то есть как минимум один даст результат НЕТ). Наконец, из третьей посылки делаем вывод, что (iii) если оба детектора настроены на регистрацию ←, то они оба не могут получить противоположный результат →; иными словами, по меньшей мере на одном детекторе мы должны зафиксировать ← (то есть ДА).

Давайте проверим, можно ли построить классическую локальную модель (то есть объяснение в стиле «носков Бертлманна»), которая удовлетворяла бы

всем этим требованиям. Предположим, что серия частиц, вращающихся по часовой стрелке, вылетает из O по направлению к детекторам L и R. При этом частицы «запрограммированы» так, что при попадании в детектор дадут строго определенный результат, а сами результаты зависят от того, какой окажется ориентация каждого отдельного детектора, но механические компоненты, определяющие поведение каждой частицы, *не могут* обмениваться сигналами после того, как частицы покинули точку O. Так, наши частицы могут оказаться в ситуации, когда оба детектора настроены на регистрацию ←, поэтому если механизм должен давать результаты, соответствующие посылке (iii), то он должен работать так, чтобы либо одна, либо другая частица гарантированно обеспечила ответ ДА (то есть ←), как только попадет в детектор с ориентацией ←. Однако в данном случае может оказаться, что *другой* детектор будет настроен уже на регистрацию ↓, а посылка (ii) требует, чтобы частица, оказавшись в таком ↓-детекторе, обязательно дала ответ НЕТ (то есть мы должны получить ↑). Следовательно, при *каждом* запуске пары частиц из O либо одна, либо другая частица должна быть настроена на ответ ↑ при измерении, заточенном под регистрацию ↓. Но это не соответствует посылке (i), а именно утверждению о том, что *иногда* (в среднем с частотой 1/12), если оба детектора настроены на регистрацию ↓, мы *должны* получать сразу два ответа ДА (↓, ↓)! Следовательно, не существует возможности учесть все ожидания квантовой механики на уровне какой бы то ни было классической (по типу «носков Бертлманна») локальной инженерии.

Все это приводит нас к выводу о том, что зачастую отдельные квантовые объекты (независимо от того, как далеко они разнесены) по-прежнему остаются взаимосвязанными и ведут себя не как отдельные объекты. Квантовое состояние такой пары объектов называется *запутанностью* (термин Шрёдингера) — этот термин уже попадался нам в разделе 2.7 (а вскользь упоминался также в разделе 2.1). На самом деле такая запутанность не редкость в квантовой механике. Совсем наоборот: при взаимодействиях квантовых частиц (или систем, которые ранее не были запутаны) почти неизбежно возникает состояние квантовой запутанности. А поскольку частицы запутаны, они почти не имеют шансов вновь распутаться в ходе обычной унитарной эволюции (**U**).

Однако взаимная зависимость, существующая в паре далеко разнесенных квантовых объектов, находящихся в состоянии запутанности, — вещь довольно тонкая. Ведь оказывается, что запутанность гарантированно не позволяет передавать новую информацию от одного объекта к другому. Возможность передачи информации со сверхсветовой скоростью нарушала бы требования теории относительности. Квантовая запутанность — это явление, не имеющее аналогов в классической физике. Она обитает на странной квантовой «ничейной терри-

тории» между двумя классическими альтернативами: коммуникабельностью и полной независимостью.

Квантовая запутанность — очень тонкий эффект, и надо как следует постараться, чтобы зафиксировать такую запутанность, в чем мы уже убедились. Тем не менее квантово-запутанные системы, которые практически всегда возникают в результате квантовой эволюции, дают нам примеры *целостного* поведения, когда, строго говоря, целое больше суммы всех частей. Такое поведение является более тонким и отчасти таинственным по сравнению с обычными условиями, в которых мы вообще не замечаем эффектов квантовой запутанности. Это действительно очень загадочный вопрос: почему наблюдаемая нами Вселенная обладает такими целостными свойствами, которые, однако, почти никогда не проявляются. Я вернусь к этому вопросу в разделе 2.12, но прежде было бы поучительно рассмотреть, насколько обширно пространство запутанных состояний по сравнению с подмножеством тех состояний, которые не являются запутанными. Здесь мы вновь подходим к проблеме *функциональной свободы*, которая обсуждалась и будет обсуждаться в разделах А.2, А.8, 1.9, 1.10 и 2.2, но мы убедимся, что с ней связаны и некоторые другие фундаментально важные вопросы, касающиеся интерпретации функциональной свободы в квантовом контексте.

2.11. Квантовая функциональная свобода

Как говорилось в разделе 2.5, квантовое описание отдельной частицы — так называемая *волновая функция* этой частицы — довольно похоже на классическое поле с определенным числом независимых компонентов на точку пространства и (подобно магнитному полю) детерминированно эволюционирует со временем в соответствии с уравнениями поля. Уравнением поля волновой функции является уравнение Шрёдингера (см. раздел 2.4). Насколько много функциональной свободы в волновой функции? Согласно идеям и нотации, приведенным в разделе А.2, волновая функция одномерной частицы обладает функциональной свободой вида $\infty^{A\infty^3}$, где A — некоторое положительное число, а число 3 — это размерность обычного пространства.

В принципе, величина A — это число независимых компонентов поля (см. раздел А.2), но здесь есть дополнительная проблема: волновые функции комплексные, а не вещественные, поэтому логично ожидать, что A будет вдвое больше, чем число вещественных компонентов для классического поля. Однако есть и еще одна проблема, которую в этой книге я лишь вкратце затрагиваю (раздел 2.6): волновая функция свободной частицы должна описываться комплексной функ-

цией с *положительной частотой*, что фактически сокращает функциональную свободу вдвое и возвращает нас к той самой величине A, которую мы получили бы для классического поля. Наконец, еще один момент: умножение на произвольный коэффициент не изменяет физическое состояние, но в контексте функциональной свободы это не играет существенной роли.

Если бы у нас были две независимые частицы разного вида, причем у одной из них состояние функциональной свободы описывалось бы выражением $\infty^{A\infty^3}$, а у другой — $\infty^{B\infty^3}$, то свобода *незапутанных* квантовых состояний в этой паре была бы просто произведением двух этих значений свободы (поскольку любое состояние одной частицы может сопровождаться любым состоянием другой):

$$\infty^{A\infty^3} \times \infty^{B\infty^3} = \infty^{(A+B)\infty^3}.$$

Однако, как было показано в разделе 2.5, чтобы получить *все* разнообразные квантовые состояния, которые могут возникать у пары частиц, в том числе и запутанные, нужно иметь возможность указать отдельную амплитуду для каждой пары местоположений (где положения двух частиц варьируются независимо друг от друга), поэтому теперь наша волновая функция содержит вдвое больше переменных, нежели те три, которые у нас были раньше (то есть шесть). Более того, каждая пара значений соответствующих возможностей, обеспечиваемых A и B, учитывается отдельно (поэтому теперь общее число этих возможностей выражается произведением AB, а не суммой $A+B$). В контексте понятий, рассматриваемых в разделе А.2, наша волновая функция представляет собой функцию в *конфигурационном пространстве* (см. раздел А.6, рис. А.18) пары частиц, которое (без учета дискретных параметров, например таких, которые описывают состояния спина) является шестимерным произведением трехмерного пространства с самим собой (см. раздел А.7, особенно рис. А.25, где объясняется, что такое *произведение пространств*). Соответственно, такое пространство, в котором определяется наша двумерная волновая функция, является шестимерным, и в нем мы обнаруживаем несравнимо более высокую функциональную свободу:

$$\infty^{AB\infty^6}.$$

В случае, если частиц три, четыре и т. д., мы получим соответственно такие значения функциональной свободы:

$$\infty^{ABC\infty^9}, \quad \infty^{ABCD\infty^{12}} \text{ и т. д.}$$

В случае N идентичных частиц на функциональные степени свободы накладываются некоторые ограничения, связанные со статистикой Бозе — Эйнштейна и Ферми — Дирака и рассмотренные в разделе 1.14, согласно которым волновая функция должна быть соответственно симметричной или антисимметричной. Однако это не снижает функциональной свободы по сравнению с уровнем, который наблюдался бы без такого ограничения, а именно $\infty^{\Lambda N \infty^{3N}}$, поскольку это ограничение свидетельствует всего лишь о следующем: функция определяется некоторой подобластью произведения пространств такой же размерности, а значения в остальной области выводят исходя из требований симметрии или антисимметрии.

Как мы видим, функциональная свобода, связанная с квантовой запутанностью, полностью поглощает свободу незапутанных состояний. Возможно, читателя смутит тот поразительный факт, что, несмотря на ошеломляющее преобладание запутанных состояний среди результатов обычной квантовой эволюции, на уровне повседневного опыта мы, пожалуй, практически всегда игнорируем эффекты такой запутанности! Нужно попытаться понять это, казалось бы, колоссальное несоответствие и другие тесно связанные с ним проблемы.

Для того чтобы как следует заняться проблемами квантовой функциональной свободы, в частности вышеупомянутым, казалось бы, вопиющим несоответствием между квантовым формализмом и физическим опытом, давайте сделаем шаг назад и попытаемся понять, какую именно «реальность» на самом деле должен представить нам квантовый формализм. Здесь уместно вернуться к исходной ситуации, которая была описана в разделе 2.2 и с которой началась вся наша дискуссия о квантовой механике. Тогда мы уделили должное внимание принципу равнораспределения энергии: представляется, что частицы и излучение способны сосуществовать в состоянии термодинамического равновесия, лишь если физические поля и системы частиц являются в известном смысле равноценными объектами и обладают сопоставимой функциональной свободой. Вспомните об ультрафиолетовой катастрофе, упомянутой в разделе 2.2, которая случилась бы при наступлении равновесия между классическим (электромагнитным) полем и совокупностью (заряженных) классических частиц. В силу огромной разницы между функциональной свободой поля (здесь $\infty^{4\infty^3}$) и функциональной свободой системы частиц, описываемых с классической точки зрения (всего лишь ∞^{6N} для N частиц, не имеющих внутренней структуры, что гораздо меньше), при приближении к точке равновесия вся энергия частиц полностью утекла бы в обширный «резервуар» функциональной свободы поля — и наступила бы *ультрафиолетовая катастрофа*. Эту путаницу смогли разрешить Планк и Эйнштейн, выдвинув постулат, в соответствии с которым непрерывное, на первый взгляд, электромагнитное поле должно приобрести частицеподобный

квантовый характер в соответствии с формулой Планка $E = h\nu$, где энергия E соответствует моде колебаний поля с частотой ν.

Однако с учетом вышеизложенного мы теперь, по-видимому, вынуждены трактовать и сами частицы в их совокупности таким образом, как будто их можно описывать как единое целое, а именно как *волновую функцию* всей системы частиц; при этом система таких частиц обладает несравнимо большей функциональной свободой, чем классическая система частиц. На это следует обратить особое внимание, если учесть случаи запутанности между частицами, где функциональная свобода у N частиц имеет вид $\infty^{\bullet\infty^{3N}}$ (здесь • означает некое положительное число), тогда как свобода классического поля (при $N > 1$) будет значительно меньше $\infty^{\bullet\infty^{3}}$. Может показаться, что ситуация изменилась на противоположную и что теперь равнораспределение энергии свидетельствует о следующем: из-за чрезмерной свободы в системе квантовых частиц вся энергия поля полностью бы истощилась. Но эта проблема возникает лишь потому, что мы были непоследовательны и трактовали поле как классическую сущность, при этом используя квантовое описание частиц. Чтобы разрешить эту проблему, давайте разберемся, как именно с физической точки зрения нужно подсчитывать степени свободы системы согласно квантовой теории. Более того, мы вкратце рассмотрим и то, как квантовая теория интерпретирует физические поля согласно процедурам квантовой теории поля (КТП; см. разделы 1.3–1.5).

В принципе, допустимы два способа рассмотрения КТП в данном контексте. В основе современных теоретических подходов лежит процедура, при которой используются *интегралы по траекториям*. Она основана на идее Дирака, выдвинутой им в 1933 году [Dirac, 1933] и впоследствии развитой Фейнманом, который превратил ее в очень эффективный и мощный метод КТП [см. Feynman et al., 2010]. (Коротко и ясно эта идея изложена в книге ПкР (см. раздел 26.6).) Правда, при всей мощности и полезности эта процедура остается очень формальной (и строго говоря, не так уж непротиворечива с математической точки зрения). Как бы то ни было, эти формальные процедуры напрямую обеспечивают вычисления по фейнмановским диаграммам (вскользь упомянутым в разделе 1.5), а эти вычисления лежат в основе стандартных расчетов КТП, применяемых физиками для определения квантовых амплитуд, прогнозируемых теорией для простейших процессов рассеяния частиц. С точки зрения той проблемы, которую я здесь рассматриваю (это проблема функциональной свободы), следовало бы ожидать, что функциональная свобода в квантовой теории должна быть точно такой же, как и в классической теории, к которой применяется процедура квантования по интегралам траекторий. Действительно, вся процедура построена так, чтобы дать классическую теорию в качестве первого приближения, но с соответствующими квантовыми поправками (по-

рядка $\hbar$) к этой классической теории, и такие вещи вообще не повлияли бы на функциональную свободу.

В физике есть и более простой способ трактовки следствий КТП: достаточно допустить, что поле состоит из неопределенного количества частиц, а именно *квантов поля* (в случае электромагнитного поля такими частицами являются фотоны). Общая амплитуда (то есть полная волновая функция) представляет собой *сумму* — квантовую суперпозицию — различных частей, каждая из которых соответствует разному количеству частиц (то есть квантов поля). Часть, содержащая N частиц, давала бы нам частную волновую функцию с функциональной свободой вида $\infty^{\bullet\infty^{3N}}$. Однако число N следует считать неопределенным, поскольку кванты поля, в нашем случае фотоны, постоянно рождаются и уничтожаются в результате взаимодействий с источниками поля — электрически заряженными или намагниченными частицами. Как раз поэтому наша общая волновая функция должна представлять собой суперпозицию частей с различными значениями N. Теперь, если попытаться рассмотреть функциональную свободу, присущую каждой частной волновой функции, и сделать это точно так же, как в случае с классической системой, причем постараться применить равнораспределение энергии, как это делалось в разделе 2.2, то мы столкнемся с некоторыми серьезными сложностями. Функциональная свобода сравнительно больших совокупностей квантов поля совершенно поглотила бы функциональную свободу более мелких совокупностей (поскольку $\infty^{\bullet\infty^{3M}}$ несоизмеримо больше $\infty^{\bullet\infty^{3N}}$, когда $M > N$). Если бы мы стали рассматривать эту функциональную свободу волновой функции точно так же, как и функциональную свободу классической системы, то обнаружили бы, что для системы, находящейся в равновесии, в соответствии с теоремой о равнораспределении энергии вся энергия пошла бы на вклад в то состояние, где накапливалось бы все больше и больше частиц, полностью перекрыв вклад в состояние с любым конечным числом частиц, что опять привело бы к катастрофической ситуации.

Вот здесь мы и упираемся в проблему: как формализм квантовой механики соотносится с физическим миром. Мы не можем считать, что функциональная свобода волновых функций зиждется на тех же основах, что и функциональная свобода в классической физике, несмотря на то что (обычно чрезвычайно запутанная) волновая функция оказывает безусловное, хотя и очень тонкое влияние на физические явления. Функциональная свобода по-прежнему играет ключевую роль в квантовой механике, но ее нужно рассматривать вместе с ключевой идеей, выдвинутой Максом Планком в 1900 году и заложенной в знаменитой формуле

$$E = h\nu,$$

за которой последовали глубокие озарения Эйнштейна, Бозе, Гейзенберга, Шрёдингера, Дирака и др. Формула Планка говорит нам, что поля, которые существуют в природе, характеризуются некоторой дискретностью, благодаря чему их поведение становится похожим на поведение системы частиц. При этом более высокочастотные моды колебаний, содержащиеся в этом поле, соответствуют более высоким энергиям. Следовательно, согласно квантовой механике, такие поля, которые в природе представляют собой проявления волновых функций, совершенно не похожи на классическое поле, описанное, например, в разделе A.2 и там же проиллюстрированное, в частности, на примере классического магнитного поля. При очень высоких энергиях квантовое поле начинает проявлять дискретные, или корпускулярные, свойства.

В данном контексте было бы правильно считать, что квантовая физика позволяет описать некую зернистую структуру фазового пространства $\mathcal{P}$ в системе (см. раздел A.6). В сущности, было бы неточно утверждать, что при этом континуум пространства-времени уступает место чему-то дискретному, наподобие дискретной упрощенной модели Вселенной, где континуум $\mathbb{R}$ вещественных чисел можно заменить конечной системой **R** (рассматривается в разделе A.2), состоящей из исключительно большого целого числа элементов N. Тем не менее подобная картина уже не будет совершенно неприменимой к фазовым пространствам, играющим важную роль в квантовых системах. Как подробнее объясняется в разделе A.6, фазовое пространство $\mathcal{P}$ для системы из M точечных классических частиц, обладающих M координатами положения и M координатами импульса, содержало бы $2M$ измерений. Единица объема — $2M$-мерного гиперобъема — таким образом, содержала бы M мер расстояния, которые можно было бы выразить, например, в метрах (м), и M мер импульса, выраженных, например, в килограммах (кг), умноженных на метр в секунду (м·с$^{-1}$). Соответственно, наш гиперобъем рассматривался бы в терминах M-й степени произведения этих величин, а именно кг^{M}·м^{2M}·с$^{-M}$, которое бы зависело от конкретных выбранных нами единиц. Однако в квантовой механике у нас есть естественная единица — *постоянная Планка* h, причем более удобно использовать ее редуцированную версию, называемую *постоянной Дирака* $\hbar = h/2\pi$, значение которой составляет:

$$\hbar = 1{,}05457\ldots \times 10^{-31} \quad \text{кг м}^2\text{с}^{-1}.$$

Величина $\hbar$ позволяет получить естественную меру гиперобъема для $2M$-мерного пространства $\mathcal{P}$, а именно использовать единицы $\hbar^M$.

В разделе 3.6 мы поговорим о *естественных единицах* (они также называются планковскими), которые были выбраны так: различные фундаментальные

константы природы просто брались за единицу. Не обязательно переводить все единицы в естественные, но мы можем сделать это по меньшей мере для массы, длины и времени так, чтобы

$$\hbar = 1$$

(это можно делать самыми разными способами, и вдобавок полный набор естественных единиц предлагается в разделе 3.6). В таком случае выясняется, что *любой* гиперобъем фазового пространства попросту выражается некоторым *числом*. Действительно, можно предположить, что $\mathcal{P}$ присуща некая естественная зернистость, где каждая отдельная ячейка, или гранула, пространства берется за единицу в контексте таких физических единиц, где $\hbar = 1$. Соответственно, объем в фазовом пространстве всегда будет иметь некое целочисленное значение, получить которое можно, просто подсчитав число ячеек в этом объеме. Ключевой момент в данном случае заключается в следующем: теперь мы можем непосредственно сравнивать гиперобъемы фазовых пространств, имеющие *разную размерность* $2M$, просто подсчитывая ячейки и не обращая внимания на значение M.

Почему это важно для нас? Потому что при работе с квантовым полем, находящимся в равновесии с системой частиц, взаимодействующих с полем и позволяющих изменять количество квантов поля, нам требуется способ сравнивать гиперобъемы фазовых пространств различных размерностей. Однако если говорить о классических фазовых пространствах, гиперобъемы высших размерностей тотально превосходят объемы с малым числом измерений (например, трехмерный объем обычной плавной кривой в евклидовом трехмерном пространстве всегда *нулевой* независимо от длины этой кривой); следовательно, такие состояния с бóльшим числом измерений свободы полностью бы вытянули энергию из состояний с меньшим их числом согласно требованиям равнораспределения энергии. Зернистость, обеспечиваемая квантовой механикой, позволяет решить эту проблему: измерение объемов сводится к обычному *подсчету*, так что гиперобъемы высших размерностей получаются гораздо больше гиперобъемов низших размерностей, но не *бесконечно* больше.

Все это непосредственно касается ситуации, с которой Макс Планк столкнулся в 1900 году. Итак, есть состояние, которое, на наш взгляд, состоит из сосуществующих компонентов, причем каждый компонент содержит различное количество квантов поля; эти кванты мы называем *фотонами*. Для любой конкретной частоты υ революционный планковский принцип (раздел 2.2) подразумевает, что фотон с такой частотой должен обладать соответствующей энергией согласно формуле

$$E = h\nu = 2\pi\hbar\nu.$$

Именно такой процедурой подсчета воспользовался прежде малоизвестный индийский физик Шатьендранат Бозе, отправивший Эйнштейну письмо в июне 1924 года. В этом письме он показал, как напрямую вывести планковскую формулу излучения (без привлечения какой-либо электродинамики), причем дополнительно к $E = h\nu$ и тому факту, что количество фотонов не должно быть фиксированным (число фотонов не сохраняется), он указывал всего лишь следующие требования: у фотона должно быть два разных состояния (см. разделы 2.6 и 2.9) и, что наиболее важно, фотоны должны подчиняться так называемой *статистике Бозе* (или *статистике Бозе — Эйнштейна*; см. раздел 1.14), чтобы состояния, отличающиеся друг от друга только перестановкой пар фотонов, не считались разными с физической точки зрения. Две последние характеристики были революционными для своего времени, и Бозе по праву увековечили, назвав в его честь *бозон* — это любая элементарная частица с целочисленным спином (которая благодаря этому свойству подчиняется статистике Бозе).

Существует и другое обширное семейство элементарных частиц, но с полуцелым спином — это фермионы, названные так в честь итальянского физика-ядерщика Энрико Ферми. Они подсчитываются немного иначе — по правилам *статистики Ферми — Дирака*. Она отчасти напоминает статистику Бозе — Эйнштейна, но в нее не включаются состояния, в которых содержатся две (или более) частицы одного рода, состояние которых при этом совпадает (принцип запрета Паули). В разделе 1.4 подробнее рассказано, как бозоны и фермионы трактуются в стандартной квантовой механике (читатель может не обращать внимания на описанную там экстраполяцию стандартной теории на спекулятивную, но все-таки исключительно модную картину суперсимметрии, принятую в физике частиц).

С учетом всех этих оговорок идея функциональной свободы хорошо сочетается как с квантовыми системами, так и с классическими, но здесь необходима осторожность. Величина ∞, фигурирующая в наших выражениях, на самом деле не бесконечна; вполне можно считать, что в обычных обстоятельствах она выражалась бы просто очень большим числом. Не вполне очевидно, что делать с проблемой функциональной свободы в общем квантовом контексте, особенно с учетом того, что находящаяся в квантовой суперпозиции система может содержать различные компоненты; эти компоненты содержали бы разное количество частиц и с классической точки зрения были бы сопряжены с фазовыми пространствами разных размерностей.

Однако если говорить об излучении, находящемся в термодинамическом равновесии с заряженными частицами, то можно вернуться к рассуждениям Планка, Эйнштейна и Бозе. Согласно этим рассуждениям, получаем планковскую фор-

мулу *интенсивности* излучения (в равновесии с материей) для каждой частоты ν (см. раздел 2.2):

$$\frac{8\pi h\nu^3}{c^3(e^{h\nu/kT}-1)}.$$

В разделе 3.4 будет показано, какое огромное космологическое значение имеет эта формула и сколь хорошо она согласуется со спектром космического микроволнового фонового излучения.

В главе 1 (конкретно в разделах 1.10 и 1.11) обсуждалась проблема роли дополнительных пространственных измерений. Мы говорили о том, насколько правдоподобно было бы увеличить число пространственных измерений до той величины, которой требует теория струн (сверх трех непосредственно наблюдаемых измерений). Сторонники таких теорий о высших размерностях порой утверждают, что по причинам квантового характера избыточная функциональная свобода не могла бы непосредственно влиять на физические явления, наблюдаемые в обычных условиях, поскольку для возбуждения этих степеней свободы потребовалась бы исключительно высокая энергия. В разделах 1.10 и 1.11 я указывал, что при рассмотрении степеней свободы в геометрии пространства-времени (то есть гравитации) эта точка зрения получается (в лучшем случае) весьма спорной. Но я не касался отдельной проблемы избыточной функциональной свободы, которая возникала бы под действием *не*гравитационных полей, например электромагнитного, то есть *материальных* полей, которые можно было бы считать *наполнением* этих дополнительных пространственных измерений. Поэтому интересно рассмотреть, насколько наличие такой пространственной надразмерности могло бы повлиять на космологическое использование вышеупомянутой формулы.

При наличии дополнительных пространственных измерений — допустим, всего их будет D (в традиционной теории струн Шварца — Грина $D = 9$) — *интенсивность* излучения как функция частоты v описывалась бы формулой

$$\frac{Qh\nu^D}{c^D(e^{h\nu/kT}-1)},$$

где Q — числовая константа (зависящая от D). Сравните эту формулу с вышеприведенной, касающейся трех измерений (см. [Cardoso and de Castro, 2005]). На рис. 2.25 мы сравниваем случай $D = 9$ с традиционным планковским $D = 3$, рассмотренным ранее на рис. 2.2. Однако, учитывая крайний дисбаланс между масштабами пространственной геометрии в различных направлениях, вряд ли кто-то ожидал бы, что такая формула могла бы иметь какое-то прямое космо-

логическое значение. Тем не менее в таких моделях на самых ранних стадиях развития Вселенной, в эпоху порядка планковского времени (около 10^{-43} с) или чуть позже, *все* пространственные измерения были искривлены примерно в таких масштабах, что обеспечивало некоторое равенство между всеми девятью гипотетическими пространственными измерениями. Поэтому можно предположить, что в самом начале времен такая версия планковской формулы для высших размерностей действительно могла играть важную роль.

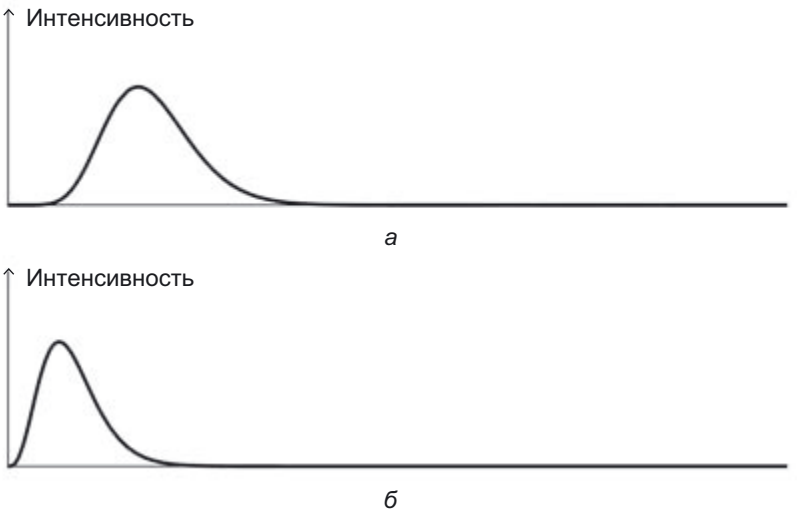


Рис. 2.25. Форма планковского спектра: *а*) для девяти пространственных измерений ($\text{const} \times \nu^9 (e^{h\nu/kT} - 1)^{-1}$); *б*) обычный случай для трех измерений ($\text{const} \times \nu^3 (e^{h\nu/kT} - 1)^{-1}$)

Как мы убедимся в разделах 3.4 и 3.6, тогда, на заре времен, должен был существовать еще один крайний дисбаланс, а именно между гравитационными степенями свободы и степенями свободы всех остальных полей. Тогда как гравитационные степени свободы, по-видимому, вообще *не* возбуждались, степени свободы остальных полей, кажется, возбуждались по максимуму! Такая ситуация наблюдалась как минимум в эпоху *разделения вещества и излучения* (последнего рассеяния), наступившую примерно через 380 000 лет после Большого взрыва. О таком исключительном дисбалансе напрямую свидетельствует сама природа космического микроволнового фона — это будет показано в разделах 3.4 и 3.6. Выясняется, что степени свободы вещества и излучения находились в *высокотермальном* (то есть максимально возбужденном) состоянии, а вот степени свободы гравитационного поля, то есть касающиеся геометрии пространства-времени, по-видимому, совершенно не вовлекались во все эти процессы. Сложно

представить себе, как такой дисбаланс мог возникнуть всего за 380 000 лет, прошедших со времени Большого взрыва, поскольку логично было бы ожидать, что термодинамическое равновесие в течение этого времени могло эволюционировать только в сторону более равновесных состояний — это напрямую следует из второго закона термодинамики (см. раздел 3.3). Поэтому напрашивается вывод: такая инертность гравитационных степеней свободы должна была возникнуть именно на самых ранних этапах (в период планковского времени — около 10^{-43} с), а гравитационные степени свободы вступают в игру на гораздо более позднем этапе (существенно позже эпохи последнего рассеяния) и обусловлены возникшей впоследствии неравномерностью распределения вещества.

Правда, тут уместен вопрос: могла ли вышеприведенная формула для высших измерений ($D = 9$), даже если считать, что она действовала лишь на протяжении кратчайшего периода сразу после Большого взрыва, сохранить какое-то остаточное влияние вплоть до времени последнего рассеяния (380 000 лет после Большого взрыва), когда и возникло то самое излучение, которое мы сегодня наблюдаем в виде реликтового микроволнового фона (см. раздел 3.4)? Фактически микроволновое реликтовое излучение имеет немного иной спектр интенсивности, чем можно было бы ожидать исходя из вышеприведенного выражения для высших размерностей (см. рис. 2.25), и исключительно точно согласуется с вариантом $D = 3$ (см. раздел 3.4), поэтому можно принять, что, когда Вселенная стала расширяться, спектр излучения $D = 9$ должен был целиком превратиться в вариант $D = 3$. Кривая отражает распределение частот, при котором достигается максимальная энтропия, то есть максимально равномерное распределение энергии материальных полей по доступным степеням свободы, *с учетом* геометрии того пространства-времени, в котором находятся эти поля. Если все пространственные измерения расширялись с одинаковой скоростью, то сохранялся бы спектр, соответствующий $D = 9$, и энтропия излучения оставалась бы более или менее постоянной, держась вблизи неимоверно более высокого значения, которое выводилось бы из вышеприведенной формулы для $D = 9$.

Однако наблюдаемый ныне реликтовый микроволновый фон соответствует $D = 3$, и с точки зрения второго закона термодинамики (см. раздел 3.3) гипотетические исключительно возбужденные степени свободы материальных полей, обусловленные наличием шести дополнительных измерений, должны куда-то деться; возможно, они пошли на активацию крошечных дополнительных шестимерных пространственных измерений в виде гравитационных либо материальных степеней свободы. Как бы то ни было, мне сложно увязать эту картину со струнно-теоретической точкой зрения, которая заключается в том, что дополнительные шесть измерений сейчас находятся на стабильном минимуме (см. раздел 1.11 и 1.14). Как могло получиться, что при исключительно

высокой температуре степени свободы материи в первозданной Вселенной с девятью пространственными измерениями почему-то улеглись именно так, что шесть дополнительных измерений словно остались совершенно невозбужденными, ведь, по-видимому, именно этого требует теория струн? Также следует спросить, какая гравитационная динамика могла дать столь огромное несоответствие между различными пространственными измерениями, и в особенности задуматься о том, как шесть свернутых невозбужденных измерений могли так аккуратно отделиться от трех расширяющихся.

Я не утверждаю, что в этих выкладках есть очевидное противоречие, но картина получается определенно очень странная, и ее требуется как-то объяснить в динамике. Хочется надеяться, что из всего этого удастся выжать побольше количественной информации. Происхождение столь колоссального дисбаланса между двумя классами пространственно-временных измерений, естественно, само по себе представляет огромную загадку струнно-теоретической картины мира. Можно задуматься и о том, каким образом на том этапе не произошло достаточной термализации *гравитационных* степеней свободы, а вместо этого возникла только лишь очень интересная система четко разграниченных измерений двух типов, в пользу которой, по-видимому, выступает современная теория струн, предусматривающая существование высших размерностей.

2.12. Квантовая реальность

Согласно стандартной квантовой механике, для вероятностного прогнозирования результатов экспериментов, проводимых в некоторой системе, нужна информация о квантовом состоянии этой системы, или *волновая функция* ψ. Однако, как мы убедились в разделе 2.11, волновая функция связана с такой функциональной свободой, которая гораздо выше, чем проявляется в наблюдаемой реальности или как минимум в том аспекте реальности, который считается результатом квантового измерения. Следует ли нам считать, что волновая функция ψ действительно отражает физическую реальность? Или считать ψ просто вычислительным инструментом для работы с вероятностными исходами результатов тех экспериментов, которые *могли бы* быть выполнены, и в таком случае результаты получаются «вещественными», а сама волновая функция — нет?

Как было упомянуто в разделе 2.4, в копенгагенской интерпретации квантовой механики за основу берется именно последний подход (также и в других научных школах) и ψ считается вычислительной уловкой без какого-либо онтологического статуса, только как «элемент мыслительных процессов экспериментатора или теоретика, обеспечивающий вероятностную оценку конкретных результатов

эксперимента». Представляется, что такие убеждения во многом возникают из-за резко отрицательного отношения физиков к самой возможности внезапных «скачков» тех или иных состояний реального мира в казалось бы непредсказуемых направлениях, а для законов квантового измерения свойственно именно это (см. разделы 2.4 и 2.8). Вспомните, с каким отчаянием Шрёдингер высказывался на эту тему (см. раздел 2.8). Как упоминалось ранее, в копенгагенской интерпретации предлагается считать, что все эти скачки происходят «в уме», поскольку чье-либо представление о мире действительно может мгновенно измениться, как только выяснятся новые факты об этом мире (то есть результат эксперимента).

На этом перекрестке я должен обратить внимание читателя на другую точку зрения, альтернативную копенгагенской; она называется *теория де Бройля — Бома* [de Broglie, 1956; Bohm, 1952; Bohm and Hiley, 1993]. Здесь я буду именовать ее распространенным термином *бомовская механика*. Она предлагает интересную альтернативную онтологию, отличную от той, что дается (или *не* дается!) в копенгагенской интерпретации. Бомовская механика широко изучается, хотя, конечно, не может считаться модной теорией. В ней не предлагается никаких альтернативных наблюдательных эффектов, которые отсутствовали бы в традиционной квантовой механике, но гораздо четче очерчивается «реальная картина» мира. Если коротко, в бомовской картине присутствуют *два* уровня онтологии. Более слабый из двух уровней описывается универсальной *волновой функцией* ψ (она также называется *волна-пилот*). Наряду с волновой функцией ψ также учитывается конкретное местоположение каждой частицы, обозначаемое как точка P в *конфигурационном пространстве* $\mathcal{C}$ (см. раздел A.6), которое можно представить как $\mathbb{R}^{3n}$, если предполагается, что существует n (различимых скалярных) частиц в плоском фоновом пространстве-времени. Мы считаем, что ψ — это функция в $\mathcal{C}$ и ее значение выражается комплексным числом; это удовлетворяет уравнению Шрёдингера. Но сама точка P, то есть местоположение конкретной частицы, в бомовском мире является элементом более *прочной* «реальности». Частицы обладают строго определенной динамикой, описываемой ψ (поэтому и ψ приходится присвоить некоторую реалистичность, пусть и более слабую, чем P). Функция ψ не испытывает никакого обратного воздействия со стороны местоположений частиц (описываемых P). Так, в эксперименте с двумя щелями, описанном в разделе 1.4, каждая из частиц действительно проходит или через одну, или через другую щель, но именно в функции ψ содержится информация об альтернативном маршруте, направляющая частицы таким образом, что на экране складывается правильный дифракционный рисунок. Однако как ни интересна эта версия с философской точки зрения, она не играет ключевой роли в моей книге, поскольку ее предсказания идентичны предсказаниям традиционной квантовой механики.

Даже устоявшаяся копенгагенская интерпретация на самом деле не избавлена от следующей проблемы: в ней нам приходится принимать ψ в качестве фактического воплощения чего-то «реального», существующего в мире. Один из аргументов в пользу такой реалистичности проистекает из принципа, предложенного Эйнштейном вместе с его коллегами Подольским и Розеном в их знаменитой статье об эффектах ЭПР, упоминавшейся в разделах 2.7 и 2.10. Эйнштейн настаивал, что в квантовом формализме присутствует «элемент реальности», если этот формализм с совершенной *определенностью* подразумевает некоторый результат измерения.

> В полной физической теории существует определенный элемент, соответствующий каждому элементу реальности. Достаточным условием реальности той или иной физической величины является возможность предсказать ее с достоверностью, не нарушая систему... *Если мы можем без какого бы то ни было возмущения системы предсказать с достоверностью (то есть с вероятностью, равной единице) значение некоторой физической величины, то существует элемент физической реальности, соответствующий этой физической величине.*

Однако в рамках стандартного квантового формализма существует такой феномен: в принципе, для любого вектора квантового состояния, допустим для $|\psi\rangle$, можно выполнить такое измерение, при котором $|\psi\rangle$ будет единственным вектором состояния, который с точностью до коэффициента пропорциональности гарантированно дает результат ДА. Почему так? С математической точки зрения нам просто требуется найти такое измерение, при котором один из векторов ортогонального базиса $\boldsymbol{\varepsilon}_1$, $\boldsymbol{\varepsilon}_2$, $\boldsymbol{\varepsilon}_3$, ... (упоминались в разделе 2.8), допустим $\boldsymbol{\varepsilon}_1$, фактически является заданным вектором состояния $|\psi\rangle$, и измерение должно дать результат ДА, если будет найден $\boldsymbol{\varepsilon}_1$, либо НЕТ, если будет найден любой из векторов $\boldsymbol{\varepsilon}_2$, $\boldsymbol{\varepsilon}_3$, ... Это крайний пример вырожденного измерения (о таких измерениях см. в конце раздела 2.8). Те, кто знаком с нотацией Дирака для операторов стандартной квантовой механики (см. раздел 2.9) [Dirac, 1930], поймут, что это измерение выполняется при помощи эрмитова оператора $\mathbf{Q} = |\psi\rangle\langle\psi|$ для любого заданного нормированного $|\psi\rangle$, где ДА соответствует собственному значению 1, а НЕТ — 0. Волновая функция ψ (с точностью до некоего ненулевого комплексного коэффициента) будет единственным образом определяться тем условием, что при этом измерении она обязательно должна давать ответ ДА. Поэтому в соответствии с вышеуказанным эйнштейновским принципом приходим к выводу, что действительно в любой волновой функции ψ так или иначе должен присутствовать несомненный элемент реальности!

Правда, может оказаться, что на практике совершенно невозможно сконструировать измерительное устройство именно такого типа, но общий аппарат

квантовой механики постулирует, что такое измерение должно быть возможным. Более того, экспериментатору, естественно, потребовалось бы заранее знать, какова функция ψ, чтобы он понимал, какое измерение делать. Но, в принципе, эту информацию можно было бы теоретически выяснить исходя из шрёдингеровской эволюции некоего ранее измеренного состояния. Следовательно, принцип Эйнштейна приписывает *элемент реальности* любой волновой функции, которая была вычислена по эволюционному уравнению Шрёдингера, то есть по **U**, исходя из некоего ранее известного состояния (оно могло быть выяснено по результатам какого-то более раннего измерения). При этом предполагается, что уравнение Шрёдингера (то есть унитарная эволюция) действительно является истинным в окружающем мире, как минимум в рассматриваемых нами квантовых системах.

Хотя для огромного множества потенциальных ψ было бы невозможно сконструировать измерительное устройство (при современном уровне технологического развития), существует немало экспериментальных ситуаций, в которых это *совершенно* осуществимо. Соответственно, было бы полезно исследовать пару именно таких ситуаций. Первая возникает при измерении спина у частиц с полуцелым спином — или, допустим, у атома с полуцелым спином и с магнитным моментом, ориентированным в том же направлении, что и спин. Можно воспользоваться прибором Штерна — Герлаха (см. раздел 2.9 и рис. 2.22), ориентированным в некотором направлении $\leftarrow$, чтобы измерить спин атома в этом направлении. Если получим ответ ДА, то придем к выводу, что спин атома действительно обладает вектором состояния (с точностью до коэффициента пропорциональности) $|\leftarrow\rangle$. Допустим, далее мы соотнесем это состояние с известным магнитным полем и при помощи шрёдингеровского уравнения вычислим, что через секунду состояние этого поля превратится в $|\nearrow\rangle$. Будем ли мы считать такое результирующее состояние спина «реальным»? Это определенно кажется оправданным, поскольку вращающийся аппарат Штерна — Герлаха, настроенный на измерение спина в направлении $|\nearrow\rangle$, в этот момент действительно гарантированно получит результат ДА. Разумеется, это очень простая ситуация, но она явно распространяется и на другие, значительно более сложные случаи.

Но бывают и другие, несколько более загадочные ситуации, связанные с квантовой запутанностью, и мы вполне можем рассмотреть различные примеры, в которых проявляются ЭПР-эффекты, например такие, как были продемонстрированы в разделе 2.10. Для определенности давайте рассмотрим пример Харди, на котором можно предположить, что мы создали заранее подготовленное состояние со спином 1, майораново описание которого (раздел 2.9) имеет вид $|\leftarrow\nearrow\rangle$, о чем подробно рассказано в разделе 2.10. Допустим, что это состо-

яние распадается на два атома со спинами $\frac{1}{2}$, причем один атом движется влево, а другой — вправо. Мы видели, что в данном примере не может быть непротиворечивого варианта наблюдения, при котором мы могли бы присвоить *независимые* квантовые состояния каждому из двух этих атомов отдельно. При любом таком присваивании мы обязательно получим неверные результаты при возможных измерениях спина, которые кто-либо может выполнить для левоориентированных и правоориентированных атомов. Определенно, существует квантовое состояние, применимое к обоим атомам, но это состояние *запутанное*, и оно касается *пары* атомов, а не каждого из атомов в отдельности. Можно произвести и такое измерение, которое действительно подтвердит данное запутанное состояние, — и функция ψ, рассмотренная выше, окажется в таком запутанном состоянии, в которое вовлечены две частицы. При таком измерении, возможно, придется каким-то образом вновь собрать оба атома вместе и выполнить измерение, которое подтвердило бы исходное состояние $|\leftarrow\nearrow\rangle$. Вполне вероятно, что это будет сложно осуществить технически, но в принципе такое возможно. В результате мы присвоим эйнштейновский «элемент реальности» запутанному состоянию, в котором пребывает разделенная пара. Однако для этого может быть недостаточно попросту измерить спины независимо, допустим, при помощи двух отдельных приборов Штерна — Герлаха (рис. 2.26 *a*), каждый из которых измерял бы спин всего одного из двух атомов. Квантовое состояние непременно распространяется на оба атома, которые пребывают в запутанном виде.

С другой стороны, предположим, что спин левоориентированного атома действительно измеряется прибором Штерна — Герлаха независимо от правоориентированного атома. В результате правоориентированный атом немедленно перейдет в состояние с конкретным спином. Допустим, например, что состояние левоориентированного атома измеряется в направлении $\leftarrow$ и мы получаем ответ ДА $|\leftarrow\rangle$; тогда оказывается, что правоориентированный атом автоматически попадет в состояние со спином $|\uparrow\rangle$. Если же, наоборот, при измерении левоориентированного атома мы получим ответ НЕТ, то правоориентированный атом автоматически окажется в состоянии $|\leftarrow\rangle$ (здесь применяется нотация, описанная в разделе 2.10). Все эти любопытные следствия напрямую проистекают из свойств примера Харди, приведенного в разделе 2.10.

В любом случае гарантируется, что *после* измерения левоориентированной частицы состояние частицы с правоориентированным спином обязательно приобретет конкретное независимое значение. Как мы могли бы это подтвердить? Постулируемое здесь состояние правоориентированной частицы можно подтвердить по меньшей мере при помощи соответствующего измерения прибором

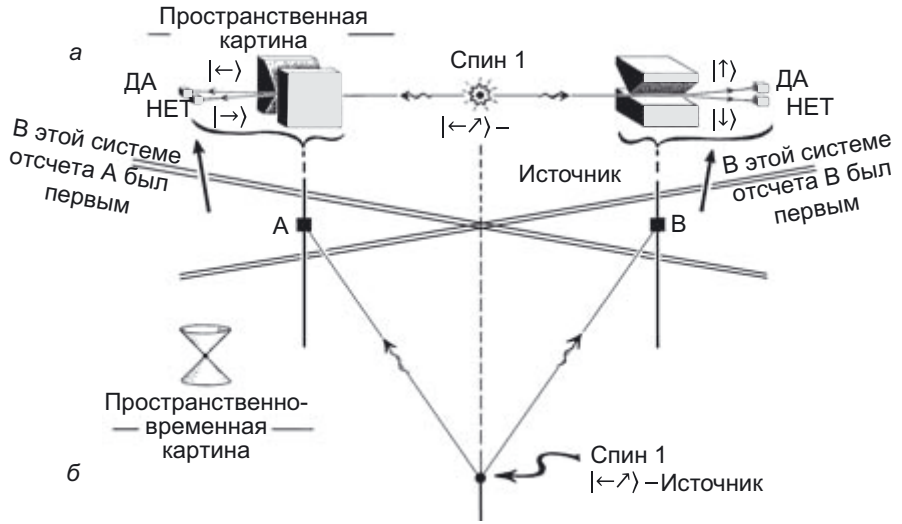


Рис. 2.26. Нелокальный эксперимент Харди (см. рис. 2.24) в пространственном представлении (*а*), но в сочетании с экспериментом, показанным в пространственно-временном контексте (*б*), представляет проблему на уровне объективного пространственно-временного описания реальности

Штерна — Герлаха. Однако такое измерение правоориентированного атома не продемонстрировало бы наличие значения (в данном случае $|\uparrow\rangle$) у состояния с правоориентированным спином, если бы при измерении спина левоориентированной частицы обнаружился результат ДА. С другой стороны, не подтвердилось бы значение $|\leftarrow\rangle$, если бы при измерении спина левоориентированной частицы был зафиксирован результат НЕТ. «Реальность» данного конкретного правоориентированного состояния теперь кажется подтвержденной: то есть имеем $|\uparrow\rangle$, если левоориентированное измерение ← даст результат ДА, и $|\leftarrow\rangle$, если получим результат НЕТ. Допустим, левоориентированное измерение ← действительно дает результат ДА. Соответствующий результат ДА для правоориентированного измерения ↑ не гарантирует, что спин правоориентированного состояния действительно был $|\uparrow\rangle$, поскольку ответ ДА, полученный «справа», мог получиться просто случайно. Если мы получим при правоориентированном измерении ← ответ ДА, то этот результат убедит нас лишь в том, что измеренное состояние не равно $|\downarrow\rangle$. Любое другое возникающее состояние спина, приближенное к $|\downarrow\rangle$, могло бы лишь с небольшой вероятностью привести к результату ДА в ответ на ↑-измерение, причем эта вероятность возрастала бы тем сильнее, чем ближе к $|\uparrow\rangle$ оказывалось бы измерение правоориентированного состояния. Чтобы убедительно подтвердить в эксперименте, что возникающее правоориентированное

состояние действительно было равно $|\uparrow\rangle$, нужно очень много раз повторить весь эксперимент, собрав таким образом статистику. Если бы всякий раз (при результате ДА слева) в правом ↑-измерении действительно регистрировалось ДА, это убедительно (по критерию Эйнштейна) свидетельствовало бы, что возникающее правоориентированное состояние равно $|\uparrow\rangle$, несмотря на то что при этом нам пришлось бы полагаться на статистику. В конце концов, многие научные знания об окружающей реальности, которые кажутся нам достоверными, накапливались именно так — статистически.

Этот пример иллюстрирует и еще одно свойство квантового измерения. Левоориентированное измерение ← требовалось нам для того, чтобы «распутать» предшествовавшее запутанное состояние. До того как было выполнено левоориентированное измерение, два атома не могли расцениваться как две единицы, каждая со своим квантовым состоянием, и концепция «состояния» применялась лишь к обоим атомам сразу как к единому целому. Но мы измеряли состояние одной частицы именно для того, чтобы «распутать» другую и позволить ей приобрести собственное квантовое состояние. Вероятно, такой момент должен немного нас обнадеживать, поскольку подсказывает, почему квантовая запутанность не встречается везде и всюду (в противном случае вообще невозможно было бы говорить об отдельных объектах).

Однако существует и другая проблема, небезосновательно волнующая многих физиков. Когда измерению подвергается некий элемент A из квантово-запутанной пары, причем элемент B из этой пары находится очень далеко от A, возникает вопрос: в какой момент A и B распутываются (B приобретает собственное состояние)? Состояние другого элемента также может быть измерено отдельно, и возникает следующий вопрос: а вдруг пара распуталась в результате измерения B, а не измерения A? Если расстояние между этими элементами достаточно велико, то можно полагать, что интервал между этими элементами был *пространственноподобным* (см. раздел 1.7), что (в специальной теории относительности) означает «одновременным в некоей выбранной нами системе координат». Однако в такой ситуации будут существовать и другие системы координат, в которых измерение A считалось бы первым по времени, а также такие системы координат, где первым произошло бы измерение B (см. рис. 2.26 *б*). Иными словами, *информация* о результате измерения должна была бы перемещаться со сверхсветовой скоростью — только так она успела бы повлиять на результат второго измерения! Следовало бы считать, что два этих акта измерения применяются к некоему *нелокальному* объекту и этот объект — целостное запутанное состояние пары атомов.

Такая (часто встречающаяся) нелокальность — один из самых загадочных и интригующих аспектов запутанных состояний. У нее нет аналога в классической физике. В классической физике у нас может быть система с двумя отдельными элементами A и B, которые некогда находились вместе, причем A мог бы постфактум сообщить B, что с ним (A) происходит, либо B мог сообщить об этом A, либо после разделения они могли бы действовать совершенно независимо друг от друга. Но квантовая запутанность — другое дело. Когда квантовая запутанность сохраняется между A и B, они *не* являются независимыми; тем не менее они не могут использовать такую зависимость, существующую между ними в силу квантовой запутанности, чтобы обмениваться информацией. Именно такая невозможность обмена фактической информацией через квантовую запутанность позволяет считать, что запутанность передается мгновенно, не нарушая положений теории относительности, согласно которой передача информации со сверхсветовой скоростью невозможна. На самом деле такая передача запутанности должна считаться не мгновенной, а вневременной, поскольку неважно, в каком направлении такая передача происходит: от A к B или от B к A. Единственное ограничение касается совместного поведения A и B, когда оба этих элемента подвергаются независимому измерению. (Такая «запутанная передача» иногда именуется *квантовой информацией*. В других работах я называл ее *кванглемен-ция* [ПкР, раздел 23.10, с. 509–514]). К этой проблеме мне придется вернуться в следующем разделе.

Однако прежде я хотел бы заострить ваше внимание на другой (родственной) проблеме: можно ли считать, что волновая функция обладает подлинной онтологической реальностью? При этом мы опираемся на оригинальную идею, предложенную Якиром Аароновым и развитую Львом Вайдманом и др. Эта идея позволяет исследовать квантовые системы иначе, нежели при помощи традиционных измерительных процедур, описанных в разделе 2.8. По мысли Ааронова, это не квантовое состояние подвергается измерению (и в результате возникает новое квантовое состояние); мы просто выбираем системы с заранее заданными, почти ортогональными состояниями — начальным и конечным. В таком случае мы можем рассмотреть так называемые *слабые измерения*, не нарушающие структуру системы. Этот метод позволяет проверить такие свойства квантовой системы, которые ранее считались недоступными. В частности, можно картировать фактическую интенсивность стационарной волновой функции в пространстве. Подробное описание данной процедуры выходит за рамки этой книги, но я должен был здесь о ней упомянуть, так как она открывает путь к исследованию многих других загадочных свойств квантовой реальности [Aharonov et al., 1988; Ritchie et al., 1991].

2.13. Объективная редукция квантового состояния: предел квантовой веры?

До сих пор, пусть я и описывал ситуацию в немного нетрадиционном ракурсе, я практически не противоречил квантовой вере, которая касается фактических результатов измерений. Я подчеркнул некоторые наиболее загадочные свойства, например тот факт, что квантовые частицы нередко могут находиться в нескольких местах одновременно (в силу вездесущего принципа квантовой суперпозиции); также, согласно этому принципу, частицы могут принимать вид волн, а волны — словно состоять из неопределенного количества частиц. Более того, если система состоит из двух или более частей, то абсолютное большинство ее квантовых состояний, скорее всего, будут *запутанными*, так что мы не сможем непротиворечивым образом интерпретировать эти части как взаимно независимые элементы.

Я признаю все эти удивительные аспекты квантово-механической догматики по меньшей мере на уровне современных наблюдений, поскольку эти свойства замечательно подтвердились в многочисленных высокоточных экспериментах. Тем не менее я не могу не отметить, что существует фундаментальная несогласованность между двумя краеугольными процедурами квантовой теории, а именно унитарной (то есть шрёдингеровской) эволюцией **U** и редукцией **R**, происходящей после квантового измерения. Большинству практикующих специалистов по квантовой физике такая несогласованность кажется чем-то *мнимым*, и избавиться от нее помогает правильная интерпретация квантового формализма. В разделах 2.4 и 2.12 я уже упоминал *копенгагенскую интерпретацию*, согласно которой квантовое состояние не относится к объективной реальности, а является полностью субъективным допущением, которое просто упрощает расчеты. Однако меня по разным причинам совершенно не устраивает такая крайне субъективная точка зрения — так, еще в разделе 2.12 я говорил о том, что квантовое состояние (с точностью до коэффициента пропорциональности) должно получать подлинно объективный онтологический статус.

Другая распространенная точка зрения связана с декогеренцией под действием окружения: считается, что квантовое состояние системы не следует рассматривать в отрыве от окружающей среды. Высказывается мнение, что в нормальных условиях квантовое состояние большой квантовой системы — допустим, некоего детектора — быстро «намертво» запутается со значительной частью окружающей среды, в том числе с молекулами воздуха, причем движения большинства из этих молекул будут фактически случайны, неразличимы в каких-либо деталях и, в принципе, несущественны для работы детектора. Соответственно, квантовое

состояние этой системы (детектора) деградирует, и детектор вместе со всеми его свойствами лучше будет интерпретировать просто как классический объект.

Для точного описания такой ситуации используется математическая конструкция, именуемая *матрицей плотности*. Эту оригинальную концепцию сформулировал Джон фон Нейман, и она позволяет исключить несущественные окружающие степени свободы из описания, просто все просуммировав [*von Neumann*, 1932]. На данном этапе матрица плотности уже помогает описать «реальность» ситуации. Затем путем хитроумных математических манипуляций эту реальность можно повторно интерпретировать как вероятностную смесь различных альтернатив. *Наблюдаемая* альтернатива, возникающая в результате измерения, считается всего лишь одной из многих; вероятность ее возникновения соответствует определенной величине, которую можно верно вычислить при помощи стандартной квантово-механической **R**-процедуры, описанной в разделе 2.4.

Действительно, матрица плотности — это вероятностная смесь различных квантовых состояний, но в то же время эта смесь представлена сразу многими способами. Вышеупомянутые математические манипуляции касаются так называемого *двойного сдвига онтологии* [ПкР, в конце раздела 29.8]. Изначально считается, что вероятностная матрица воплощает вероятностную смесь различных «реальных» альтернативных состояний окружения. Затем онтология *сдвигается*, и «реальность» приписывается уже матрице как таковой. Таким образом, свобода может получить иную онтологическую интерпретацию (на основе вращения базиса гильбертова пространства), где та же самая матрица плотности уже с третьей онтологической точки зрения расценивается как описание вероятностной смеси альтернативных результатов измерения. Как правило, приводятся такие объяснения, в которых особое внимание уделяется математической стороне вопроса, и лишь поверхностное — *онтологической* согласованности. Лично я *действительно* усматриваю подлинную значимость в представлении о декогеренции матрицы плотности под действием окружающей среды — в том, как разворачивается ее математика, правда есть что-то примечательное. Тем не менее если говорить о явлениях, фактически происходящих в физическом мире, то в этой картинке чего-то решительно не хватает. Чтобы найти верное решение для измерительного парадокса, требуется изменить саму *физику*, а не только какую-то умную математику, используемую в качестве замазки для онтологических трещин! Вот как описал эту ситуацию Джон Белл [2004]:

> Когда [наиболее уверенные в себе] специалисты по квантовой физике все-таки признают, что в их обычных формулировках присутствует некоторая неоднозначность, они при этом, вероятно, будут настаивать, что обычная квантовая механика «вполне хороша для решения любых практических

задач». Я соглашусь с ними: ОБЫЧНАЯ КВАНТОВАЯ МЕХАНИКА (насколько мне известно) ВПОЛНЕ ХОРОША ДЛЯ РЕШЕНИЯ ЛЮБЫХ ПРАКТИЧЕСКИХ ЗАДАЧ.

Декогеренция под действием окружающей среды просто дает нам ДРЛПЗ-картинку (аббревиатура от белловского «для решения любых практических задач»); это может быть частный ответ, которым вполне можно на время удовлетвориться, но не окончательный. Для окончательного ответа, думаю, требуется что-то гораздо более глубокое, порывающее с нашей догматической квантовой *верой*!

Если мы постараемся сохранить непротиворечивую онтологию, продолжая беспрекословно придерживаться **U** на всех уровнях, то неизбежно придем к той или иной *многомировой* интерпретации, которую впервые ясно сформулировал Хью Эверетт III [Everett, 1957][1]. Давайте вновь рассмотрим ситуацию с котом Шрёдингера, описанную в конце раздела 2.7 (и вспомним рис. 2.15), где мы старались обеспечить согласованную **U**-онтологию на протяжении всего процесса. В той ситуации мы представили себе высокоэнергетический фотон, выпущенный из лазера L и направленный в светоделитель M. Если бы фотон прошел *через* M и активировал детектор в точке A, то открылась бы дверь A и кот проскочил бы через эту дверь в комнату, где его ждет угощение. С другой стороны, если бы фотон *отразился*, то детектор B открыл бы дверь B и кот проник бы в комнату уже через эту дверь. Однако M — не просто зеркало, а светоделитель, поэтому, согласно **U**-эволюции, на выходе из M фотон находился бы в состоянии *суперпозиции* и летел сразу по двум траекториям: MA и MB. В результате возникла бы и такая суперпозиция: ситуации «A открыта, B закрыта» и «A закрыта, B открыта» существовали бы одновременно. Можно представить, что, согласно U-эволюции, наблюдающий за всем этим человек, сидящий в комнате с кошачьим угощением, должен был бы увидеть *суперпозицию*, в которой кот выбегает сразу из двери A и из двери B. Разумеется, это была бы абсурдная ситуация, никогда и ни с кем не происходившая; более того, **U**-эволюция разворачивается совсем не так. Напротив, мы получим ситуацию, в которой человек-наблюдатель *тоже* окажется в квантовой суперпозиции двух состояний сознания, при одном из которых кот появится из двери A, а при другом — из двери B.

Так в суперпозиции друг с другом окажутся два «мира» — именно в этом и заключается эвереттовская интерпретация. Принято утверждать (логика, на мой взгляд, здесь хромает), что ощущения наблюдателя разделяются на два сосуще-

[1] См. следующее примечание Дж. А. Уилера [Wheeler, 1967; см. также De Wittand Graham, 1973; Deutsch, 1998; Wallace, 2012; Sanders et al., 2012].

ствующих отдельных восприятия, *не* находящихся в суперпозиции. Вот что меня в данном случае смущает: я не понимаю, почему два так называемых варианта восприятия должны избежать суперпозиции. Почему наш наблюдатель будет не способен ощутить квантовую суперпозицию? Да, такой опыт для нас непривычен, но *почему* нет? Можно возразить, что мы недостаточно хорошо представляем себе природу человеческого опыта, чтобы иметь право рассматривать такие проблемы с разных точек зрения. Но мы определенно можем задаться вопросом: почему человеческий опыт якобы позволяет расплетать конкретное квантовое состояние в два параллельно-мировых состояния, а не улавливать одно цельное состояние суперпозиции, которое, в сущности, и дает нам **U**-описание? Можно вспомнить состояния с полуцелым спином из раздела 2.9. Рассматривая состояние спина $|\nearrow\rangle$ как суперпозицию $|\uparrow\rangle$ и $|\downarrow\rangle$, мы не считаем, что существуют два параллельных мира, в одном из которых актуально состояние $|\uparrow\rangle$, а в другом — $|\downarrow\rangle$. У нас единственный мир с состоянием $|\nearrow\rangle$.

Более того, существует и другая проблема, связанная с вероятностями. Почему восприятие человека-наблюдателя, находящегося в суперпозиции, должно разделяться на две линии с вероятностями, соответствующими правилу Борна? Что бы это ни означало, смысла здесь практически нет! На мой взгляд, экстраполяция **U**-эволюции на такие экстремальные ситуации, как пример с котом, — это чересчур вольная игра нашего воображения, и я занимаю противоположную позицию: считаю подобные ситуации с якобы неограниченной применимостью **U** обычным *доведением до абсурда*. При этом, сколь бы хорошо ни были проверены все следствия **U**-эволюции, мы пока не наблюдали ни одного эксперимента, в котором происходило бы нечто хотя бы отдаленно напоминающее такие ситуации.

Как уже отмечалось в разделе 2.7, ключевая проблема заключается в *линейности* **U**. Такая универсальная линейность — очень необычная штука для физических теорий. В разделе 2.6 я указывал, что классические максвелловские уравнения электромагнитного поля действительно *линейны*, но при этом необходимо оговориться, что такая линейность не распространяется на классические динамические уравнения электромагнитного поля, взаимодействующего с заряженными частицами или сплошными средами. Полностью универсальная линейность, которая требуется в **U**-эволюции по представлениям современной квантовой механики, — явление в высшей степени беспрецедентное. Как уже упоминалось в разделе 1.1 (см. также раздел A.11), ньютоновское гравитационное поле также удовлетворяет линейным уравнениям, но эта линейность, опять же, не распространяется на движения тел, находящихся под действием ньютоновских сил тяготения. Вероятно, более показателен для данной проблемы тот факт, что в уточненной эйнштейновской теории гравитации, то есть в общей теории относительности, само гравитационное поле является фундаментально *не*линейным.

Вот к чему я веду: существуют действительно веские причины полагать, что линейность современной квантовой теории на самом деле может лишь *приблизительно* отражать реальное положение вещей, поэтому разделяемая многими физиками *вера* в универсальность современного аппарата квантовой механики (и его линейность в частности), а следовательно, и в *унитарность* **U**, может быть неуместной. Часто приводится довод, что не наблюдалось никаких контрпримеров, противоречащих квантовой теории, и что все поставленные по сей день эксперименты, охватывающие предельно разнообразные явления в самых разных масштабах, по-прежнему полностью подтверждают квантовую теорию, в том числе и **U**-эволюцию квантового состояния. Можно вспомнить (см. разделы 2.1 и 2.4), что тонкие квантовые эффекты (запутанность) были подтверждены на отрезке длины до 143 км [Xiao et al., 2012]. На самом деле этот эксперимент, поставленный в 2012 году, подтвердил еще более затейливые свойства квантовой механики, нежели простые ЭПР-эффекты (рассмотренные в разделе 2.10), а именно стал доказательством так называемой квантовой телепортации (см. [Zeilinger, 2010; Bennett et al., 1993; Bouwmeester et al., 1997]). Кроме того, эксперимент показал, что квантовая запутанность *действительно* сохраняется на таких расстояниях. Соответственно, пределы квантовой теории, каковы бы они ни были, по-видимому, не зависят от обычных физических расстояний — а расстояния, которые могли бы фигурировать в моем примере с котом Шрёдингера, определенно были куда меньше 143 км. Нет, я ставлю вопрос о пределах точности современной квантовой механики на совершенно иной шкале — на такой, где в совершенно особенном смысле становятся заметны *смещения масс* между компонентами, находящимися в суперпозиции.

Мои аргументы в пользу такого ограничения связаны с фундаментальным противоречием принципов квантовой механики (в первую очередь принципа линейной суперпозиции) и принципов эйнштейновской общей теории относительности. Здесь я изложу один аргумент [Penrose, 1996], который впервые сформулировал еще в 1996 году. Он основан на эйнштейновском принципе общей ковариантности (см. разделы A.5 и 1.7). В разделе 4.2 я приведу более сложный и современный аргумент, основанный на эйнштейновском принципе эквивалентности (см. раздел 1.12).

Ситуация, о которой я здесь рассказываю, охватывает квантовую суперпозицию двух состояний, каждое из которых, если рассматривать его в отрыве от второго, считалось бы *стационарным*, то есть никогда бы не изменялось. Хочу продемонстрировать следующее: если мы полагаем, что принципы общей теории относительности в данной ситуации соблюдаются, то оказывается, что имеется некоторый «предел стационарности» для суперпозиции двух состояний. Однако для того чтобы продолжить такую аргументацию, нам придется определиться

с понятием *стационарности* в квантовой механике и затронуть некоторые смежные аспекты этой дисциплины. Здесь я еще особо не углублялся в формализм квантовой теории, и прежде чем переходить к этим вопросам, нужно подробнее познакомиться с ее основополагающими идеями.

Из раздела 2.5 мы помним, что в некоторых идеализированных квантовых состояниях могут существовать очень четко определенные положения, а именно *координатные состояния*, задаваемые волновыми функциями вида $\psi(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{q})$, где $\mathbf{q}$ — это трехмерный радиус-вектор пространственной точки Q, в которой локализована волновая функция. Менее локализованными будут такие состояния, которые возникают в виде суперпозиций нескольких других локализованных состояний, возможно, описываемых разными трехмерными векторами $\mathbf{q}'$, $\mathbf{q}''$ и т. д. Такие суперпозиции даже могут быть связаны с целым *континуумом* альтернативных положений и, таким образом, заполнять всю трехмерную область пространства. Крайний случай, очень отличающийся от состояния положения, — это *импульсное* состояние, которое мы рассматривали в разделе 2.6 и позже в разделе 2.9; такие состояния описываются формулой $y(\mathbf{x}) = e^{-i\mathbf{p}\cdot\mathbf{x}/\hbar}$ для некоторого трехмерного вектора импульса $\mathbf{p}$. Эти состояния, как и координатные состояния, также являются *идеализированными* и не имеют конечных норм (см. в конце раздела 2.5). Они *равномерно* распределяются по всему пространству, однако обладают фазой, которая равномерно вращается по единичной окружности в плоскости Весселя (см. раздел A.10) в направлении $\mathbf{p}$ и со скоростью, пропорциональной импульсу частицы.

Импульсные состояния являются совершенно неопределенными относительно координат, и наоборот, любое координатное состояние остается совершенно неопределенным относительно соответствующего импульса. В данном случае координаты и импульсы являются так называемыми *канонически сопряженными* переменными, и чем лучше то или иное свойство определено относительно одной из них, тем хуже оно должно быть определено относительно другой согласно *принципу неопределенности* Гейзенберга. Он обычно записывается в виде следующей формулы:

$$\Delta\mathbf{x}\Delta\mathbf{p} \geq \frac{1}{2}\hbar,$$

где $\Delta\mathbf{x}$ и $\Delta\mathbf{p}$ — меры *неопределенности* координаты и импульса соответственно. Дело в том, что в алгебраическом формализме квантовой механики канонически сопряженные переменные становятся *некоммутативными* операторами квантового состояния (см. в конце раздела 2.8). Для операторов $\mathbf{p}$ и $\mathbf{x}$ имеем $\mathbf{xp} \neq \mathbf{px}$, где как $\mathbf{p}$, так и $\mathbf{x}$ ведут себя как *дифференциальные* операторы в отношении другого (см. раздел A.11). Однако если бы мы попытались подробнее рассмотреть эту

проблему, то серьезно вышли бы за рамки этой книги (см. Dirac [1930] или, если вас интересует современная компактная книга об основах квантово-механического формализма, Davies and Betts [1994]). Простейшее введение в тему дается в главах 21 и 22 книги ПкР.

В известном смысле (в соответствии с требованиями специальной теории относительности) время t и энергия E также являются канонически сопряженными, и принцип неопределенности Гейзенберга между временем и энергией формулируется следующим образом:

$$\Delta t \Delta E \geq \frac{1}{2}\hbar.$$

Строгая интерпретация этого соотношения порой вызывает споры, однако существует один общепризнанный вариант его использования — в случае с радиоактивными ядрами. Для такого нестабильного ядра Δt можно считать его *временем жизни*, а вышеприведенное отношение указывает, что должна существовать неопределенность энергии ΔE или, что аналогично, неопределенность массы не менее $c^{-2}\Delta E$ (в соответствии с эйнштейновским уравнением $E = mc^2$).

Теперь давайте вернемся к нашей суперпозиции двух стационарных состояний. В квантовой механике стационарным именуется такое состояние, *энергия которого* точно определена, поэтому в соответствии с принципом неопределенности Гейзенберга между временем и энергией данное состояние должно быть совершенно равномерно распределено во времени, что на самом деле свидетельствует о его стационарности (рис. 2.27). Более того, как и в случае с импульсным состоянием, его фаза равномерно вращается по единичной окружности в плоскости Весселя со скоростью, пропорциональной значению энергии состояния E. Фактически в данном случае зависимость от времени описывается формулой $e^{Et/i\hbar} = -\cos(Et/\hbar) - i\sin(Et/\hbar)$, так что частота данного фазового вращения равна $E/2\pi\hbar$.

Здесь я рассмотрю простейшую ситуацию, связанную с квантовой суперпозицией, а именно суперпозицию двух состояний, каждое из которых было бы стационарным, если бы рассматривалось отдельно. Для простоты картины представьте себе камень, находящийся в суперпозиции в двух положениях, которым соответствуют состояния $|1\rangle$ и $|2\rangle$. Камень лежит на горизонтальной плоской поверхности, а два положения отличаются друг от друга лишь тем, что камень перемещается из положения в состоянии $|1\rangle$ в положение в состоянии $|2\rangle$ путем горизонтального смещения, а энергия E этих состояний одинакова (рис. 2.28). Мы рассматриваем общую суперпозицию

$$|\psi\rangle = \alpha|1\rangle + \beta|2\rangle,$$

где α и β — ненулевые комплексные константы. Отсюда следует, что $|\psi\rangle$ также стационарно[1] и имеет определенное значение энергии E. Когда энергии $|1\rangle$ и $|2\rangle$ различаются, возникает интересная новая ситуация, рассмотренная в разделе 4.2.

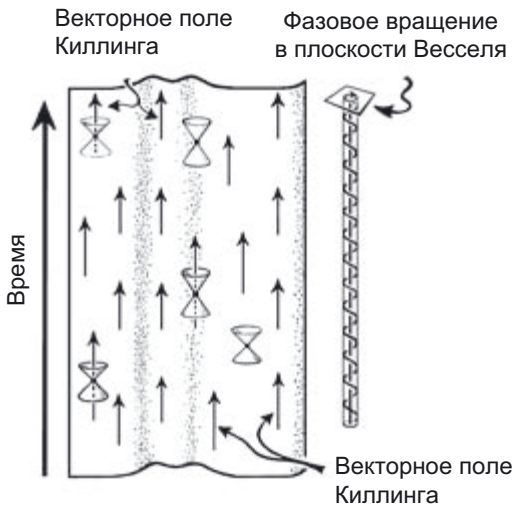


Рис. 2.27. Классическое и квантовое представление о стационарности. В классических терминах пространства-времени стационарное пространство-время обладает времениподобным векторным полем Киллинга **k**, вдоль которого геометрия пространства-времени не изменяется при каких бы то ни было (локальных) перемещениях, порождаемых **k**, а время считается ориентированным в направлении **k**. С квантово-механической точки зрения данное состояние обладает строго определенной энергией E, поэтому изменяется со временем лишь на общую фазу $e^{Et/i\hbar}$, соответствующую вращению по единичной окружности в плоскости Весселя с частотой $E/2\pi\hbar$

В общей теории относительности стационарность выражается немного иным (хотя и схожим) образом. Мы по-прежнему считаем стационарное состояние подлинно равномерно распределенным во времени (хотя и без какой-либо комплексной фазы для вращения), но время не определено однозначно. Общая концепция темпоральной равномерности для пространства-времени $\mathcal{M}$ обычно выражается в терминах *временоподобного вектора Киллинга* **k**. Вектор Киллинга — это времениподобное поле в пространстве-времени (см. раздел A.6, рис. A.17), вдоль которого структура пространственно-временной метрики

[1] Те из вас, кто знаком со стандартным квантовым формализмом, могут представить ситуацию так: если взять $\mathbf{E} = (i\hbar)^{-1}\partial/\partial t$ в качестве оператора энергии, то получим $\mathbf{E}|1\rangle = E|1\rangle$ и $\mathbf{E}|2\rangle = E|2\rangle$, где $\mathbf{E}|\psi\rangle = E|\psi\rangle$.

остается совершенно неизменной, а **k**, будучи времениподобным, позволяет отождествить вектор **k** с направлением хода *времени* в соответствующей системе координат (см. раздел А.7, рис. А.29). Как правило, на **k** налагается дополнительное ограничение: этот вектор должен быть *невращающимся*, то есть *ортогональным гиперповерхности*, но в данном случае это не играет особой роли.

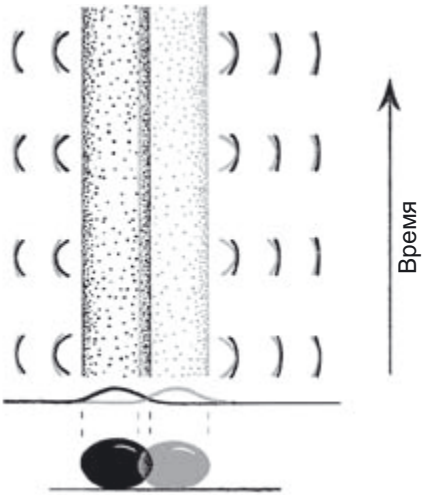


Рис. 2.28. Гравитационное поле камня, находящегося в суперпозиции двух пространственных положений и претерпевающего горизонтальное смещение (положения камня обозначены черным и серым тоном). В таком случае получаем суперпозицию двух пространств-времен с немного различающимися значениями ускорения свободного падения (показаны в виде черной и серой пространственно-временных кривых). Усредненный по пространству квадрат разности этих ускорений дает меру ошибки E_G, связанной с идентификацией пространства-времени

Выше мы уже сталкивались с вектором Киллинга (см. разделы 1.6 и 1.9), когда говорили об оригинальной пятимерной теории Калуцы — Клейна. В этой теории требовалось наличие непрерывной симметрии вдоль дополнительного пространственного измерения, а поле вектора Киллинга должно было указывать в направлении этой симметрии, так чтобы все пятимерное пространство могло бы скользить само по себе в этом направлении без изменения геометрии его метрики. Идея вектора Киллинга **k** для стационарного четырехмерного пространства-времени $\mathcal{M}$ совершенно аналогична, только теперь четырехмерное пространство-время может скользить само по себе в направлении **k**, соответствующем времени, и сохранять при этом структуру своей пространственно-временной метрики (см. рис. А.29 в разделе А.7).

Эта картина очень напоминает квантово-механическое определение стационарности (без фазового вращения), но теперь мы должны рассматривать ее в контексте искривленного пространства-времени общей теорией относительности. В общей теории относительности поле вектора Киллинга *не* просто дано нам в качестве движения вдоль предполагаемой оси времени. С другой стороны, в стандартном формализме квантовой механики эволюция во времени считается *данностью* (и протекает относительно некоей заранее обозначенной временной координаты). Это специфический ингредиент уравнения Шрёдингера. Именно в этой разнице коренится фундаментальная проблема, возникающая при попытках рассматривать квантовую суперпозицию в контексте общей теории относительности.

Следует пояснить: чтобы можно было рассматривать проблему в общей теории относительности, будем считать, что каждое *индивидуальное* состояние, рассматриваемое нами (в данном случае это состояния $|1\rangle$ и $|2\rangle$) уместно будет трактовать как *классический* объект, подчиняющийся классическим законам общей теории относительности (на соответствующем уровне приближения). Действительно, если бы это было *не* так, то это основательно подорвало бы нашу веру в универсальную применимость законов квантовой механики, поскольку *наблюдения* подтверждают, что макроскопические тела с весьма точным приближением подчиняются классическим законам. Классические законы *действительно* исключительно точно описывают макроскопические объекты, так что если этот факт нельзя увязать с квантовыми процедурами, то, пожалуй, именно с последними что-то неладно. То же самое касается классических процедур общей теории относительности, и в разделе 1.1 мы уже отмечали, с какой исключительной точностью теория Эйнштейна описывает крупные «чисто гравитационные» системы (то есть динамику двойных нейтронных звезд). Следовательно, если мы собираемся признать непогрешимость **U**-процедур квантовой механики, то также должны согласиться, что эти процедуры (например, рассматриваемые здесь) вполне можно применять и в контексте общей теории относительности.

Теперь стационарность индивидуальных состояний $|1\rangle$ и $|2\rangle$ должна описываться векторами Киллинга $\mathbf{k}_1$ и $\mathbf{k}_2$ в *различных* пространственно-временных многообразиях $\mathcal{M}_1$ и $\mathcal{M}_2$, причем такие векторы будут описывать гравитационные поля каждого из многообразий. Действительно, два пространства-времени должны считаться отличными друг от друга, точно так же как камни расположены в разных точках относительно окружающей земной геометрии. Соответственно, не существует никакого способа однозначно отождествить $\mathbf{k}_1$ с $\mathbf{k}_2$ (то есть считать их одинаковыми) и тем самым постулировать стационарность суперпозиции. Это один из аспектов эйнштейновского принципа *общей ковариантности* (см. раз-

делы A.5 и 1.7), запрещающего любое осмысленное *поточечное отождествление* двух по-разному искривленных пространственно-временных геометрий (то есть нельзя утверждать, что точка в одном пространстве-времени *идентична* некоей точке в другом пространстве-времени всего лишь потому, что их пространственно-временные координаты совпадают). Мы не будем пытаться разрешить эту проблему на каком-либо более глубоком уровне, а просто оценим *погрешность*, возникающую при попытке отождествить $\mathbf{k}_1$ с $\mathbf{k}_2$ в ньютоновском пределе (где $c \to \infty$). Технически схема такого ньютоновского предела была описана Эли Картаном [1945] и Куртом Фридрихсом [1927]; см. также [Ehlers, 1991].

Как измерить такую погрешность? В каждой «отождествляемой» точке у нас будут два разных ускорения свободного падения f_1 и f_2 (они берутся относительно вектора Киллинга $\mathbf{k}_1 = \mathbf{k}_2$, который теперь является общим для двух систем пространства-времени; см. рис. 2.28). Это будут соответствующие ньютоновские гравитационные поля, каждое из которых действует в своем пространстве-времени, а квадрат этой разности $|f_1 - f_2|^2$ будет считаться локальной мерой расхождения (или *погрешностью*), возникающей при отождествлении пространств-времен. Данная локальная мера погрешности интегрируется (то есть суммируется) по всему трехмерному пространству. Результирующая *полная* погрешность, получаемая таким образом, — это величина E_G, которую в данной ситуации можно получить при помощи сравнительно простых вычислений. Это будет энергия, которую пришлось бы затратить на разделение двух камней, которые сначала лежали вместе, а затем были разнесены по точкам $|1\rangle$ и $|2\rangle$, где из всех сил, действующих между ними, учитывается лишь *гравитация*. В более общем смысле величина E_G считалась бы *собственной гравитационной энергией разницы* между распределениями масс в $|1\rangle$ и $|2\rangle$ (подробнее см. [Penrose, 1996], а также в разделе 4.2). Лайош Диоши (Lajos Diósi) [1984, 1987] пришел к подобной идее несколькими годами ранее, но при этом он не отталкивался от общей теории относительности. В разделе 4.2, где мы подробнее рассмотрим эти проблемы, более убедительно показано, что E_G является препятствием для общей стационарности суперпозиции в силу эйнштейновского принципа эквивалентности.

Величина погрешности E_G считается фундаментальной неопределенностью *энергии* в суперпозиции. Поэтому, воспользовавшись принципом неопределенности Гейзенберга применительно ко времени и энергии (как в вышеприведенном примере с нестабильной частицей), мы делаем вывод о том, что суперпозиция $|\psi\rangle$ неустойчива и должна распасться до состояния $|1\rangle$ или $|2\rangle$ за среднее время τ порядка:

$$\tau \approx \frac{\hbar}{E_G}.$$

Итак, оказывается, что квантовые суперпозиции не вечны. Если смещение масс между парой состояний, находящихся в суперпозиции, исчезающе мало, а именно так и обстоит дело со всеми поставленными на данный момент квантовыми экспериментами, то именно по этим причинам суперпозиция получится очень долговечной, и никакого конфликта с законами квантовой механики мы не заметим. Но при значительном смещении масс между состояниями такая суперпозиция должна спонтанно распадаться до одного или другого варианта, и такое противоречие основным квантовым законам должно быть наблюдаемо. До сих пор ни один квантовый эксперимент еще не достиг уровня, на котором можно было бы заметить такое расхождение, но подобные эксперименты готовятся уже несколько лет, и один из них я вкратце опишу в разделе 4.2. Надеюсь, что его результаты будут получены уже в пределах следующего десятилетия — несомненно, это будет захватывающее достижение.

Даже если результаты таких экспериментов укажут на недостатки квантовой веры (возможно, эмпирически подтвердят вышеуказанный критерий $t \approx \hbar/E_G$), мы все еще будем очень далеки от создания расширенной квантовой теории, в которой **R** и **U** были бы отличными приближениями: **U** — при малом смещении масс между состояниями в суперпозиции, **R** — при большом. Однако при этом может быть получена некая оценка *границ* (пока не установленных) применимости современной квантовой веры. Я бы предположил, что все случаи редукции квантового состояния — это гравитационные эффекты вышеупомянутого типа. Во многих стандартных ситуациях, связанных с квантовыми измерениями, смещение масс будет происходить главным образом в *окружающей среде*, запутанной с измерительным прибором, поэтому традиционная точка зрения, связанная с *декогеренцией под действием окружения*, может обрести непротиворечивую онтологию. Ключевые свойства моделей редукции, одна из которых рассматривается здесь, рассмотрели Гирарди, Римини и Вебер в своей революционной разработке 1986 года [Ghirardi et al., 1986, 1990]. Но эти идеи гораздо шире, и они вполне могут быть проверены в рамках различных экспериментов, активно разрабатываемых сегодня [Marshall et al., 2003; Weaver et al., 2016; Eerkens et al., 2015; Pepper et al., 2012; см. также Kaltenbaek et al., 2015; Li et al., 2011; Bedingham and Halliwell, 2014]. Вполне вероятно, эти результаты будут получены в ближайшее десятилетие или в результате разработки новых, пока еще не оформившихся идей.

ГЛАВА 3

Фантазия

3.1. Большой взрыв и вселенные Фридмана

Может ли *фантазия* играть какую-либо неиллюзорную роль в наших попытках понять физическую реальность? Определенно, фантазия — полная противоположность науки как таковой, и ей не место в серьезном научном дискурсе. Однако остается ощущение, что от этого вопроса не так легко отмахнуться, как могло бы показаться, — многие природные процессы покажутся фантастичными, если исходить из выводов, к которым может привести нас рациональный научный опыт, основанный на надежных экспериментальных исследованиях. Как мы убедились, особенно в предыдущей главе, мир действительно устроен самым фантастическим образом, если изучать его на микроуровне, где царят квантовые явления. Конкретный материальный объект может находиться в нескольких местах и, подобно сказочному вампиру (способному превращаться из летучей мыши в человека и обратно, когда ему вздумается), может проявлять то корпускулярные, то волновые свойства словно по собственному выбору. При этом его «поведение» подчиняется таинственным числам, в которых содержится мнимый квадратный корень из –1.

Более того, в предельно крупных масштабах опять обнаруживаются явления, многие из которых могут показаться фантастическими — возможно, даже поразительнее всех находок литературной фантастики. Например, иногда наблюдается столкновения между целыми галактиками, и приходится считать, что они неотвратимо поглощают друг друга (а мы фиксируем это по возникающим искажениям пространства-времени, провоцируемым обеими галактиками). Действительно, такие искажения пространства-времени иногда можно наблюдать даже напрямую — по грубому искривлению снимков очень отдаленных галактик. Вдобавок самые экстремальные из известных нам искажений пространства-времени могут приводить к возникновению массивных черных дыр в космическом пространстве: недавно удалось наблюдать, как две такие дыры поглощают друг

друга и образуют еще более крупную [Abbott et al., 2016]. Есть черные дыры, которые в миллионы или десятки тысяч миллионов раз тяжелее Солнца, поэтому такие дыры могли бы с легкостью заглатывать целые солнечные системы. Тем не менее эти монстры весьма малы по сравнению с самими галактиками, в центрах которых они и обретаются. Часто такая черная дыра выдает свое существование, порождая два коллимированных пучка высокоэнергетических частиц. Эти пучки исторгаются из черной дыры в противоположных направлениях из крошечного центрального региона той галактики, в которой сидит эта дыра; частицы летят со скоростью, которая может достигать до 99,5 % скорости света [Tombesi et al., 2012; Piner, 2006]. Однажды удалось наблюдать, как такой пучок вылетел из одной галактики и прицельно попал в другую, как будто эта была колоссальная межгалактическая война.

В еще более крупных масштабах обнаруживаются целые регионы, заполненные невидимым нечто, пронизывающим космос. Складывается впечатление, что на эту совершенно неизвестную субстанцию приходится около 84,5 % всей материи, имеющейся во Вселенной. При этом существует еще *что-то*, достигающее самых дальних пределов наблюдаемой Вселенной и словно растаскивающее ее в разные стороны с нарастающей скоростью. Словно от безысходности ученые дали двум этим сущностям довольно туманные названия — «темная материя» и «темная энергия» соответственно. Именно темная материя и темная энергия в основном определяют общую структуру известной Вселенной. Еще тревожнее кажется следующий факт: современная космология почти наверняка доказывает, что вся известная нам Вселенная возникла из одного гигантского взрыва, до которого совсем ничего не было — если вообще можно говорить о чем-то «до» возникновения пространственно-временного континуума, который, как мы полагаем, лежит в основе всей материальной реальности. Поистине, такая концепция *Большого взрыва* — фантастическая идея!

Так и есть; но в нашем распоряжении все больше эмпирических доказательств в пользу того, что на заре существования наша Вселенная действительно была невероятно плотной и стремительно расширялась. В ней было заключено не только все материальное содержимое того космоса, который нам известен, но и все пространство-время, на фоне которого сейчас разыгрывается бытие физической реальности и которое, по-видимому, неограниченно далеко простирается во все стороны. Все, что нам известно, по-видимому, возникло в результате этого Большого взрыва. Каковы доказательства? Мы должны оценить достоверность этой идеи и попытаться понять, куда она может нас привести.

В этой главе мы обсудим некоторые современные идеи, касающиеся происхождения самой Вселенной, и в частности коснемся следующей проблемы: в какой

степени оправданно прибегать к фантазиям для объяснения эмпирических фактов. В последние годы многочисленные эксперименты действительно дали нам огромные объемы данных, непосредственно значимых для понимания истоков Вселенной. Вещи, которые ранее казались набором в основном непроверенных спекуляций, перешли в категорию точной науки. Важнее всего упомянуть спутники COBE[1], запущенный в 1989 году, WMAP[2], запущенный в 2001 году, а также космическую обсерваторию им. Планка, которая работает с 2009 года. Упомянутые спутники постепенно исследовали реликтовый космический микроволновый фон (см. раздел 3.4) все подробнее и подробнее. Однако нерешенные вопросы остаются, и в поисках ответов на них некоторые специалисты по теоретической космологии углубились в дебри, которые вполне уместно назвать совершенно фантастическими.

Да, в некоторой степени фантазия, безусловно, оправданна, но не слишком ли рьяно устремились современные теоретики в этом направлении? В разделе 4.3 я озвучу мою собственную довольно нетрадиционную версию, позволяющую разрешить многие из этих загадок. Идеи, на которых замешан мой ответ, кому-то также могут показаться дикими, и я вкратце опишу, почему их следовало бы воспринимать всерьез. Тем не менее в этой книге меня больше интересуют устоявшиеся к настоящему времени представления о самых ранних этапах эволюции нашей замечательной Вселенной, и я хотел бы обсудить, насколько правдоподобны отдельные направления, в которых ведут свои исследования некоторые современные космологи.

Для начала у нас есть величественная эйнштейновская общая теория относительности, которая, насколько известно, исключительно точно описывает структуру нашего искривленного пространства-времени и движение небесных тел (см. разделы 1.1 и 1.7). В 1922 и 1924 годах вслед за первыми попытками Эйнштейна применить эту теорию для описания целостной структуры Вселенной русский математик Александр Фридман впервые нашел решения для эйнштейновских уравнений поля в контексте пространственно однородного (гомогенного и изотропного) распределения расширяющейся материи, причем приблизительной моделью такой материи считалась *идеальная жидкость* (*пылевое решение*), представляющая усредненное распределение массы-энергии галактик [Rindler, 2001; Wald, 1984; Hartle, 2003; Weinberg, 1972]. Действительно, с эмпирической точки зрения кажется, что в таком случае получается достаточно хорошее

[1] COBE — COsmic Background Explorer («Исследователь космического реликтового излучения»). — *Примеч. пер.*

[2] WMAP — Wilkinson Microwave Anisotropy Probe («Зонд им. Уилкинсона для изучения микроволновой анизотропии»). — *Примеч. пер.*

общее приближение к усредненному распределению материи в существующей Вселенной, и выводится тензор энергии **T**, который был необходим Фридману для представления гравитации в уравнении Эйнштейна $\mathbf{G} = 8\pi\gamma\mathbf{T} + \Lambda\mathbf{g}$ (см. раздел 1.1). Характерная черта моделей Фридмана заключается в том, что расширение начинается с *сингулярности* (теперь этот момент именуется *Большим взрывом*). Тогда кривизна пространства-времени была *бесконечной*, а плотность массы-энергии источника материи **T** устремлялась бы в бесконечность, если бы мы попытались отмотать время назад до этой пространственно-временной сингулярности. (Удивительно, что общеупотребительный ныне термин «Большой взрыв» задумывался как уничижительный; придумал его Фред Хойл, ярый сторонник альтернативной теории *стационарной Вселенной*; см. раздел 3.2.) Он впервые упомянул слова «Большой взрыв» в радиоинтервью компании BBC — серию таких выступлений он сделал в 1950 году. В разделе 3.10 эти интервью упоминаются и в ином контексте; позже на их основе была составлена книга [Hoyle, 1950].

Пока я стану условно считать, что очень малая эйнштейновская космологическая постоянная Λ — именно она и обусловливает *ускоряющееся* расширение Вселенной, упомянутое ранее (см. также раздел 1.1), — равна *нулю*. Тогда нам потребуется рассмотреть всего три отдельные ситуации, определяемые пространственной геометрией: кривизна пространства K может быть положительной ($K > 0$), нулевой ($K = 0$) или отрицательной ($K < 0$). В авторитетных книгах по космологии принято *нормировать* величину K, приводя ее к одному из трех значений: 1, 0, –1. Здесь рассказ будет понятнее, если считать K вещественным числом, характеризующим фактическую кривизну пространства. Мы можем представлять K как величину, указывающую такую пространственную кривизну в какой-то специально подобранный момент времени t. Например, можно условиться, что t будет соответствовать эпохе *последнего рассеяния* (см. раздел 3.4), когда образовался космический микроволновый фон, но выбор конкретного момента в данном случае не важен. Суть в том, что *знак* K не будет меняться со временем, поэтому положительное, отрицательное или нулевое значение K характеризует модель в целом, независимо от выбранного «момента отсчета».

Однако следует отметить, что значение K само по себе не вполне характеризует геометрию пространства-времени. Также существуют нестандартные «свернутые» варианты таких моделей, пространственная геометрия которых бывает довольно сложной, причем в некоторых примерах Вселенная может быть *конечной*, даже если $K = 0$ или $K < 0$. Некоторые ученые интересовались такими моделями (см. Levin [2012], Luminet et al., [2003], исходно Schwarzschild [1900]). Однако здесь эти модели для нас не важны; эта проблема существенно не влияет на большинство аргументов, которые я в данном случае излагаю.

Если не учитывать топологические сложности, то у нас получится всего три типа однородной геометрии, которые (на плоскости) очень красиво изображал голландский художник М. К. Эшер (рис. 3.1; ср. также с рис. 1.38 в разделе 1.15). Трехмерная картина выглядит так же.

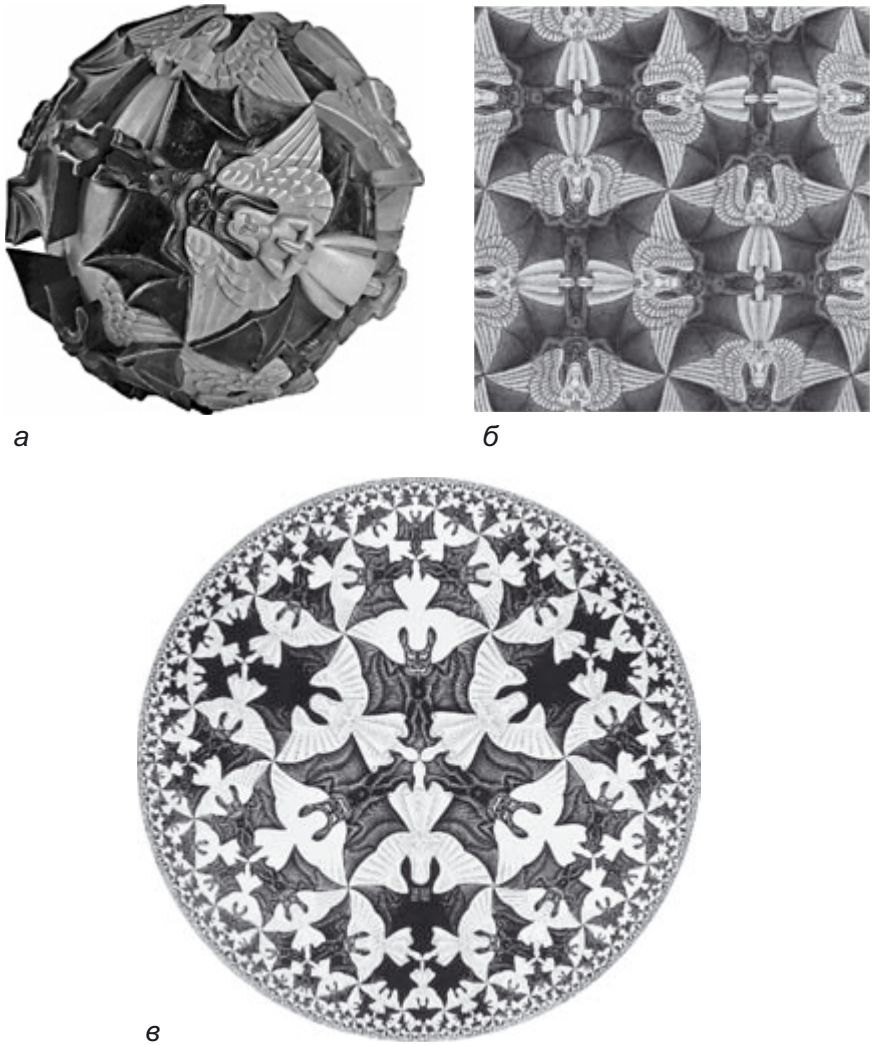


Рис. 3.1. Эшер иллюстрирует (для двумерного случая) три типа однородной геометрии: *а*) положительная кривизна ($K > 0$); *б*) плоская евклидова геометрия ($K = 0$); *в*) отрицательная кривизна ($K < 0$) с использованием предложенного Бельтрами конформного представления гиперболической геометрии, также показанного на рис. 1.38

Проще всего понять случай $K = 0$, так как в этом случае пространственное сечение будет представлять собой обычное трехмерное евклидово пространство, хотя, чтобы описать расширяющуюся Вселенную, нам понадобится множество таких последовательных сечений: см. рис. 3.2 *б*. (Это расширение можно понимать в терминах расходящихся времениподобных линий, которые соответствуют мировым линиям идеализированных галактик, описываемых этой моделью. Это будут линии времени, о которых мы поговорим далее.) Трехмерные пространства, являющиеся пространственными сечениями в случае $K > 0$, представить немного сложнее, поскольку они являются *3-сферами* (S^3), каждая из которых в трех измерениях аналогична двумерной поверхности обычной сферы (S^2), а расширение Вселенной выражается как увеличение радиуса сферы со временем (рис. 3.2 *а*). В случае отрицательной кривизны ($K < 0$) трехмерные пространства обладают *гиперболической* геометрией (она же *геометрия Лобачевского*). Такую геометрию можно точно представить, воспользовавшись *конформным представлением* (Бельтрами — Пуанкаре), которое в двумерном случае описывается как область, ограниченная окружностью $\S$ в евклидовой плоскости, где прямые линии представлены в виде дуг окружности, пересекающих ограничивающую окружность под прямыми углами (рис. 3.2 *в* и рис. 1.38 в разделе 1.15) (см., в частности, ПкР, разделы 2.4–2.6; Needham [1997]). Трехмерная гиперболическая геометрия выглядит схоже, однако в ней вместо окружности $\S$ наличествует сфера (обычная 2-сфера), ограничивающая область (3-шар) в евклидовом трехмерном пространстве.

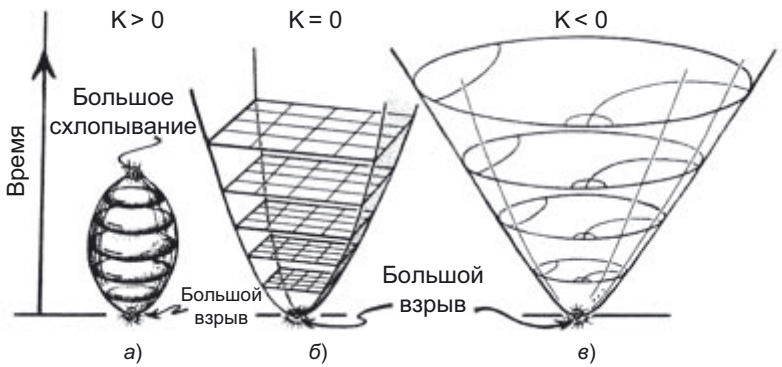


Рис. 3.2. Фридмановские пылевые космологические модели для случая нулевой космологической постоянной Λ: *а*) $K > 0$, *б*) $K = 0$, *в*) $K < 0$

Термин «конформная», применяемый в этих моделях, употребляется потому, что в гиперболической геометрии величина *угла* между двумя гладкими кривыми в точке их пересечения будет такой же, как и в фоновой евклидовой гео-

метрии (так, например, углы на кончиках рыбьих плавников на рис. 1.38 *а* или крылья чертей на рис. 3.1 *в* представлены без искажений, независимо от того, как близко к ограничивающей окружности они расположены). Другая (грубая) формулировка того же принципа звучит следующим образом: *формы* (но не размеры) очень мелких деталей в таких представлениях всегда отображаются без искажений (см. также рис. А.39 в разделе А.10).

Как отмечалось ранее, уже найдены некоторые убедительные доказательства того, что в нашей Вселенной космологическая постоянная Λ имеет небольшое *положительное* значение, поэтому мы должны рассматривать фридмановские модели, соответствующие $\Lambda > 0$. На самом деле, как ни ничтожна Λ, ее значение все-таки достаточно велико (при этом, согласно уравнениям Эйнштейна, мы продолжаем считать ее константой), чтобы преодолеть коллапс и «большое схлопывание», показанное на рис. 3.2 *а*. Вместо этого при *всех трех* возможных вариантах значения K, допускаемых современными наблюдениями, Вселенная в итоге должна пуститься в расширение *с ускорением*. При такой положительной постоянной Λ расширение Вселенной будет продолжаться до бесконечности и в конечном итоге станет *экспоненциальным* (см. рис. А.1 в разделе А.1). Согласно таким выкладкам мы представляем себе общую историю Вселенной такой, как показано на рис. 3.3. Задний план изображен неопределенно, чтобы показать, что наблюдения допускают все три варианта пространственной кривизны K.

Варианты *отдаленного будущего* во всех этих моделях при $\Lambda > 0$ даже при наличии в них некоторых нерегулярностей очень похожи и хорошо описываются конкретной пространственно-временной моделью, которая называется *де-ситтеровским* пространством. Тензор Эйнштейна $\mathbf{T}$ в нем равен просто $\Lambda\mathbf{g}$. Эта модель была найдена Виллемом де Ситтером (и независимо от него Туллио Леви-Чивитой) в 1917 году (см. [de Sitter, 1917a, b; Levi-Cività, 1917; Schrödinger, 1956]; ПкР, п. 28.4). В настоящее время принято считать, что эта модель хорошо аппроксимирует отдаленное будущее *именно нашей* Вселенной, когда тензор энергии полностью определяется Λ, поэтому в предельно далеком будущем сложится ситуация $\mathbf{G} \approx \Lambda\mathbf{g}$.

Разумеется, при этом мы предполагаем, что эйнштейновские уравнения ($\mathbf{G} = 8\pi\gamma\mathbf{T} + \Lambda\mathbf{g}$) будут действовать неограниченно долго и значение Λ, определенное в наше время, останется константой. В разделе 3.9 будет показано, что, согласно экзотическим идеям инфляционной космологии, модель де Ситтера должна была описывать Вселенную и на значительно более раннем этапе, непосредственно после Большого взрыва, однако значение Λ на тот момент должно было колоссально превышать нынешнее. Эти вопросы обретут для нас важность позднее (см. разделы 3.7–3.9 и 4.3), а пока мы не будем подробно на них останавливаться.

Рис. 3.3. Пространственно-временное представление ожидаемой эволюции нашей Вселенной с изменениями, учитывающими наблюдаемое (достаточно большое) значение $\Lambda > 0$. Неопределенность, показанная на заднем плане картинки, отражает неопределенность пространственной геометрии вообще, что не играет особой роли в эволюционном плане

Де-ситтеровское пространство — это высокосимметричное пространство-время, которое можно описать как (псевдо-)сферу в пятимерном пространстве Минковского (рис. 3.4 *a*). Эта (псевдо-)сфера возникает в точке $t^2 - w^2 - x^2 - y^2 - z^2 = -3/\Lambda$, получая локальную метрическую структуру от охватывающего пятимерного пространства Минковского с координатами (t, w, x, y, z) (Те, кто знает, как стандартным образом записываются метрики при помощи дифференциалов, понимают, что эта пятимерная метрика Минковского приобретает вид $ds^2 = dt^2 - dw^2 - dx^2 - dy^2 - dz^2$.) Де-ситтеровское пространство полностью повторяет симметрию четырехмерного пространства Минковского; в обоих случаях имеем 10-параметрическую группу симметрий. Также можно вспомнить гипотетическое анти-де-ситтеровское пространство, рассмотренное в разделе 1.15. Оно весьма тесно связано с де-ситтеровским пространством и обладает группой симметрии такого же порядка.

Де-ситтеровское пространство — пустая модель, в которой тензор энергии **T** равен нулю, поэтому там нет никаких (идеализированных) галактик, которые могли бы определять линии времени, чьи ортогональные трехмерные пространственные сечения позволяли бы определять конкретные трехмерные геометрии «синхронного времени». На самом деле, довольно примечателен факт: оказывается, в де-ситтеровском пространстве можно выбирать такие трехмерные пространственные сечения (с синхронным временем) тремя принципиально разными способами, так что де-ситтеровское пространство можно интерпретировать как равномерно расширяющуюся в пространстве Вселенную с каждым из трех альтернативных типов пространственной кривизны в зависимости от того, каким образом ее рассекают такие трехмерные сечения, соответствующие одному

и тому же космическому времени: $K > 0$ (при $t = \text{const}$), $K = 0$ (при $t - w = \text{const}$) и $K < 0$ (при $-w = \text{const}$) (рис. 3.4 *б*–*г*). Это смог красиво продемонстрировать Эрвин Шрёдингер в своей книге «Расширяющиеся вселенные» (1956). Более ранняя модель стационарной Вселенной, которую мы обсудим в разделе 3.2, описывается пространством де Ситтера в соответствии с сечением $K = 0$, показанным на рис. 3.4 *в* (и конформно представленным на рис. 3.26 *б* в разделе 3.5). В большинстве версий инфляционной космологии (до которой мы дойдем в разделе 3.9) также используется такое сечение $K = 0$, поэтому инфляция может однородно и экспоненциально продолжаться в течение неограниченного времени.

На самом деле, что касается крупномасштабной структуры нашей реальной Вселенной, современные наблюдения не позволяют однозначно ответить, какой из этих вариантов пространственной геометрии точнее всего ее описывает. Тем не

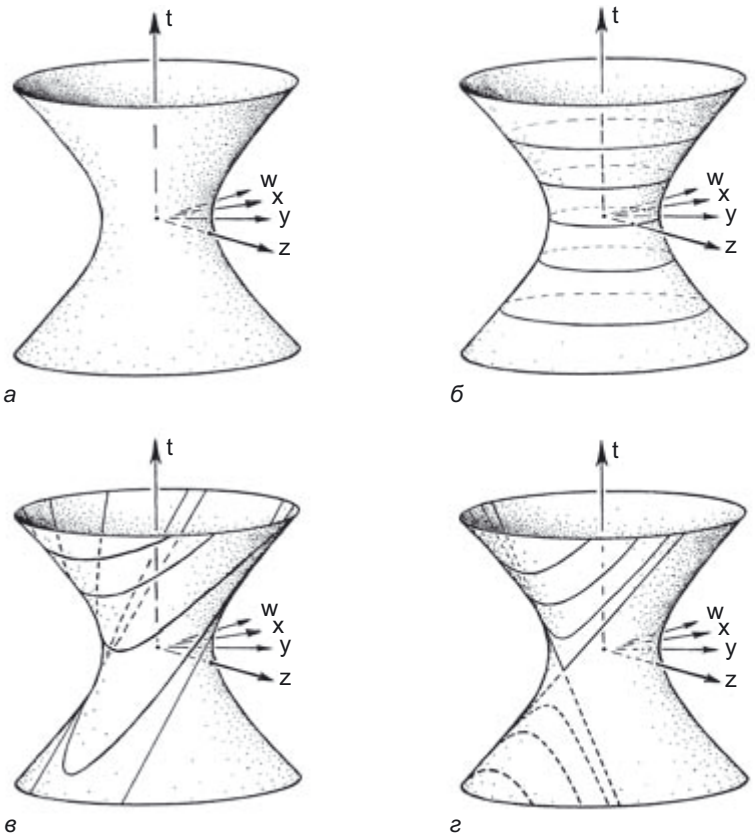


Рис. 3.4. Де-ситтеровское пространство (*а*); $K > 0$ (при $t = \text{const}$) (*б*); $K = 0$ (при $t - w = \text{const}$) – модель стационарной Вселенной (*в*); $K < 0$ (при $-w = \text{const}$) (*г*)

менее каков бы ни был окончательный ответ, сейчас не кажется, что вариант $K = 0$ так уж близок к истине (примечательно, особенно с учетом на первый взгляд убедительных доказательств в пользу $K < 0$, появившихся к концу XX века). В некотором смысле данная ситуация крайне неудовлетворительна с эмпирической точки зрения; ведь если мы всего лишь можем сказать, что значение K очень близко к нулю, то по-прежнему остается вероятность того, что более тщательные наблюдения (или более убедительная теория) впоследствии покажут, что наша Вселенная точнее соответствует какой-то другой пространственной геометрии (то есть сферической или гиперболической). Так, если в конце концов появятся хорошие доказательства в пользу $K > 0$, это будет подлинно важно с философской точки зрения, поскольку будет означать, что пространственные размеры Вселенной конечны. Однако по состоянию на настоящий момент принято просто утверждать следующее: согласно наблюдениям $K = 0$. Это может быть очень хорошим приближением, но в любом случае мы не знаем, насколько близка реальная Вселенная к подлинной пространственной однородности и изотропии, особенно с учетом определенных противоречивых данных, полученных при наблюдении космического микроволнового фона (например, [Starkman et al., 2012; Gurzadyan and Penrose, 2013, 2016]).

Для того чтобы построить картину полного пространства-времени в соответствии с фридмановскими моделями и их обобщениями, нужно знать, как будут со временем изменяться «размеры» нашей пространственной геометрии, причем с самого начала. В стандартных космологических моделях, например у Фридмана, либо в обобщенных моделях, вкратце именуемых ФЛРУ (*Фридмана — Леметра — Робертсона — Уокера*), — во всех моделях этого общего класса пространственные сечения однородны и изотропны и полное пространство-время обладает такой же симметрией, как и сами сечения. В них есть четкое определение *космического времени* t, которое описывает эволюцию такой вселенской модели. Данное космическое время стартует в момент $t = 0$ (Большой взрыв) и отсчитывается идеализированными часами, следующими по мировым линиям идеализированных галактик (рис. 3.5, а также рис. 1.17 в разделе 1.7). Я буду именовать эти мировые линии *линиями времени* в модели ФЛРУ (в работах по космологии они также иногда называются мировыми линиями *фундаментальных наблюдателей*). Линии времени — это геодезические кривые, ортогональные пространственным сечениям, которые, в свою очередь, являются 3-плоскостями с одинаковым значением t.

Случай де-ситтеровского пространства имеет важную особенность: поскольку, как упоминалось ранее, пространство пустое, то есть тензор энергии-импульса $\mathbf{T}$ в уравнении $\mathbf{G} = 8\pi\mathbf{T} + \Lambda\mathbf{g}$ равен нулю, то у нас нет никаких мировых линий, связанных с материальными телами, которые позволили бы нам определить

временные линии или, соответственно, пространственную геометрию. Поэтому локально у нас есть выбор, как трактовать данную модель описания Вселенной: соответствует ли она $K > 0$, $K = 0$ или $K < 0$. Тем не менее глобально эти три ситуации различаются, что видно на рис. 3.4 *б–г*: в каждом из этих случаев нарезка захватывает иную часть целостного де-ситтеровского пространства. Далее я буду исходить из того, что **T** не равен нулю и обеспечивает положительную плотность энергии вещества, что позволяет хорошо определить и линии времени, и пространственноподобные 3-поверхности постоянного времени для каждого значения t, как показано на рис. 3.2.

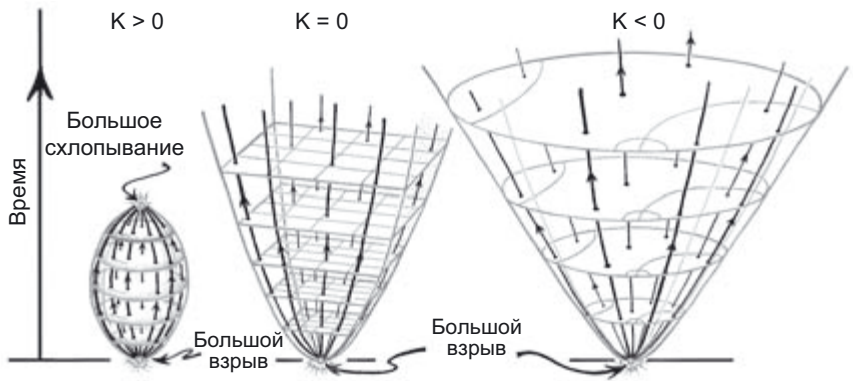


Рис. 3.5. Фридмановские модели с рис. 3.2 с добавлением линий времени (мировых линий идеализированных галактик)

В случае с положительной кривизной пространства ($K > 0$) в стандартной Вселенной Фридмана, наполненной пылью, охарактеризовать ее «размер» можно, используя радиус R 3-сферических пространственных сечений, и исследовать этот размер как функцию t. При $\Lambda = 0$ находим функцию $R(t)$, описывающую *циклоиду* в плоскости (R, t) (при этом скорость света принимается за единицу: $c = 1$). Циклоида — это кривая с простой геометрической характеристикой: она описывается точкой окружности, катящейся по оси t (рис. 3.6 *б*). Отметим, что (после момента времени $\pi R_{\max}$) значение R вновь достигает нуля, как и при Большом взрыве, поэтому вся модель Вселенной с $0 < t < \pi R_{\max}$ вновь схлопывается в сингулярность, и этот момент часто именуется *большим схлопыванием*.

В оставшихся случаях $K < 0$ и $K = 0$ (с нулевой Λ) Вселенная будет бесконечно расширяться и *большого схлопывания* не будет. В случае $K < 0$ есть «радиус», аналогичный R, но для $K = 0$ можно просто выбрать произвольную пару миро-

вых линий идеализированных галактик и взять в качестве R отрезок, разделяющий их в пространстве. В случае $K = 0$ скорость расширения асимптотически стремится к нулю, а в случае $K < 0$ — к некоторому положительному значению. Современные наблюдения свидетельствуют в пользу того, что Λ, скорее всего, положительно и его величина достаточна, чтобы играть решающую роль в скорости расширения Вселенной, поэтому значение K утрачивает важность для этой динамики, и Вселенная в конечном итоге срывается в ускоренное расширение, как показано на рис. 3.3.

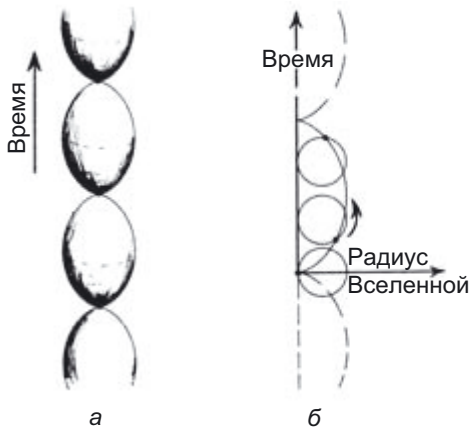


Рис. 3.6. Осциллирующая модель Фридмана ($K > 0$, $\Lambda = 0$) (*а*) и ее радиус как функция времени, представляющая собой циклоиду (*б*)

На заре релятивистской космологии модель с положительным значением K (и $\Lambda = 0$) часто именовалась *осциллирующей* (рис. 3.6 *а*), поскольку циклоидная кривая будет продолжаться бесконечно, если мы позволим «обручу» сделать более одного оборота (штриховая кривая на рис. 3.6 *б*). Можно предположить, что непрерывно сменяющие друг друга участки циклоиды могут соответствовать последовательным циклам в истории реальной Вселенной, где под действием некоей встряски каждое схлопывание, которое претерпевает Вселенная, сменяется новым Большим взрывом. Подобная возможность возникает также при $K \leq 0$, и можно предположить, что на более раннем этапе пространство-время претерпело коллапс, идентичный повороту времени вспять на этапе расширения, и Большое схлопывание того этапа совпадает с Большим взрывом, который мы считаем началом нынешнего расширения Вселенной. Опять же, пришлось бы вообразить себе некий отскок , который каким-то образом позволяет превратить схлопывание в расширение.

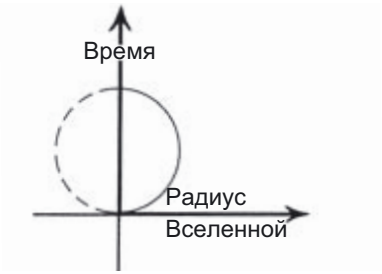


Рис. 3.7. В модели Толмана, где доминирует излучение ($K > 0$, $\Lambda = 0$), радиус как функцию времени можно изобразить в виде полукруга

Однако чтобы такая картина стала физически правдоподобной, требуется представить некоторую убедительную математическую схему, которая бы находилась в согласии с современными физическими представлениями и методами и в которую вписывался бы такой отскок. Например, предположим, что можно изменить взятые на вооружение Фридманом уравнения состояния, при помощи которых он пытался описать общее распределение материи в его «равномерно размазанных» галактиках. Фридман пользовался приблизительной моделью, которая иногда называется *пылевой*; в этой модели не учитываются никакие взаимодействия (кроме гравитации) между «составными элементами» (то есть «галактиками»), мировые линии которых являются линиями времени. Если изменить уравнения состояния, это может существенно отразиться на свойствах $R(t)$ вблизи $t = 0$. Еще более точной аппроксимацией, нежели пыль Фридмана (в период сразу после Большого взрыва), представляется такое уравнение состояния, которым позже пользовался Ричард Чейз Толман [1934], американский специалист по математической физике и космологии. В толмановских ФЛРУ-моделях использовалось уравнение состояния *чистого излучения*. Считается, что оно хорошо аппроксимирует состояние материи на самых ранних этапах развития Вселенной, когда было так жарко, что на каждую частицу приходилось значительно больше энергии, чем по уравнению $E = mc^2$ для массы m даже наиболее тяжелых частиц, которые могли существовать сразу после Большого взрыва. В схеме Толмана для случая $K > 0$ кривая $R(t)$ является не дугой циклоиды, а (при правильно подобранном масштабе R и t) образует *полукруг* (рис. 3.7). В случае с пылевой моделью можно было бы обосновать переход от Схлопывания к Взрыву, прибегнув к *аналитическому продолжению* (см. раздел А.10), которое действительно позволяет переходить от одной дуги циклоидной кривой к следующей таким математическим методом. Но в толмановской модели с чистым излучением аналитическое продолжение просто дополнило бы полукруг и превратило его в круг, а это не имеет никакого смысла, если данная

процедура интересует нас для описания отскока, то есть должна допускать продолжение в сторону отрицательных значений t.

Для того чтобы новое уравнение состояния описывало механизм отскока, необходимо что-то гораздо более радикальное, чем излучение Толмана. В данном случае заслуживает внимания такой серьезный момент: если отскок происходит при некотором несингулярном переходе, в процессе которого сохраняются гладкость пространства-времени и пространственная симметрия модели, то сходящиеся временные линии фазы сжатия могут превратиться в расходящиеся временные линии фазы расширения, пройдя через «бутылочное горлышко», которое объединило бы обе эти фазы. Если бы это горлышко было гладким (несингулярным), то превращение такого экстремального схождения временных линий в экстремальное расхождение было бы достижимо при невероятной кривизне горлышка, что приводило бы к сильному отталкиванию, а это грубо противоречит стандартным условиям положительности энергии, которым удовлетворяет обычная классическая материя (см. разделы 1.11, 3.2 и 3.7; [Hawking and Penrose, 1970]).

Поэтому нельзя рассчитывать, что какое-либо разумное классическое уравнение состояния позволило бы нам описать отскок в контексте ФЛРУ-моделей, и неизбежно встает вопрос: а не помогли бы нам продвинуться в этом направлении уравнения квантовой механики? Необходимо учитывать, что вблизи классической ФЛРУ-сингулярности кривизна пространства-времени становится неопределенно велика. Если бы мы попытались описать такую кривизну в терминах ее радиуса, то этот радиус (величина, обратная кривизне), был бы соответственно мал. Продолжая придерживаться понятий классической геометрии, мы с приближением к классической сингулярности получали бы все меньшие радиусы кривизны пространства-времени, и в итоге радиус стал бы даже меньше планковских масштабов порядка 10^{-33} см (см. разделы 1.1 и 1.5). Большинство теоретиков, размышляя о квантовой гравитации, предполагают, что при таких масштабах пространство-время уже резко отличалось бы от своего привычного вида (гладкое многообразие) (хотя в разделе 4.3 я выдвину совершенно иные аргументы на этот счет). Так это или нет, но нет никаких оснований сомневаться в том, что процедуры общей теории относительности неизбежно придется модифицировать, чтобы они сочетались с методами квантовой механики на подступах к такой радикально искривленной пространственно-временной геометрии. То есть нам нужна подходящая к нашему случаю теория *квантовой гравитации*, которая позволила бы справиться с ситуациями, в которых классические эйнштейновские процедуры приводят к сингулярности (но ср. с разделом 4.3).

Часто приходится слышать утверждения, что подобный прецедент уже был. Как отмечалось в разделе 2.1, в начале XX века возникла серьезная проблема

с классическими представлениями об атоме, поскольку, согласно теории, атомы должны были катастрофически схлопываться в сингулярное состояние, когда электроны по спирали падали бы на ядро (с генерацией импульса излучения), и решить эту проблему удалось лишь с появлением квантовой механики. Не следует ли ожидать, что и при обсуждении подобного катастрофического коллапса всей Вселенной ситуация могла бы проясниться на уровне квантовой механики? Но вот загвоздка: даже сейчас не существует общепризнанной гипотезы квантовой гравитации. Еще серьезнее тот факт, что большинство из уже выдвигавшихся гипотез не решают проблему сингулярностей — сингулярности остаются даже в квантованной теории. Есть некоторые заслуживающие внимания исключения — гипотезы, связанные с несингулярным квантовым отскоком [Bojowald, 2007; Ashtekar et al., 2006], но мне придется вернуться к этой теме в разделах 3.9 и 3.11 (а также в разделе 4.3), где я утверждаю, что подобные гипотезы на самом деле не дают особых надежд на решение проблемы сингулярности именно в нашей Вселенной.

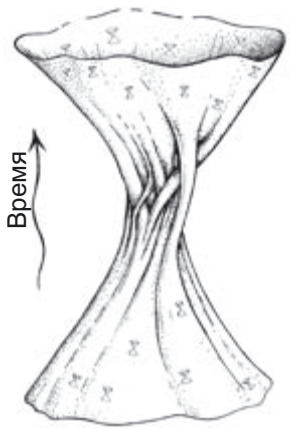


Рис. 3.8. Гипотетическая отскакивающая Вселенная, в которой допускается крайняя нерегулярность, которая позволила бы перейти от коллапса к расширению, минуя сингулярность

Совершенно иная возможность избежать сингулярности связана с ожиданиями, что небольшие отклонения от точной симметрии, присутствующие на этапе коллапса Вселенной, могли бы радикально расти по мере приближения Большого схлопывания, поэтому непосредственно перед полным коллапсом структура пространства-времени далеко не точно соответствовала бы ФЛРУ-модели. Поэтому часто озвучивается надежда на то, что сингулярность, которая проявляется в ФЛРУ-моделях, может быть ложной и что в более общей асимметричной

ситуации такие классические пространственно-временные сингулярности просто не возникнут; следовательно, есть основания ожидать, что в общем случае коллапсирующая Вселенная ввиду некоей сложной промежуточной пространственно-временной геометрии (рис. 3.8) может перейти к нерегулярному расширению. Даже сам Эйнштейн пытался выдвигать такие аргументы — о том, что сингулярности можно избежать путем отскока от нерегулярного коллапса [Einstein, 1931; Einstein and Rosen, 1935] или потому, что конечному коллапсу и сингулярности могут как-то воспрепятствовать орбитальные движения небесных тел [Einstein, 1939].

Можно утверждать, что после такого почти сингулярного (но не строго сингулярного) коллапса возникнет состояние, возмущения которого станут постепенно разглаживаться, и в результате оно станет сильно напоминать расширяющуюся ФЛРУ-модель (как на рис. 3.8). В 1963 году эту задачу подробно проанализировали два советских физика-теоретика — Евгений Михайлович Лифшиц и Исаак Маркович Халатников [Лифшиц и Халатников, 1963]. Их работа показывает, что в обычных условиях такие сингулярности, по-видимому, *не* возникают, что подкрепляет описанную выше гипотезу несингулярного отскока. Соответственно, утверждалось, что в общей теории относительности пространственно-временные сингулярности, возникающие при гравитационном коллапсе и проявляющиеся в известных точных решениях коллапсирующих моделей Фридмана или других ФЛРУ-моделей, порождаются лишь потому, что известные решения обладают нереалистичными специфическими свойствами, например строгой симметрией. Поэтому такие сингулярности не сложились бы в условиях типичных асимметричных пертурбаций. Однако это предположение *не* подтвердилось, о чем и пойдет речь в следующем разделе.

3.2. Черные дыры и локальные нерегулярности

В 1964 году я стал всерьез обдумывать смежную проблему, связанную с более локальными явлениями, такими как гравитационный коллапс звезды или совокупности звезд, в результате которого образуется так называемая *черная дыра*. Черные дыры оставались на втором плане, пока в 1930 году выдающийся 19-летний физик Субрахманьян Чандрасекар не показал (см. [Wali, 2010; Chandrasekhar, 1931]), что существует предел массы (составляющий около 1,4 солнечной), которого может достичь белый карлик, — при превышении этого предела он катастрофически схлопнется под действием гравитации. Белый карлик — это исключительно компактная звезда. Одним из первых обнаруженных белых карликов оказался загадочный объект, сопровождающий ярчайшую звезду

Сириус. Эта крошечная звездочка, Сириус B, по массе примерно равна Солнцу, однако по диаметру не больше Земли, которая примерно в 10^6 раз меньше Солнца по объему. По сути, такой белый карлик уже израсходовал весь запас своего ядерного топлива и держится исключительно под действием так называемого *давления вырожденного электронного газа*. Такое давление обусловлено принципом запрета Паули (см. раздел 1.14), которому подчиняются электроны, и именно поэтому электроны не могут сближаться слишком плотно. Чандрасекар показал, что существует фундаментальный предел для максимальной массы белого карлика, при превышении которого давление вырожденного электронного газа уже не в состоянии предотвратить катастрофическое сжатие звезды.

Иногда могут возникать еще более плотные звезды, удерживаемые от коллапса так называемым *давлением вырожденного нейтронного газа*. При чрезвычайно больших давлениях электроны соединяются с протонами, в результате чего образуются нейтроны, и уже на них начинает действовать принцип запрета Паули [Ландау, 1932]. Нам уже удалось наблюдать множество таких *нейтронных* звезд, плотность которых достигает невероятных значений — она сравнима с плотностью атомного ядра, но иногда бывает и выше. В таком случае масса целого солнца оказывается сконцентрирована в радиусе не более 10 км, то есть нейтронная звезда в 10^{14} раз меньше Солнца по объему. Часто нейтронные звезды обладают сильнейшими магнитными полями и могут стремительно вращаться. В результате взаимодействия быстро вращающегося магнитного поля с окружающим нейтронную звезду заряженным веществом возникают электромагнитные импульсы, которые можно зафиксировать на Земле, даже если их источник удален от нее на 10^5 световых лет. Такой источник называется пульсаром, он ведет себя подобно проблесковому маячку. Тем не менее и в данном случае есть предел — подобный пределу Чандрасекара, но именуемый в честь *Ландау*, соответствующий максимальной массе, которую может набрать нейтронная звезда. Точное значение этого предела пока неизвестно, но вряд ли оно больше двух солнечных масс. Наиболее массивная из известных нейтронных звезд (на момент написания книги) — это пульсар, вращающийся вокруг белого карлика по очень близкой орбите (с периодом 2,5 часа) и образующий вместе с ним двойную систему J0348+0432; его масса оценивается как две солнечных.

Что касается локальных физических процессов, насколько мы их сегодня понимаем с теоретической точки зрения, действительно, нет никакой иной возможности сдержать коллапс более массивных объектов. Однако известны гораздо более массивные звезды и концентрированные звездные скопления, поэтому возникают фундаментальные вопросы: какая судьба может их ожидать, когда гравитационный коллапс наконец постигнет такие объекты — например, если очень большая звезда окончательно израсходует свое ядерное горючее. Чандра-

секар, написавший на эту тему в 1934 году революционную статью, осторожно прокомментировал проблему так:

> «Жизненный цикл звезды с небольшой массой должен принципиально отличаться от жизненного цикла очень массивной звезды. Для небольшой звезды естественная стадия белого карлика — это первый шаг к полному исчезновению. Если же масса звезды превышает некое критическое значение m, то такая звезда не может стать белым карликом, поэтому остается рассматривать другие возможности».

Напротив, многие другие по-прежнему относились к вопросу скептически. Наиболее характерен комментарий выдающегося британского астрофизика сэра Артура Эддингтона, сделанный им в 1935 году:

> «Звезда будет продолжать излучать, сжимаясь и сжимаясь, пока, как я полагаю, она не уменьшится до размера в несколько километров, когда гравитация станет столь сильной, что будет удерживать излучение, и звезда наконец упокоится с миром... Думаю, должен существовать закон природы, не позволяющий звездам вести себя столь абсурдным образом!»

Эта проблема стала особенно актуальна в начале 1960-х годов, и наиболее активно ее прорабатывал выдающийся американский физик Джон Арчибальд Уилер, в особенности потому, что как раз в 1963 году голландский астроном Мартен Шмидт открыл первый *квазар* 3C 273 (позже квазарами стали называться все подобные объекты). С учетом того как далеко от нас находился этот объект (расстояние можно было определить по его красному смещению), его собственная яркость была экстраординарной: более чем в 4×10^{12} раз выше, чем у Солнца, так что излучение квазара более чем в 100 раз превышало излучение всей нашей Галактики. Учитывая, что сам квазар при этом был относительно мал — сравним по размеру с нашей Солнечной системой, что можно было понять по характерным стремительным изменениям его светимости с периодичностью в несколько дней, — такой колоссальный выход энергии позволил астрономам заключить, что центральный объект, порождающий эти выбросы, должен обладать чудовищной массой, но при этом быть исключительно компактным и не превышать собственный *радиус Шварцшильда*. Этот критический радиус сферически симметричного тела с массой m вычисляется по формуле:

$$\frac{2\gamma m}{c^2},$$

где γ — гравитационная постоянная Ньютона, а c — скорость света.

Здесь следует вкратце объяснить, что представляет собой этот радиус. Он связан с хорошо известным *решением Шварцшильда* уравнений Эйнштейна (**G = 0**, см. раздел 1.1) в вакууме для гравитационного поля, окружающего статическое сферически симметричное массивное тело (идеализированную звезду). Это решение нашел немецкий физик и астроном Карл Шварцшильд вскоре после того, как в конце 1915 года Эйнштейн полностью сформулировал общую теорию относительности. Через некоторое время сам Шварцшильд трагически скончался от пузырчатки, поразившей его на Восточном фронте в Первую мировую войну. Если вообразить, что коллапсирующее тело стало бы симметрично схлопываться и что решение Шварцшильда так и оставалось бы верным на протяжении всего этого процесса — чего, собственно, и требуют уравнения, то метрика в координатном выражении достигла бы сингулярности на радиусе Шварцшильда. Большинство физиков (включая самого Эйнштейна) полагали, что геометрия реального пространства-времени в этой точке обязательно станет сингулярной по структуре.

Однако позже удалось осознать, что радиус Шварцшильда — это не пространственно-временная сингулярность, а тот самый радиус, по достижении которого коллапсирующая (сферически симметричная) материя становится черной дырой. Любой сферический объект, сжатый до собственного радиуса Шварцшильда, необратимо схлопывается и стремительно исчезает из вида. Высказывалась версия, что энергетические выбросы, наблюдаемые на 3C 273, должны возникать в ходе бурных процессов, сопровождающих такой гравитационный коллапс, и происходить в областях, непосредственно прилегающих к радиусу Шварцшильда. Прежде чем звезды и другая материя окажутся поглощены черной дырой, они претерпевают действие колоссальных сил и чрезвычайно разогреваются.

Гравитационный коллапс с образованием черной дыры (которая предположительно обладает идеальной сферической симметрией) во многом напоминает ситуацию, происходящую с фридмановскими моделями, и для этой ситуации известно точное решение эйнштейновских уравнений — на этот раз речь идет о решении, полученном Оппенгеймером и Снайдером в 1939 году. Уравнение дает полную геометрическую пространственно-временную картину такого коллапса для симметричной сферы. Тензор энергии **T** для коллапсирующей материи точно такой, как у фридмановской пыли. Действительно, «материальная» часть их решения в точности соответствует одной из составляющих фридмановской пылевой модели — как часть коллапсирующей Вселенной. В решении Оппенгеймера — Снайдера есть сферически симметричное распределение вещества (пыли), которое в процессе коллапса пересекает радиус Шварцшильда, и уже в центре этой структуры возникает пространственно-временная сингулярность.

Там плотность коллапсирующей материи и кривизна пространства-времени становятся бесконечными.

Сам радиус Шварцшильда оказывается сингулярностью только в таких статических координатах, которыми пользовался сам Шварцшильд, хотя долго существовало ошибочное мнение о том, что этот радиус и есть подлинная физическая сингулярность. Любопытно отметить: первым человеком, осознавшим, что радиус Шварцшильда — это всего лишь координатная сингулярность и что данное решение можно свободно расширить за пределы этого радиуса к истинной сингулярности, локализованной в центре, по-видимому, был математик Поль Пенлеве, пришедший к такому выводу в 1921 году [Painlevé, 1921]. При этом Пенлеве дважды ненадолго становился премьер-министром Франции — в 1917 и 1925 годах. Однако создается впечатление, что это теоретическое заключение осталось практически незамеченным среди физиков, занимавшихся относительностью, — действительно, в те времена возникла большая путаница в отношении интерпретаций теории Эйнштейна. Впоследствии, в 1932 году, аббат Жорж Леметр четко показал, что падающее вещество может пересечь радиус Шварцшильда, не встретив сингулярности [Lemaître, 1933]. Более простое описание такой геометрии гораздо позже (в 1958 году) дал Дэвид Финкельштейн [Finkelstein, 1958], воспользовавшийся формой шварцшильдовской метрики; что любопытно, это описание еще в 1924 году было найдено Эддингтоном для совершенно иной цели [Eddington, 1924], не связанной с гравитационным коллапсом!

Сегодня поверхность радиуса Шварцшильда именуется (абсолютным) *горизонтом событий*. По причинам, которые я проясню чуть позже, вещество может упасть внутрь горизонта, но выйти обратно не может. Возникает вопрос: а если у нас будут отклонения от точной сферической симметрии и/или мы воспользуемся более общими уравнениями состояния, нежели уравнения для несжимаемой пыли, задействованные в данном случае Оппенгеймером и Снайдером, то возможно ли, что коллапс не приведет к возникновению сингулярности и мы сможем представить себе очень сложную — однако несингулярную — промежуточную конфигурацию, через которую произойдет «отскок» коллапса к нерегулярному расширению первоначально падавшей материи?

Пространственно-временное описание модели Оппенгеймера — Снайдера проиллюстрировано на рис. 3.9 (одно пространственное измерение я убрал). Ключевые геометрические свойства иллюстрируются нулевыми конусами (см. рис. 1.8 *б*), и любое распространение информации может происходить только в пределах этих конусов (см. раздел 1.7). Эта картинка основана на вышеупомянутом описании Финкельштейна [Finkelstein, 1958]. Отметим, что из-за присутствия очень плотного коллапсирующего вещества конусы будут значительно

искажены и сужены, причем тем больше, чем ближе они расположены к центру. Поэтому на определенном расстоянии от центра образующая конуса, направленная в будущее, становится вертикальной, и это означает, что испущенные в этой точке сигналы уже не могут попасть во внешний мир. Это расстояние и есть шварцшильдовский радиус коллапсара. Рисунок позволяет убедиться в том, что коллапсирующее вещество может провалиться внутрь горизонта, но после этого теряет всякую возможность обмениваться информацией с окружающим миром. В центре находим пространственно-временную сингулярность, где искривления пространства-времени *действительно* расходятся до бесконечности, а коллапсирующее вещество достигает бесконечной плотности. Все (времениподобные) мировые линии, оказавшись внутри сферы Шварцшильда, попадают в сингулярность — независимо от того, лежат ли они внутри коллапсирующего вещества либо просто следуют за ним. Пути назад нет!

Возникает интересная проблема, если сравнить эту теорию с ньютоновской. Нередко отмечалось, что именно этот радиус важен и для ньютоновской гравитации, — впервые на это указал еще в 1783 году британский ученый преподобный

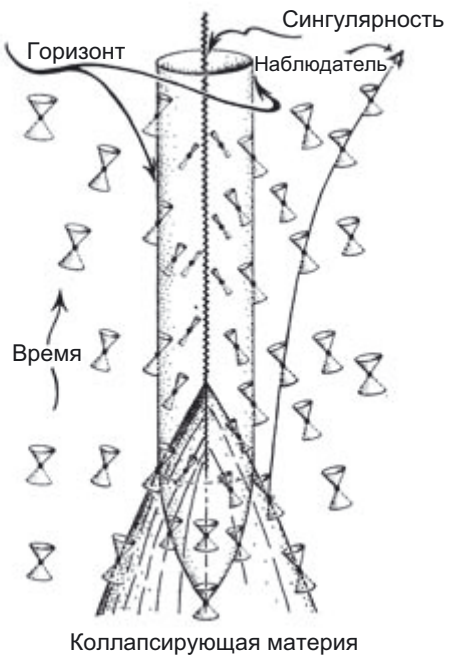


Рис. 3.9. Стандартное пространственно-временное изображение гравитационного коллапса с образованием черной дыры. Наблюдатель, находящийся за пределами горизонта, не может видеть, что происходит под горизонтом

Джон Мичелл [1783]. Воспользовавшись теорией Ньютона, он получил как раз радиус Шварцшильда, рассуждая, что если массивное сферическое тело обладает таким радиусом, то излучаемый им свет будет падать обратно и не сможет от него оторваться. Со стороны Мичелла эта догадка была удивительно провидческой, но вывод его можно серьезно критиковать, поскольку, согласно теории Ньютона, скорость света не постоянна, и для ньютоновского тела такого размера скорость света была бы гораздо выше, как если бы свет падал на эту поверхность с очень большого расстояния. На самом деле описание черной дыры можно сформулировать лишь исходя из конкретных свойств общей теории относительности, и в теории Ньютона такой объект невозможен (см. [Penrose, 1975a]).

Далее, примерно так же как в случае с коллапсирующими ФЛРУ-космологиями, возникает вопрос: могут ли отклонения от идеальной сферической симметрии давать принципиально иную картину? Можно предположить, что если коллапсирующее вещество не обладает подразумеваемой здесь идеальной сферической симметрией, то эти отклонения вполне могут усиливаться по мере приближения к центру, и, следовательно, удается избежать бесконечных плотностей и бесконечной кривизны пространства-времени (см. рис. 3.8 в разделе 3.1). С этой точки зрения сингулярности возникают лишь благодаря такой точной фокусировке поступающей материи, стремящейся к центральной точке. Соответственно, хотя плотность все равно может быть очень высока, не предполагается, что она может достигнуть бесконечных значений. Не исключено, что после некоего бурного и сложного завихрения и расплескивания материя может вновь сформироваться в каком-либо виде, а истинная сингулярность при этом так и не будет достигнута. Такая идея как минимум возникала.

Осенью 1964 года эта проблема стала меня всерьез беспокоить, и я задумался, смогу ли ответить на этот вопрос, если воспользуюсь некоторыми математическими идеями, которые ранее развивал в контексте модели стационарной Вселенной (в 1950-е годы эту модель описали Герман Бонди, Томас Голд и Фред Хойл (см. [Sciama, 1959, 1969])). Согласно этой модели, Вселенная не имела начала, расширялась непрерывно и до бесконечности. В результате расширения материя распределялась все более тонким слоем, но такое разрежение компенсировалось за счет новой материи (прежде всего водорода), которая постоянно, но очень медленно возникала во Вселенной. Я хотел проверить, можно ли избежать явного противоречия между моделью стационарной Вселенной и стандартной общей теорией относительности (согласно которой обычная материя должна обладать положительной энергией — об этом требовании я упоминал выше, в разделе 3.1), если допустить наличие отклонений от полной симметрии, которая предполагается в стандартных представлениях о стационарной Вселенной. Вооружившись геометрическим/топологическим аргументом, я пришел к убеж-

дению, что такие отклонения от симметрии *не снимают* противоречий. Я никогда не публиковал эти рассуждения, но опирался на схожие идеи в ином контексте, применяя их (довольно строго, но не абсолютно строго) при описании асимптотической структуры гравитационно излучающей системы [Penrose, 1965 b, приложение]. Данные методы очень отличаются от тех, что обычно применяются в общей теории относительности, приемы которой зачастую связаны с поиском точных частных решений или с масштабным численным моделированием.

Что касается гравитационного коллапса, я также стремился показать, что всякий раз, когда такой коллапс приводит к падению под горизонт, наличие отклонений от симметрии (и применение более общих уравнений состояния, нежели те, которые описывают пыль Фридмана, излучение Толмана и т. п.) существенно не влияет на традиционную картину Оппенгеймера — Снайдера: всегда должна возникать какая-либо сингулярность, препятствующая по-настоящему плавной эволюции. Необходимо учитывать, что есть и множество других ситуаций, в которых тело может относительно мягко сжиматься под действием гравитации, и в таких случаях, пожалуй, наличие иных сил может приводить к возникновению стабильной конфигурации или же к какому-либо отскоку. Следовательно, нам нужен подходящий критерий, который позволил бы характеризовать необратимый коллапс, проявляющийся в ситуациях типа описанной Оппенгеймером и Снайдером (см. рис. 3.9). Разумеется, к этому критерию предъявляется следующее ключевое требование: его нельзя выводить из каких-либо предположений о симметрии.

Серьезно поразмыслив, я стал понимать, что никакая полностью *локальная* характеристика не позволяет достичь того, что нам здесь требуется, равно как никакое из свойств полной или средней меры, скажем кривизны пространства-времени, не имеет практического применения. В итоге я сформулировал понятие *ловушечная поверхность* , и наличие такой поверхности в пространстве-времени — серьезный признак того, что необратимый коллапс действительно имел место. (Тех, кто заинтересовался, отсылаю к работе Penrose [1989, p. 420], где описаны любопытные обстоятельства, в которых меня посетила эта идея.) Технически ловушечная поверхность — это замкнутая пространственноподобная 2-поверхность, все *нуль-нормальные* направления на которой (данная концепция проиллюстрирована на рис. 3.10 *a* (см. также раздел 1.7)) сходятся в будущем. Термин «нормальный» означает в обычной евклидовой геометрии «под прямым углом» (см. рис. 1.18 в раздел 1.7), а из рис. 3.10 видно, что направленные в будущее нулевые нормали соответствуют направлениям лучей света (то есть нулевым геодезическим), которые выходят из 2-поверхности перпендикулярно ей, если рассматривать их в любой мгновенной пространственноподобной 3-поверхности, содержащей заданную 2-поверхность.

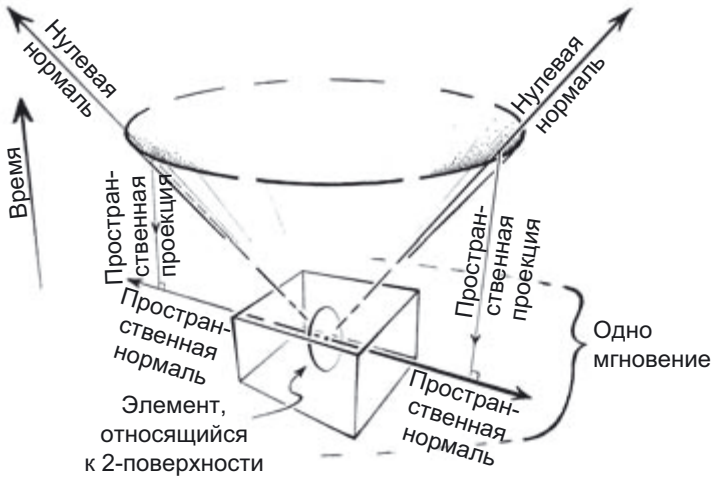


Рис. 3.10. Нуль-нормальные направления, ведущие к пространственноподобной 2-поверхности, — это направления лучей света, исходящих из 2-поверхности под прямым углом к ней, если рассматривать их в любой мгновенной пространственноподобной 3-поверхности, содержащей заданную 2-поверхность

Чтобы уяснить, что все это означает в пространственном контексте, представьте себе гладкую двумерную искривленную поверхность $\mathcal{S}$ в обычном евклидовом трехмерном пространстве. Вообразите себе вспышку света, которая одновременно охватывает всю поверхность $\mathcal{S}$, и посмотрите, как волновой фронт излученного света станет распространяться от $\mathcal{S}$ с одной ее стороны или с другой (рис. 3.11 *а*). Там, где поверхность $\mathcal{S}$ искривлена, волновой фронт с вогнутой стороны начнет сходиться, а на выпуклой стороне — расходиться. Однако вот что происходит с ловушечной поверхностью S: с *обеих* ее сторон волновые фронты начинают сходиться (рис. 3.11 *б*)! На первый взгляд может показаться, что такое локальное условие неосуществимо на обычной пространственноподобной 2-поверхности, однако на самом деле в пространстве-времени дела обстоят иначе. Даже в плоском пространстве-времени (пространство Минковского; см. рис. 1.23 в разделе 1.7) можно с легкостью строить *локально* ловушечные 2-поверхности. Простейший пример: возьмем $\mathcal{S}$ на пересечении двух световых конусов прошлого, вершины которых P и Q пространственноподобно разделены (рис. 3.11 *в*). Здесь все нулевые нормали, идущие к $\mathcal{S}$, сходятся в будущем в направлении к P или Q (а вот почему это противоречит нашим интуитивным представлениям о двумерных плоскостях в евклидовом трехмерном пространстве: ведь такая $\mathcal{S}$ не может содержаться в пределах единственного плоского евклидового трехмерного пространства или «среза времени»). Правда, данная конкретная $\mathcal{S}$ не

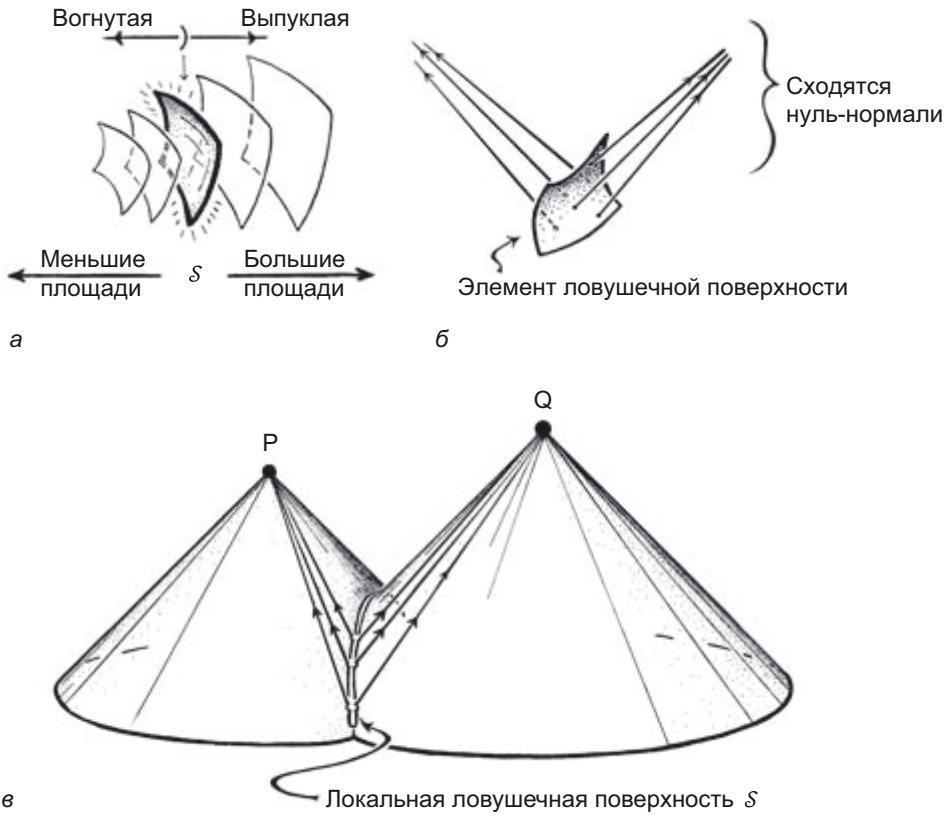


Рис. 3.11. Условия ловушечной поверхности: *а*) в обычном евклидовом трехмерном пространстве вспышка света, происходящая одновременно на всей искривленной 2-поверхности S, будет сокращать площадь волнового фронта, если происходит на вогнутой стороне, и увеличивать, если происходит на выпуклой; *б*) для любого локального участка S на ловушечной поверхности лучи света будут сходиться с *обеих* сторон; *в*) такое явление «локального захвата» в пространстве-времени для некомпактной поверхности S не является аномальным, оно происходит в пространстве Минковского на пересечении двух световых конусов прошлого

ловушечная поверхность, поскольку она не является *замкнутой* (то есть компактной; см. раздел А.3). В некоторых работах термином *замкнутая ловушечная поверхность* именуется то, что я называю просто *ловушечной поверхностью*; см., например, [Hawking and Ellis, 1973]. Следовательно, условие, необходимое для того, чтобы в пространстве-времени могла содержаться ловушечная поверхность, действительно является нелокальным. В пространстве-времени Оппенгеймера — Снайдера имеются настоящие (то есть замкнутые) ловушечные поверхности, и они возникают в области, ограниченной радиусом Шварцшильда, после того как

произойдет коллапс. При этом сама природа ловушечной поверхности такова, что любые достаточно малые возмущения, ведущие к такому коллапсу, также должны содержать ловушечные поверхности независимо от каких-либо соображений симметрии. (Сложнее, что здесь мы имеем дело с так называемым *открытым* условием, то есть достаточно малые изменения не нарушают это условие.)

Теорема [Penrose, 1965a], которую мне удалось сформулировать в конце 1964 года, в сущности, показала, что если в пространстве-времени возникает ловушечная поверхность, это приводит к пространственно-временной сингулярности. Точнее, я показал, что если пространство-время (подчиняющееся некоторым ограничениям, обоснованным с физической точки зрения, — к ним я перейду чуть позже) содержит ловушечную поверхность, то оно не может неограниченно продолжаться в будущем. Именно невозможность такого продолжения — признак наличия сингулярности. Эта теорема не приводит к бесконечной кривизне или к бесконечной плотности, но в обычных обстоятельствах сложно себе представить, какие еще преграды могли бы препятствовать эволюции пространства-времени в направлении будущего. Теоретически существуют и другие возможности, но они не возникают в типичных обстоятельствах (то есть возникают лишь с ограниченной функциональной свободой; см. разделы A.2 и A.8).

Теорема также основана на допущении, что эйнштейновские уравнения остаются применимыми (с космологической постоянной Λ или без нее) при тензоре энергии $\mathbf{T}$, который удовлетворяет так называемому *нулевому энергетическому условию* (согласно которому для любого нулевого вектора $\mathbf{n}$ величина, получаемая путем двойной свертки $\mathbf{n}$ с $\mathbf{T}$, обязательно будет неотрицательной)[1]. Это очень слабое требование к источникам гравитации, и оно соблюдается для любого физически мыслимого классического вещества. Другое допущение, которое мне потребовалось, заключалось в следующем: пространство-время должно возникать в ходе обычной временной эволюции из некоторого пространственно-неограниченного исходного состояния, то есть технически из некомпактной (а значит, «открытой», см. раздел A.5) исходной пространственноподобной 3-поверхности. В принципе, это позволяло заключить, что в локальных ситуациях гравитационного коллапса, как только возникает ловушечная поверхность, сингулярности неизбежно возникают с любыми физически возможными классическими веществами, независимо от каких-либо допущений о симметрии.

---

[1] В индексной нотации это условие записывается в виде $T_{ab}n^a n^b \geq 0$, когда $n^a n_a = 0$. В некоторых моих работах (см., например, [Penrose, 1969a], с. 264) я называл его *слабым энергетическим условием*, что вызывало некоторую путаницу, потому что Хокинг и Эллис [1973] использовали *второй* термин в ином (более строгом) смысле.

Разумеется, все равно оправдан вопрос: а могут ли вообще ловушечные поверхности возникать в каких-либо вообразимых астрофизических ситуациях? В частности, можно предположить, что если существуют физические тела еще плотнее нейтронных звезд, то известная нам физика частиц вполне может недостаточно адекватно описывать процессы, которые происходили бы в таких условиях. Однако эта проблема не является существенной, поскольку ожидается, что гравитационный коллапс может возникать и в других ситуациях, когда ловушечные поверхности должны образовываться и при вполне изученных плотностях. Дело, собственно, заключается в том, как относится теория относительности к общему изменению масштаба. Если бы у нас была такая модель пространства-времени, метрика которой задавалась бы тензорным полем **g** (см. раздел 1.1 и 1.7) и которая удовлетворяла бы эйнштейновским уравнениям для тензора источника энергии **T** (и для космологической постоянной Λ), то при замене **g** на величину $k\mathbf{g}$, где k — некоторое постоянное положительное число, окажется, что уравнения Эйнштейна вновь удовлетворяются для тензора энергии **T** (и при космологической постоянной $k^{-1}\Lambda$, но эту крошечную составляющую можно игнорировать). То, как плотность вещества ρ закодирована в тензоре **T**, подразумевает[1], что ρ, соответственно, должна быть заменена на $k^{-1}\rho$. Следовательно, если у нас есть некая модель коллапса, в которой возникает ловушечная поверхность, когда плотность достигает некоторого значения ρ, то можно построить и другую модель, где ловушечная поверхность все равно возникает, но значение ρ будет настолько мало, насколько мы захотим, — достаточно лишь правильно откалибровать метрику. Если у нас есть модель коллапса, в которой ловушечные поверхности формируются только при неких экстраординарных плотностях (например, гораздо выше, чем в нейтронной звезде), то найдется и другая, гораздо более крупномасштабная модель, где рассматриваются значительно более протяженные расстояния — например, в масштабе центральных регионов галактик, а не нейтронных звезд, — и плотности в этой модели будут не выше, чем обычно наблюдаются здесь, на Земле. Считается, что именно такие условия и должны наблюдаться в окрестностях черной дыры, которая в четыре миллиона раз тяжелее Солнца и, вероятно, расположена в центре нашей Галактики. Определенно, если говорить о квазаре 3C 273, средние значения плотности вблизи его горизонта событий будут гораздо ниже, и ничто не должно мешать образованию ловушечных поверхностей в таких условиях.

Что касается других точек зрения в отношении образования ловушечных поверхностей и относительно их математических трактовок как таковых, отсылаю

---

[1] В индексной нотации имеем $\rho = T_{ab}t^a t^b$, где t^a определяет направление времени для наблюдателя, нормированное в соответствии с $t^a t^b g_{ab} = 1$. Следовательно, t^a масштабируется коэффициентом $k^{-1/2}$, тогда как ρ — коэффициентом k^{-1}.

вас к работам Шена и Яу [1993] и Христодулу [2009]. В работе [Penrose, 1969a] (см., в частности, рис. 3) я привожу простую и интуитивно понятную аргументацию в пользу того, что в фактически эквивалентных условиях повторно сходящийся световой конус легко окажется в гравитационном коллапсе и при сравнительно невысоких плотностях. Такое альтернативное условие, характеризующее необратимый гравитационный коллапс (ведущий к сингулярности), в числе прочих обсуждается с математической точки зрения в работе [Hawking and Penrose, 1970].

Поскольку катастрофические коллапсы такого рода могут происходить в относительно локальных масштабах даже в расширяющейся Вселенной, следует ожидать, что подобные процессы определенно должны проявляться и в глобальном масштабе, например при коллапсе Вселенной, когда наблюдается серьезная иррегулярность в распределении массы коллапсирующего вещества. Соответственно, вышеизложенные соображения остаются актуальными и при глобальном коллапсе целой Вселенной, а сингулярности — характерный аспект гравитационного коллапса в классической общей теории относительности. В 1965 году Стивен Хокинг, тогда еще молодой аспирант, заметил [1965], что стандартная ФЛРУ-модель в фазе коллапса также содержала бы ловушечные поверхности, но в данном случае они были бы колоссальными — сравнимыми по масштабам со всей наблюдаемой частью Вселенной. Таким образом, мы опять приходим к выводу, что сингулярности неизбежно возникают в пространственно-открытой коллапсирующей Вселенной. (Уточнение «открытые» требуется здесь потому, что в теореме 1965 года подразумевалась некомпактная исходная поверхность.) На самом деле, формулируя свой аргумент, Хокинг рассматривал ход времени в обратном направлении, так что аргумент касался ранних стадий развития открытой расширяющейся Вселенной — фактически растревоженного Большого взрыва, а не заключительных этапов, когда Вселенная претерпевает коллапс. Тем не менее, в принципе, посыл Хокинга от этого не меняется: если ввести нерегулярности в стандартные симметричные открытые космологические модели, сингулярности все равно в них сохраняются, точно так же, как и в локальных моделях коллапса [Hawking, 1965]. В серии последующих статей [Hawking, 1966a, b, 1967] Хокинг смог развить свои приемы, так что, например, результирующие теоремы удавалось глобально применять к пространственно-закрытым моделям Вселенной (в таких случаях не требуется условие, связанное с ловушечными поверхностями). Затем, в 1970 году, мы, объединив усилия, получили очень общую теорему, охватывавшую практически все частные случаи, приводившие к сингулярности и выявленные нами ранее [Hawking and Penrose, 1970].

Как со всем этим согласуются выводы Лифшица и Халатникова, упомянутые в конце раздела 3.1? Между двумя этими картинами должны были возникнуть

серьезные противоречия, но Лифшиц и Халатников, узнав о первой из вышеупомянутых теорем о сингулярности (с этой работой они познакомились на международной конференции по проблемам теории относительности, состоявшейся в Лондоне в 1965 году), заручились весьма ценной для них помощью Владимира Белинского и (с помощью Белинского) исправили ошибки в своих более ранних работах, обнаружив, что есть и более общие решения, нежели полученные ими прежде. Их новый вывод заключался в том, что сингулярности, в принципе, должны возникать в ситуациях глобального коллапса, и это полностью соответствовало моим заключениям (к которым впоследствии пришел и Хокинг). Подробный анализ, выполненный Белинским, Лифшицем и Халатниковым, позволил им построить очень сложную модель общей сингулярности [Белинский и др., 1970, 1972]. Она часто именуется *гипотеза БКЛ*. Здесь я буду именовать ее «гипотезой БКЛМ», учитывая, что исходно на нее повлияли работы выдающегося американского физика-теоретика, специалиста по общей теории относительности Чарльза У. Мизнера, независимо предложившего собственную космологическую модель с такими же сложными свойствами немного раньше, чем это сделали советские физики [Misner, 1969].

3.3. Второй закон термодинамики

В сущности, вывод из раздела 3.2 таков: невозможно разрешить проблему пространственно-временной сингулярности в контексте уравнений классической общей теории относительности. Ведь мы убедились, что сингулярности — это не только особенность некоторых известных частных симметричных решений данных уравнений; сингулярности возникают и в совершенно глобальных ситуациях гравитационного коллапса. Однако по-прежнему остается вариант, озвученный ближе к концу раздела 3.1: возможно, нам повезет больше, если мы воспользуемся методами квантовой механики. В целом эти методы относятся к одной из разновидностей уравнения Шрёдингера (см. разделы 2.4, 2.7 и 2.12), где соответствующие физические процессы, в данном случае обусловленные свойствами эйнштейновского искривленного пространства-времени, подчиняющегося общей теории относительности, следует квантовать в соответствии с некоторой схемой квантовой гравитации.

Ключевая проблема, свойственная как уравнению Шрёдингера, так и уравнениям стандартной классической физики (в том числе и уравнениям общей теории относительности), — это свойство симметричности относительно обращения времени, которое должно сохраняться в любой разновидности квантовой гравитации, которая следует стандартным процедурам. Соответственно, какое бы решение квантовых уравнений у нас ни получилось, мы всегда должны быть

в состоянии построить другое, в котором параметр t, означающий время, меняется на $-t$, и в результате у нас всегда должно получаться новое решение этих уравнений. Однако следует отметить, что в случае с уравнением Шрёдингера (в отличие от ситуации с классическими уравнениями) при такой замене мы также должны поменять местами мнимые единицы i и –i. Это означает, что при обращении времени необходимо заменить все присутствующие в уравнениях комплексные величины на комплексно сопряженные им. (Если мы хотим, чтобы эта мера t означала «время, истекшее с момента Большого взрыва», и требуем, чтобы значение t оставалось положительным, то симметрия времени сведется к замене t на $C - t$, где C — некоторое большое постоянное положительное число.) В любом случае, симметричность относительно обращения времени — как раз то свойство, которое мы определенно ожидаем найти при применении хотя бы условно обкатанных процедур квантования к теории гравитации.

Почему симметрия уравнений во времени так важна и так загадочна для нас при обсуждении проблемы пространственно-временной сингулярности? Центральная проблема — это второй закон термодинамики, который я буду здесь для краткости именовать просто «второй закон». Нам предстоит убедиться в том, что этот фундаментальный закон глубоко и неотделимо связан с природой сингулярностей в структуре пространства-времени и что он ставит перед нами вопрос, насколько многообещающими кажутся стандартные квантово-механические процедуры для полного разрешения проблемы сингулярности.

Для того чтобы интуитивно понять второй закон, представим себе какое-нибудь привычное действие, которое кажется совершенно необратимым во времени: допустим, мы разлили воду из чашки и вода впиталась в ковер. Эту ситуацию можно рассмотреть в чисто ньютоновских терминах: каждая из молекул воды в отдельности подчиняется законам стандартной ньютоновской динамики, где частицы ускоряются согласно силам, действующим между ними, а также под действием гравитационного поля Земли. На уровне отдельных частиц все их движения подчиняются законам, которые полностью обратимы во времени. Однако если мы попробуем воспроизвести конкретную ситуацию с расплескиванием воды в обратном порядке, то получим, казалось бы, абсурдную картину: молекулы воды в строгом порядке спонтанно улетучиваются с ковра и устремляются вверх, одновременно собираясь в чашке. Этот процесс полностью согласуется с законами Ньютона (то есть энергия, необходимая для взлета молекул воды и попадания их в чашку, — это та самая тепловая энергия, которая тратилась бы на их спонтанное движение по ковру). Тем не менее такая ситуация никогда не встречается на практике.

Физики описывают такую макроскопическую асимметрию во времени (возникающую, несмотря на то что все субмикроскопические действия во времени

полностью симметричны), вводя понятие *энтропии*. Грубо говоря, энтропия — это степень *явной неупорядоченности* системы. В принципе, второй закон постулирует, что при любых макроскопических физических процессах энтропия системы возрастает со временем (или как минимум не уменьшается, если не учитывать крошечных флуктуаций, несущественных для общего тренда). Итак, второй закон просто фиксирует известный и, казалось бы, довольно удручающий факт: если пустить все на самотек, то с течением времени беспорядок будет только расти!

Вскоре мы убедимся, что такая интерпретация несколько преувеличивает отрицательные аспекты второго закона, и если получше разобраться в проблеме, то сложится гораздо более позитивная и интересная картина. Сначала давайте попробуем немного аккуратнее описать, что такое энтропия некоторого состояния системы. Я должен уточнить термин *состояние*, потому что он практически не связан с концепцией квантового состояния, которую мы обсуждали в разделах 2.4 и 2.5. Здесь я говорю о так называемом *макроскопическом состоянии* (классической) физической системы. При определении макроскопического состояния конкретной системы нас не интересуют ее мельчайшие детали — например, где находятся отдельные частицы или как они могут двигаться. Напротив, нас интересуют средние величины, например распределение температуры в газе или жидкости, плотность и общее направление движения. Нас интересует общий состав среды в разных точках, например концентрации и коллективные движения, скажем, молекул азота (N_2) или кислорода (O_2) либо молекул CO_2, H_2O и т. д., то есть любых составляющих рассматриваемой системы, но не такие детали, как положение или движения отдельных молекул. Зная значения всех таких макроскопических параметров, мы сможем определить макроскопическое состояние системы. Признаться, результат у нас получится немного зыбкий, но практика показывает, что тонкая доводка этих параметров (например, совершенствование технологии измерений), по-видимому, почти не влияет на итоговое значение энтропии.

Здесь нужно пояснить один момент, который многих запутывает. Нестрого говоря, можно утверждать, что есть состояния с низкой энтропией — «менее случайные» и, соответственно, «более организованные», и, согласно второму закону, степень организованности системы постоянно снижается. Однако если рассмотреть ситуацию с другой стороны, то организация высокоэнтропийного состояния системы получается такой же, как и у исходного низкоэнтропийного. Дело в том, что при использовании детерминистичных уравнений динамики организация никогда не снижается, поскольку в конечном высокоэнтропийном состоянии имеется масса точнейших корреляций в движениях частиц, и природа их такова, что если бы мы смогли в точности обратить вспять каждое такое

движение, то вся система «откатилась бы назад», пока не достигла бы исходного «организованного» состояния с низкой энтропией. Это просто свойство динамического детерминизма, и оно всего лишь означает, что «организация» как таковая ничуть не помогает нам понять энтропию и второй закон. Ключевой момент заключается в том, что низкая энтропия соответствует *явному*, или *макроскопическому*, порядку и что неявные корреляции между положениями или движениями субмикроскопических элементов (частиц или атомов) — *не те* вещи, которые повышают энтропию системы. На самом деле это и есть центральная проблема при определении энтропии, и если бы мы попытались обойтись в вышеприведенных определениях энтропии без терминов *явный* или *макроскопический*, то ничуть не продвинулись бы к пониманию энтропии и физического смысла второго закона.

Что же в таком случае представляет собой эта «мера энтропии»? Грубо говоря, мы подсчитываем все возможные различные субмикроскопические состояния, которые могли бы образовать конкретное макроскопическое состояние, и их число N является мерой энтропии макроскопического состояния. Чем больше оказывается N, тем больше энтропия. Однако нерационально считать величину, пропорциональную N, мерой энтропии, потому что нам требуется величина, которая бы могла суммироваться, если бы мы рассматривали вместе две взаимно независимые системы. Следовательно, если Σ_1 и Σ_2 — две такие независимые системы, то нам требуется, чтобы общая энтропия S_{12} двух этих систем была равна сумме $S_1 + S_2$ соответствующих значений энтропии S_1 и S_2:

$$S_{12} = S_1 + S_2.$$

Однако число субмикроскопических состояний N_{12}, составляющих Σ_1 и Σ_2, вместе должно составлять произведение N_1N_2 числа N_1, обусловливающего энтропию Σ_1, и числа N_2, обусловливающего энтропию Σ_2 (поскольку любому варианту N_1, дающему энтропию Σ_1, может сопутствовать любой вариант N_2, дающий энтропию Σ_2). Чтобы преобразовать произведение N_1N_2 в сумму $S_1 + S_2$, нам всего лишь потребуется воспользоваться логарифмом при определении энтропии (см. раздел А.1):

$$S = k \log N,$$

где мы выберем какую-нибудь удобную константу k.

В сущности, это и есть знаменитое определение энтропии, которое дал в 1872 году великий австрийский физик Людвиг Больцман, но в нем есть еще один аспект, который нуждается в пояснении. В классической физике число N обычно получается *бесконечным*! Учитывая такую проблему, приходится представить такой

«подсчет» совершенно иным (более непрерывным) образом. Чтобы лаконично выразить эту процедуру, лучше всего вновь обратиться к формализму *фазового пространства*, который в общих чертах был представлен в разделе 2.11 (а подробнее объяснен в разделе А.6). Как вы помните, фазовое пространство $\mathcal{P}$ некоторой физической системы концептуально представляет собой пространство, обычно содержащее огромное число измерений, и каждая точка такого пространства соответствует полному описанию *субмикроскопического состояния* (допустим, классической) физической системы, которую мы рассматриваем. Это состояние охватывает все *движения* (представленные импульсами), а также все *положения* всех частиц, образующих данную систему. С течением времени точка **P** в пространстве $\mathcal{P}$, соответствующая субмикроскопическому состоянию системы, будет описывать в пространстве $\mathcal{P}$ *кривую* $\mathcal{C}$, положение которой в $\mathcal{P}$ будет определяться уравнениями динамики, как только в $\mathcal{P}$ будет выбрана конкретная (исходная) точка $\mathbf{P}_0$ на кривой $\mathcal{C}$. Любая такая точка $\mathbf{P}_0$ будет фиксировать, какая именно кривая $\mathcal{C}$ отражает эволюцию конкретной системы во времени (см. рис. А.22 из раздела А.7), описываемую **P** (причем $\mathbf{P}_0$ соответствует *исходному* субмикроскопическому состоянию системы). Такова природа детерминизма, играющего центральную роль в классической физике.

Теперь, чтобы определить энтропию, нам потребуется собрать вместе — в единую область, именуемую *регион крупнозернистого разбиения*, — все те точки в $\mathcal{P}$, в которых мы полагаем значения всех макроскопических параметров одинаковыми. Таким образом, все $\mathcal{P}$ будет разделено на такие крупнозернистые области (рис. 3.12). (Можем считать, что границы между такими областями довольно «зыбкие», поскольку всякий раз будет немного сложно их провести.) Обычно считается, что точек из $\mathcal{P}$, лежащих поблизости от таких границ, пренебрежимо мало по сравнению с общим числом точек, что позволяет их игнорировать (см. раздел 1.4 в работе [Penrose, 2010], особенно рис. 1.12). Следовательно, фазовое пространство $\mathcal{P}$ будет делиться на такие регионы, и можно сказать, что *объем* V подобного региона позволяет оценить, сколько существует вариантов заполнения данного макроскопического состояния различными субмикроскопическими состояниями в границах данного региона крупнозернистого разбиения.

К счастью, существует естественная $2n$-мерная мера объема для фазового пространства $\mathcal{P}$, определяемая в рамках классической механики (см. раздел А.6) для системы с n степенями свободы. Каждой пространственной координате x соответствует импульсная координата p, а симплектическая структура $\mathcal{P}$ дает нам меру площади для каждой пары координат, как показано на рис. А.21. Собрав вместе все координаты, получаем $2n$-мерную меру Лиувилля, упоминаемую в разделе А.6. В квантовой системе такой $2n$-мерный объем является числом, кратным $\hbar^n$ (см. раздел 2.2 и 2.11). Если число степеней свободы в рассматри-

ваемой системе очень велико, то это будет очень многомерный объем. Однако естественная квантово-механическая мера объема позволяет без труда сравнивать объемы различных фазовых пространств (см. раздел 2.11). Теперь мы можем дать замечательное больцмановское определение энтропии S макроскопического состояния, которое имеет следующий вид:

$$S = k \log V,$$

где V — объем определенной в $\mathcal{P}$ области с крупнозернистым разбиением, причем состояние определяется в зависимости от значений макроскопических параметров. Число k — фундаментальная константа, имеющая значение $1{,}38 \cdot 10^{-23}$ Дж/К (джоуль на кельвин). Она называется «постоянная Больцмана», и мы уже сталкивались с ней в разделах 2.2 и 2.11.

Для того чтобы оценить, как это помогает понять второй закон, важно прочувствовать, сколь колоссально могут различаться по размеру различные области крупнозернистого разбиения, по меньшей мере в ситуациях, которые встречаются на практике. Логарифм в формуле Больцмана вкупе с чрезвычайной малостью k в обыденных масштабах несколько маскирует всю беспредельность таких различий объема (ср. с разделом А.1). Поэтому легко упустить тот факт, что крошечная разница в энтропии по факту соответствует просто огромной разнице объемов пространств с крупнозернистым разбиением. Допустим, точка **P** движется по кривой $\mathcal{C}$ в фазовом пространстве $\mathcal{P}$, где **P** соответствует (субмикроскопическому) состоянию некоторой рассматриваемой нами системы, а $\mathcal{C}$ описывает ее эволюцию во времени согласно уравнениям динамики. Предположим, **P** переходит из одной области с крупнозернистым разбиением $\mathcal{V}_1$ в соседнюю область $\mathcal{V}_2$ и объемы этих областей равны V_1 и V_2 соответственно (см. рис. 3.12). Как отмечалось ранее, даже при крохотных различиях значений энтропии, соответствующих $\mathcal{V}_1$ и $\mathcal{V}_2$, велика вероятность того, что один из объемов V_1 и V_2 будет несравнимо больше другого. Если $\mathcal{V}_1$ окажется более крупной областью, то в ней найдется лишь ничтожно малое число точек, кривые $\mathcal{C}$ которых ведут из $\mathcal{V}_1$ непосредственно в $\mathcal{V}_2$ (на рис. 3.12 она показана как $\mathcal{V}'_2$). Более того, хотя кривая, описывающая эволюцию (субмикроскопического) состояния во времени прокладывается по $\mathcal{P}$ в соответствии с детерминистичными уравнениями классической динамики, эти уравнения практически неприменимы к крупнозернистому разбиению, поэтому мы не сильно погрешим против истины, если будем считать эту эволюцию совершенно случайной (в контексте областей с крупнозернистым разбиением). Соответственно, если $\mathcal{V}_1$ действительно будет несравнимо больше $\mathcal{V}_2$, мы можем считать крайне маловероятным, что в перспективе эволюция $\mathcal{V}_1$ вновь пойдет в направлении $\mathcal{V}_2$. С другой стороны, будь $\mathcal{V}_2$ гораздо больше $\mathcal{V}_1$ (этот случай показан на рис. 3.12), то исключительно велика вероятность того,

что кривая $\mathcal{C}$, начало которой лежит в $\mathcal{V}_1$, может далее последовать в $\mathcal{V}_2$. Затем, затерявшись в $\mathcal{V}_2$, она скорее проникнет в еще более обширный объем $\mathcal{V}_3$, нежели вернется в столь ничтожно малую область, как $\mathcal{V}_1$. Поскольку несравнимо более крупному объему соответствует обычно лишь чуть более высокая энтропия, мы уже примерно представляем себе, почему энтропия будет неудержимо возрастать с течением времени. Именно этого и следует ожидать согласно второму закону.

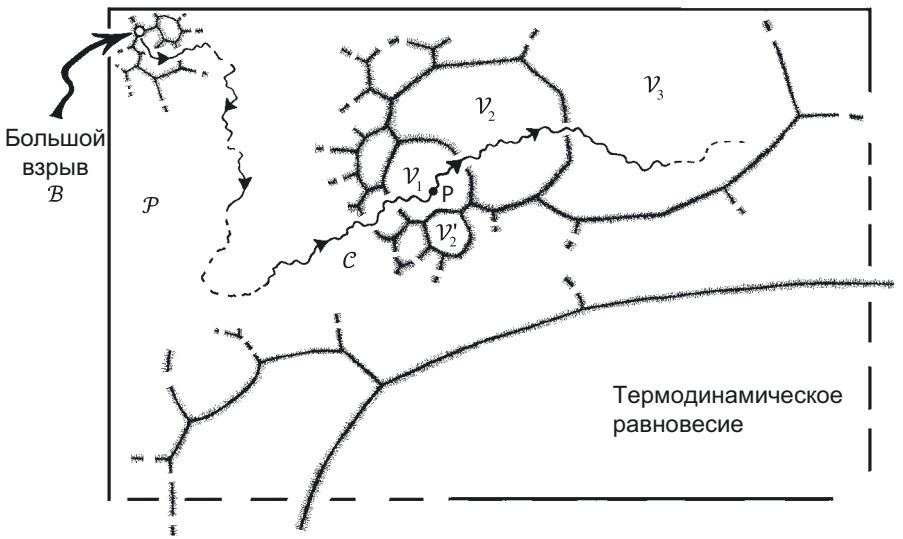


Рис. 3.12. Показанное здесь фазовое пространство $\mathcal{P}$ — многомерное многообразие, точки которого соответствуют целостному (классическому) состоянию системы (всем значениям координат и импульсов; см. рис. А.20), — разделено на области с крупнозернистым разбиением (границы между которыми являются нечеткими), и в каждой из этих областей сгруппированы все состояния с одинаковыми макроскопическими параметрами (до определенной степени точности). Энтропия по Больцману, присущая **P** в регионе $\mathcal{V}$ с объемом V, равна $k \log V$. Второй закон термодинамики понимается как тенденция к неуклонному возрастанию объемов, через которые пролегает кривая $\mathcal{C}$, отражающая эволюцию **P** во времени (см. рис. А.22). Рамки данной иллюстрации позволяют лишь условно показать возрастание этих объемов. В конечном итоге второй закон работает, потому что $\mathcal{C}$ начинается в крошечной области $\mathcal{B}$, где произошел Большой взрыв

Однако эти объяснения позволяют разобраться в ситуации лишь наполовину, причем речь идет о «простой» половине. Они помогают понять, почему, если эволюция системы начинается из состояния со сравнительно низкой энтропией, абсолютное большинство субмикроскопических состояний, образующих конкретное макроскопическое состояние, будут эволюционировать в сторону все более высокой энтропии (возможно, с небольшими флуктуациями в противо-

положном направлении). Второй закон термодинамики гласит, что энтропия возрастает, и приведенные выше приблизительные выкладки отчасти объясняют, почему все именно так. Однако, поразмыслив, мы можем счесть этот вывод в чем-то парадоксальным: ведь мы приходим к нему, изучая системы, которые подчиняются законам динамики, полностью симметричным во времени. Нет, на самом деле *не можем*. Асимметрия во времени возникает лишь потому, что сам вопрос, который мы задаем в отношении системы, по сути асимметричен, а именно: мы хотим выяснить вероятное поведение системы *в будущем*, *исходя* из ее актуального макроскопического состояния. В таком случае мы приходим к выводу, который согласуется с асимметричным во времени вторым законом термодинамики.

Однако давайте посмотрим, что произойдет, если мы переформулируем вопрос, обратив время вспять. Допустим, имеется макроскопическое состояние со сравнительно низкой энтропией — у нас высоко над ковром находится полная чашка воды, но при этом она немного неустойчива. Теперь давайте поинтересуемся не тем, что, скорее всего, произойдет с водой в будущем, а тем, в результате каких *прошедших* событий могла возникнуть такая ситуация. Предположим, у нас в фазовом пространстве $\mathcal{P}$, как и раньше, есть две смежные области $\mathcal{V}_1$ и $\mathcal{V}_2$ с крупнозернистым разбиением. Но теперь будем исходить из того, что субмикроскопическое состояние, представленное точкой **P**, на рис. 3.12 локализовано в области $\mathcal{V}_2$. Если $\mathcal{V}_2$ несравнимо больше $\mathcal{V}_1$, то лишь малая толика всех точек в $\mathcal{V}_2$ будет соответствовать положениям **P**, к которым приводят кривые $\mathcal{C}$, начинающиеся в $\mathcal{V}_1$, тогда как, если $\mathcal{V}_1$ будет гораздо крупнее, чем $\mathcal{V}_2$, то в $\mathcal{V}_2$ найдется гораздо больше точек, к которым приводят кривые $\mathcal{C}$, исходящие из $\mathcal{V}_1$. Следовательно, повторив те же рассуждения, которые успешно описывали ход времени из прошлого в будущее, мы, вероятно, придем к выводу о том, что наша точка с гораздо большей вероятностью проникнет в $\mathcal{V}_2$ из значительно более обширного объема, чем из более мелкого, – то есть из состояния с более высокой энтропией, а не с более низкой. Распространяя эти рассуждения все дальше и дальше в прошлое, мы приходим к выводу о том, что абсолютное большинство путей к точкам области $\mathcal{V}_2$ лежат на таких кривых $\mathcal{C}$, которые соответствуют более высокой энтропии, а далее — все более и более высокой (пожалуй, с минимальными противоположными флуктуациями), по мере того как мы углубляемся в прошлое.

Разумеется, это прямо противоречит второму закону, поскольку мы, кажется, только что пришли к такому выводу: если бы мы отправились в прошлое, то энтропия вокруг неуклонно стала бы возрастать. Иными словами, анализируя любую ситуацию с достаточно низкой энтропией, мы бы с огромной вероятностью обнаружили, что ранее в природе действовал закон, *обратный* второму закону

термодинамики! Естественно, это нонсенс: повседневный опыт подсказывает, что настоящее время лишено каких-либо специфических черт, касающихся второго закона, и этот закон действовал во Вселенной в прошлом, а также продолжит действовать в будущем. Более того, любое прямое наблюдение приносит нам информацию только о событиях, которые уже произошли, и мы никогда не получали никаких наблюдательных свидетельств из будущего — и это заставляет нас верить во второй закон. Получается, что наши наблюдения прямо противоречат тому теоретическому выводу, который мы только что сделали!

Возьмем наш пример с чашкой воды. Вопрос: каким наиболее вероятным способом вода могла попасть в чашку, которая установлена высоко над ковром и при этом достаточно неустойчива? Только что изложенные теоретические выкладки (картина, которая, «вероятнее всего, могла бы предшествовать» такой ситуации) помогают представить последовательность событий, в ходе которых энтропия *снижалась* с течением времени (то есть возрастала бы при движении в прошлое). Имеем: на ковре лужица воды, и вдруг капли начинают собираться, выстраиваться в единое целое и подниматься вверх, в направлении чашки, пока вся вода одновременно не окажется в сосуде. Разумеется, ничего подобного на практике не происходит. На самом деле случилась бы последовательность событий, в ходе которой энтропия возрастает, то есть события подчиняются второму закону: например, кто-то держит кувшин и наливает в чашку воду из него, либо (если мы попробуем описать ситуацию без человеческого вмешательства) какой-нибудь робот открывает и закрывает кран.

Итак, в чем же ошибочны наши рассуждения? Они не ошибочны, если рассматривать наиболее вероятную последовательность событий, которая могла бы привести к искомому макроскопическому состоянию в результате совершенно случайных флуктуаций. Однако, насколько нам известно, на самом деле мир устроен иначе. Согласно второму закону, следует ожидать, что в отдаленном будущем (на макроскопическом уровне) будет царить страшный беспорядок, и здесь нет никаких ограничений, которые могли бы развенчать наши аргументы о вероятной последовательности событий в будущем. Но если второй закон и в самом деле действовал всегда, с момента возникновения Вселенной, то отдаленное прошлое должно было выглядеть совсем не так, как настоящее. В нем должны были действовать серьезные ограничения, обеспечивающие исключительно высокую макроскопическую упорядоченность. Если учесть в наших оценках вероятности хотя бы только это единственное ограничение, касающееся исходного состояния Вселенной, а именно оно должно было обладать *исчезающе малой энтропией*, то придется отвергнуть все вышеприведенные рассуждения о том, как мог быть устроен мир в прошлом, поскольку они не удовлетворяют этому ограничению. И тогда можно согласиться с картиной, согласно которой второй закон действовал всегда.

Следовательно, ключевое свойство второго закона заключается в том, что на заре существования Вселенная обладала исключительно высокой макроскопической организацией. Но что это было за состояние? Как мы видели в разделе 3.1, современная теория, подкрепляемая убедительными данными наблюдений, которые будут рассмотрены чуть позже, в разделе 3.4, свидетельствует о том, что это состояние представляло собой всеобъемлющий *Большой взрыв*! Как такой невообразимо бурный взрыв мог быть невероятно упорядоченным макроскопическим состоянием с крайне низкой энтропией? Давайте поговорим об этом в следующем разделе, где нас, кажется, ожидает экстраординарный парадокс, связанный, по-видимому, с этим уникальным событием.

3.4. Парадокс Большого взрыва

Сначала поставим вопрос о наблюдениях. Какие есть прямые доказательства в пользу того, что когда-то вся наблюдаемая Вселенная находилась в чрезвычайно сжатом и невероятно горячем состоянии, чтобы это согласовывалось с картиной Большого взрыва, представленной в разделе 3.1? Наиболее убедительным свидетельством является *реликтовое излучение* (РИ), иногда именуемое *отблеском большого взрыва*. Реликтовое излучение представляет собой свет, но с очень большой длиной волны, так что увидеть его глазами совершенно невозможно. Этот свет льется на нас со всех сторон исключительно равномерно (но в основном некогерентно). Он представляет собой тепловое излучение с температурой ~2,725 K, то есть на два с лишним градуса выше абсолютного нуля. Считается, что наблюдаемый «отблеск» зародился в невероятно горячей Вселенной (~3000 K на тот момент) примерно через 379 000 лет после Большого взрыва — в эпоху *последнего рассеяния*, когда Вселенная впервые стала прозрачной для электромагнитного излучения (хотя это произошло совсем не *при* Большом взрыве; данное событие приходится на первую 1/40 000 общего возраста Вселенной — от Большого взрыва до наших дней). С эпохи последнего рассеяния длина этих световых волн увеличилась примерно настолько, насколько расширилась и сама Вселенная (приблизительно в 1100 раз), так что плотность энергии столь же радикально уменьшилась. Поэтому наблюдаемая температура РИ составляет всего 2,725 K.

Тот факт, что это излучение существенно некогерентное (то есть тепловое), впечатляюще подтверждается самой природой его частотного спектра, приведенного на рис. 3.13. По вертикали на графике откладывается интенсивность излучения на каждой конкретной частоте, а частота возрастает слева направо. Непрерывная кривая соответствует *планковскому спектру абсолютно черного тела*, о котором шла речь в разделе 2.2, для температуры 2,725 K. Точки на кривой — это данные конкретных наблюдений, для которых указаны планки погрешностей. При этом

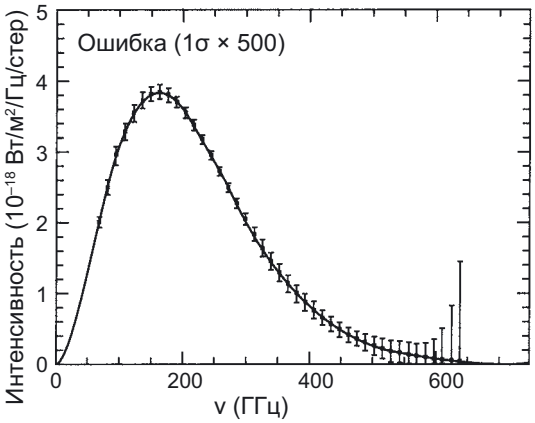


Рис. 3.13. Спектр РИ, зафиксированный телескопом COBE, почти полностью совпадает с *тепловым* (планковским) спектром (сплошная линия): планки погрешностей, соответствующие наблюдениям РИ, увеличены в 500 раз

планки погрешностей увеличены в 500 раз, поскольку иначе их было бы просто невозможно рассмотреть, даже справа, где ошибки достигают максимума. Совпадение между теоретической кривой и результатами наблюдений просто замечательное — пожалуй, это самое лучшее совпадение с тепловым спектром, найденное в природе[1].

Однако о чем свидетельствует это совпадение? О том, что мы рассматриваем состояние, которое, по-видимому, находилось очень близко к термодинамическому равновесию (поэтому ранее и использовался термин *некогерентный*). Но какой вывод следует из того, что новоиспеченная Вселенная была очень близка к термодинамическому равновесию? Вернемся к рис. 3.12 из раздела 3.3. Самая обширная область с крупнозернистым разбиением (по определению) будет гораздо больше, чем любая другая такая область, и, как правило, она настолько велика по сравнению с остальными, что значительно превзойдет по объему их все! Термодинамическое равновесие соответствует макроскопическому состоянию, в которое, надо полагать, рано или поздно придет любая система. Иногда оно называется *тепловой смертью Вселенной*, но в этом случае, как ни странно, речь должна идти о *тепловом рождении Вселенной*. Ситуация осложняется тем,

[1] Зачастую утверждают, что наблюдаемый спектр РИ наилучшим способом согласуется с теоретическим планковским спектром. Однако это не совсем так, поскольку COBE всего лишь сравнивает спектр РИ со спектром искусственно созданного теплового излучения, поэтому в действительности мы получаем лишь совпадение наблюдаемого теплового спектра РИ и наблюдаемого теплового спектра искусственного источника.

что новорожденная Вселенная стремительно расширялась, поэтому то состояние, которое мы рассматриваем, на самом деле неравновесное. Тем не менее расширение в данном случае может считаться по сути *адиабатическим* — данный момент в полной мере оценил Толман еще в 1934 году [Tolman, 1934]. Это означает, что величина энтропии при расширении не изменялась. (Ситуацию, подобную данной, когда благодаря адиабатическому расширению сохраняется термодинамическое равновесие, можно описать в фазовом пространстве как совокупность равных по объему областей с крупнозернистым разбиением, которые отличаются друг от друга лишь конкретными объемами Вселенной. Можно считать, что для этого первичного состояния была характерна максимальная энтропия — несмотря на расширение!)

По-видимому, мы столкнулись с исключительным парадоксом. Согласно аргументам, изложенным в разделе 3.3, Второй закон требует (и, в принципе, тем самым и объясняется), чтобы Большой взрыв был макроскопическим состоянием с крайне *низкой* энтропией. Однако наблюдения РИ, по-видимому, свидетельствуют о том, что макроскопическое состояние Большого взрыва отличалось колоссальной энтропией, пожалуй, даже максимально возможной. Где же мы так серьезно ошибаемся?

Вот один из распространенных вариантов объяснения этого парадокса: предполагается, что, поскольку новорожденная Вселенная была очень «маленькой», там мог существовать некий предел максимальной энтропии, и состояние термодинамического равновесия, которое, по-видимому, поддерживалось в то время, было попросту предельным уровнем энтропии, возможным на тот момент. Однако это *неверный* ответ. Такая картина могла бы соответствовать совершенно иной ситуации, в которой размеры Вселенной зависели бы от некоего внешнего ограничения, например, как в случае с газом, который заключен в цилиндре с герметичным поршнем. В таком случае давление поршня обеспечивается неким внешним механизмом, который оснащен внешним источником (или отводом) энергии. Но эта ситуация неприменима ко Вселенной в целом, геометрия и энергия которой, а также «габаритный размер» определяются исключительно внутренним устройством и управляются динамическими уравнениями эйнштейновской общей теории относительности (включая уравнения, описывающие состояние материи; см. разделы 3.1 и 3.2). В таких условиях (когда уравнения полностью детерминистичны и инвариантны по отношению к направлению времени — см. раздел 3.3) с течением времени не может меняться общий объем фазового пространства. При этом предполагается, что *само по себе* фазовое пространство $\mathcal{P}$ не должно «развиваться»! Вся эволюция попросту описывается расположением кривой $\mathcal{C}$ в пространстве $\mathcal{P}$ и в данном случае представляет полную эволюцию Вселенной (см. раздел 3.3).

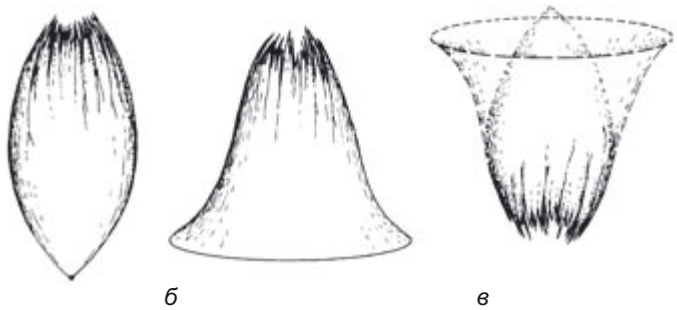


Рис. 3.14. Фридмановская модель с типичными возмущениями при $K > 0$, $\Lambda = 0$ (в отличие от рис. 3.6 *а*), развивающаяся в соответствии со Вторым законом, должна схлопнуться в результате слияния множества черных дыр, и при этом получится исключительно беспорядочная сингулярность, совсем не похожая на модель ФЛРУ (*а*); аналогичное поведение, ожидаемое в любой коллапсирующей модели с типичными возмущениями (*б*); обращение времени, как и предполагается, приводит к типичному Большому взрыву (*в*)

Пожалуй, проблема прояснится, если рассмотреть поздние стадии *коллапса* Вселенной, когда она приближается к Большому краху. Вспомните модель Фридмана при $K > 0$, $\Lambda = 0$, представленную на рис. 3.2 *а* в разделе 3.1. Теперь мы считаем, что возмущения в этой модели возникают из-за нерегулярного распределения материи, и в некоторых частях уже произошли локальные коллапсы, на месте которых остались черные дыры. Тогда следует предположить, что после этого какие-то черные дыры будут сливаться друг с другом и что схлопывание в конечную сингулярность окажется исключительно сложным процессом, не имеющим почти ничего общего со строго симметричным Большим крахом идеально шарообразной симметричной фридмановской модели, представленной на рис. 3.6 *а*. Напротив, в качественном отношении ситуация коллапса будет гораздо больше напоминать ту грандиозную мешанину, которая показана на рис. 3.14 *а*; возникающая в данном случае итоговая сингулярность может в какой-то степени согласовываться с гипотезой БКЛМ, упомянутой в конце раздела 3.2. Конечное схлопывающееся состояние будет обладать невообразимой энтропией, несмотря на то что Вселенная вновь сожмется до крохотных размеров. Хотя именно такая (пространственно-замкнутая) реколлапсирующая фридмановская модель сейчас *не* считается правдоподобным вариантом представления нашей собственной Вселенной, те же соображения актуальны и для других фридмановских моделей, с космологической постоянной или без нее. Коллапсирующая разновидность любой такой модели, испытывающая схожие возмущения из-за неравномерного распределения материи, опять же, должна превратиться во всепоглощающий хаос, сингулярность вроде черной дыры (рис. 3.14 *б*). Обратив время вспять

в каждом из этих состояний, мы дойдем до возможной исходной сингулярности (потенциального Большого взрыва), имеющей, соответственно, *колоссальную* энтропию, что противоречит высказанному здесь предположению о «потолке» энтропии (рис. 3.14 *в*).

Здесь я должен перейти к альтернативным возможностям, которые также порой рассматриваются. Некоторые теоретики предполагают, что второй закон должен каким-то образом *обращаться вспять* в таких коллапсирующих моделях, так что общая энтропия Вселенной будет становиться все меньше (после максимального расширения) по мере приближения Большого краха. Однако такую картину особенно сложно представить себе при наличии черных дыр, которые, стоит им образоваться, сами станут работать на *повышение* энтропии (что связано с асимметрией времени в расположении нулевых конусов у горизонта событий, см. рис. 3.9). Это будет продолжаться и в отдаленном будущем — как минимум до тех пор, пока черные дыры не испарятся под действием хокинговского механизма (см. разделы 3.7 и 4.3). В всяком случае, подобная возможность не отменяет представленных здесь аргументов. Существует еще и другая важная проблема, которая связана с такими сложными коллапсирующими моделями и о которой, возможно, задумывались и сами читатели: сингулярности черных дыр вполне могут возникать совсем не одновременно, поэтому при обращении времени мы не получим Большой взрыв, который происходит «весь и сразу». Однако именно таково одно из свойств (пока не доказанной, но убедительной) *гипотезы сильной космической цензуры* [Penrose, 1998a; ПкР, раздел 28.8], согласно которой в общем случае такая сингулярность будет *пространственноподобной* (раздел 1.7), а потому может считаться единовременным событием. Более того, безотносительно к вопросу о справедливости самой гипотезы сильной космической цензуры известно множество решений, удовлетворяющих этому условию, и все подобные варианты (при расширении) будут обладать относительно высокими значениями энтропии. Это существенно снижает степень беспокойства относительно справедливости наших выводов.

Соответственно, мы не находим доказательств того, что при малых пространственных размерах Вселенной в ней бы обязательно существовал некий «низкий потолок» возможной энтропии. В принципе, скопление материи в виде черных дыр и слияние «чернодырных» сингулярностей в единый сингулярный хаос — это процесс, который отлично согласуется со вторым законом, и этот финальный процесс должен сопровождаться колоссальным возрастанием энтропии. «Крошечное» по геометрическим меркам окончательное состояние Вселенной может обладать невообразимой энтропией, гораздо более высокой, чем на сравнительно ранних этапах такой коллапсирующей космологической модели, и пространственная миниатюрность сама по себе не устанавливает «потолок»

для максимального значения энтропии, хотя такой «потолок» (при обращении хода времени) как раз мог бы объяснить, почему при Большом взрыве энтропия была чрезвычайно мала. На самом деле такая картина (рис. 3.14 *а*, *б*), на которой в общем виде представлен коллапс Вселенной, подсказывает разгадку парадокса: почему при Большом взрыве была исключительно низкая энтропия по сравнению с той, что могла быть, несмотря на то что взрыв был *горячим* (а такое состояние должно обладать максимальной энтропией). Ответ заключается в том, что энтропия может радикально увеличиваться, если допустить серьезные отклонения от пространственной однородности, и величайший прирост такого рода связан с неравномерностями, обусловленными как раз возникновением черных дыр. Следовательно, пространственно-однородный Большой взрыв действительно мог обладать, условно говоря, неимоверно низкой энтропией, несмотря на то что его содержимое было невероятно горячим.

Одно из наиболее убедительных доказательств в пользу того, что Большой взрыв действительно был довольно однородным с пространственной точки зрения, хорошо согласующееся с геометрией модели ФЛРУ (но не согласующееся с гораздо более общим случаем беспорядочной сингулярности, проиллюстрированным на рис. 3.14 *в*), вновь связано с РИ, но на этот раз с его угловой однородностью, а не с термодинамической природой. Такая однородность проявляется в том, что температура РИ практически одна и та же в любой точке неба, и отклонения от однородности составляют не более 10^{-5} (с поправкой на небольшой доплеровский эффект, связанный с нашим движением сквозь окружающую материю). Кроме того, наблюдается практически всеобщая однородность в распределении галактик и другой материи; так, распределение барионов (см. раздел 1.3) в достаточно больших масштабах характеризуется значительной однородностью, хотя и есть заметные аномалии, в частности так называемые войды, где плотность видимой материи кардинально ниже среднего. В целом можно утверждать, что однородность оказывается тем выше, чем дальше в прошлое Вселенной мы заглядываем, а РИ — это древнейшее свидетельство распределения материи, которое мы можем непосредственно наблюдать.

Эта картина согласуется с точкой зрения, согласно которой на ранних этапах развития Вселенная действительно была исключительно однородной, но со слегка нерегулярной плотностью. С течением времени (и под влиянием различного рода «трения» — процессов, замедляющих относительные движения) эти плотностные неравномерности усиливались под действием гравитации, что согласуется с представлением о постепенном комковании вещества. Со временем комкование нарастает, в результате образуются звезды; они группируются в галактики, в центре каждой из которых образуется массивная черная дыра. В конечном итоге такое комкование обусловлено неотвратимым действием

гравитации. Такие процессы действительно сопряжены с сильнейшим нарастанием энтропии и демонстрируют, что с учетом гравитации тот первозданный сияющий шар, от которого сегодня осталось лишь РИ, мог обладать далеко не максимальной энтропией. Термическая природа этого шара, о которой свидетельствует планковский спектр, показанный на рис. 3.13, говорит лишь вот о чем: если рассмотреть Вселенную (в эпоху последнего рассеяния) просто как систему, состоящую из вещества и энергии, взаимодействующих друг с другом, то можно считать, что она фактически находилась в термодинамическом равновесии. Однако если при этом учесть и гравитационные воздействия, то картина драматическим образом меняется.

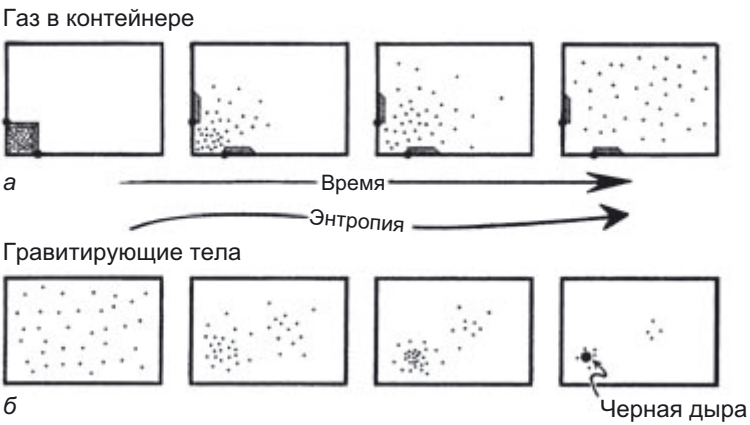


Рис. 3.15. Для молекул газа в контейнере пространственная однородность достигается при максимальной энтропии (*а*); для звезд, связанных гравитационными взаимодействиями в «контейнере» галактического масштаба, высокая энтропия достигается при комковании, а максимальная — при образовании черной дыры (*б*)

Если представить себе, например, газ в герметичном контейнере, то естественно считать, что максимальной энтропии он достигнет в том макроскопическом состоянии, когда равномерно распределится по всему контейнеру (рис. 3.15 *а*). В данном отношении он будет напоминать раскаленный шар, породивший РИ, которое равномерно распределено по небу. Однако если заменить молекулы газа обширной системой тел, связанных друг с другом гравитацией, например отдельных звезд, то получится совершенно иная картина (рис. 3.15 *б*). Из-за гравитационных эффектов звезды распределятся неравномерно, в виде скоплений. В конечном итоге наибольшая энтропия будет достигнута, когда многочисленные звезды сколлапсируют или сольются в черные дыры. Несмотря на то что на этот процесс и может потребоваться немало времени (хотя ему и будет

способствовать трение, обусловленное присутствием межзвездного газа), мы увидим, что в конечном итоге при доминировании гравитации энтропия тем выше, чем *менее* равномерно распределено вещество в системе.

Такие эффекты прослеживаются даже на уровне повседневного опыта. Возможен вопрос: а какова роль Второго закона в поддержании жизни на Земле? Часто можно услышать, что мы живем на этой планете благодаря энергии, получаемой от Солнца. Но это не вполне верное утверждение, если рассматривать Землю в целом, так как практически вся энергия, получаемая Землей днем, вскоре вновь улетучивается в космос, в темное ночное небо. (Разумеется, точный баланс будет немного корректироваться под влиянием таких факторов, как глобальное потепление и разогрев планеты под действием радиоактивного распада.) В противном случае Земля попросту раскалялась бы все сильнее и за несколько дней стала бы необитаемой! Однако фотоны, получаемые непосредственно от Солнца, обладают относительно высокой частотой (они сосредоточены в желтой части спектра), а Земля отдает в космос гораздо более низкочастотные фотоны, относящиеся к инфракрасному спектру. По формуле Планка ($E = h\nu$, см. раздел 2.2) каждый из поступающих от Солнца фотонов *в отдельности* обладает гораздо более высокой энергией, чем фотоны, излучаемые в космос, поэтому для достижения баланса с Земли должно уходить гораздо больше фотонов, чем приходит (см. рис. 3.16). Если поступает меньше фотонов, то у входящей

Рис. 3.16. Жизнь на Земле существует благодаря огромному температурному дисбалансу в небе. Мы получаем от Солнца энергию в виде относительно немногочисленных высокочастотных (желтых) фотонов с низкой энтропией, а зеленые растения преобразуют ее в гораздо более многочисленные низкочастотные (инфракрасные) фотоны, которые уносят с Земли такое же количество энергии в высокоэнтропийном виде. Таким образом растения, а значит и вся остальная жизнь на Земле, растут и сохраняют свою структуру

энергии будет меньше степеней свободы, а у исходящей — больше, и, следовательно, по формуле Больцмана ($S = k \log V$) входящие фотоны будут обладать гораздо меньшей энтропией, чем исходящие. Мы пользуемся низкоэнтропийной энергией, заключенной в растениях, чтобы понижать собственную энтропию: едим растения либо травоядных животных. Так жизнь на Земле сохраняется и процветает. (По-видимому, эти мысли впервые четко сформулировал Эрвин Шрёдингер в 1967 году, написавший свою революционную книгу «Жизнь как она есть» [Schrödinger, 2012]).

Важнейший факт, связанный с этим низкоэнтропийным балансом, заключается в следующем: Солнце — это горячее пятно в совершенно темном небе. Но как сложились такие условия? Сыграли роль многие сложные процессы, в том числе связанные с термоядерными реакциями и т. д., но самое важное то, что Солнце вообще существует. А оно возникло потому, что солнечная материя (как и материя, образующая другие звезды) развивалась в процессе гравитационного комкования, причем все начиналось с относительно однородного распределения газа и темной материи.

Здесь требуется упомянуть загадочную субстанцию под названием *темная материя*, которая, по-видимому, составляет 85 % материального (не-Λ) содержимого Вселенной, но она обнаруживается только по гравитационному взаимодействию, а состав ее неизвестен. Сегодня мы всего лишь учитываем эту материю при оценке общей массы, которая нужна при расчете некоторых числовых величин (см. разделы 3.6, 3.7, 3.9, а о том, какую более важную теоретическую роль может играть темная материя, см. раздел 4.3). Безотносительно к проблеме темной материи мы видим, насколько важной для нашей жизни оказалась низкоэнтропийная природа первоначального однородного распределения материи. Наше существование, насколько мы его понимаем, зависит от низкоэнтропийного гравитационного запаса, который характерен для исходного однородного распределения материи.

Здесь мы подходим к примечательному — на самом деле даже *фантастическому* — аспекту Большого взрыва. Тайна кроется не только в том, как он произошел, но и в том, что это было событие с чрезвычайно низкой энтропией. Более того, примечательно не столько это обстоятельство, сколько сам факт, что энтропия была низкой лишь в *одном* конкретном отношении, а именно: *гравитационные* степени свободы по какой-то причине были *полностью подавлены*. Это резко контрастирует со степенями свободы материи и (электромагнитного) излучения, поскольку они, по-видимому, были максимально возбуждены в горячем состоянии с максимальной энтропией. На мой взгляд, это, пожалуй, глубочайшая космологическая загадка, и по какой-то причине она до сих пор остается недооцененной!

Следует подробнее остановиться на том, сколь особенным было состояние Большого взрыва и какая энтропия может возникнуть в процессе гравитационного комкования. Соответственно, нужно для начала осознать, какая невероятная энтропия на самом деле присуща черной дыре (см. рис. 3.15 *б*). Этот вопрос мы обсудим в разделе 3.6. Но пока давайте обратимся к другой проблеме, связанной со следующей, довольно вероятной возможностью: ведь Вселенная на самом деле может оказаться пространственно-*бесконечной* (как в случае ФЛРУ-моделей с $K \leq 0$, см. раздел 3.1) или как минимум большая часть Вселенной может быть недоступна для непосредственного наблюдения. Соответственно, мы подходим к проблеме *космологических горизонтов*, о которой поговорим в следующем разделе.

3.5. Горизонты, сопутствующие объемы и конформные диаграммы

Прежде чем точнее определить, насколько особенным был Большой взрыв по сравнению со всем множеством возможных вариантов пространственно-временной геометрии и распределения материи, давайте обсудим другую ситуацию: ведь многие модели с бесконечной пространственной геометрией должны обладать *бесконечной общей энтропией*, и вот здесь возникает осложнение. Однако дух вышеприведенной аргументации не особенно изменится, если говорить не об общей энтропии Вселенной, а о таком понятии, как энтропия сопутствующего объема. Идея *сопутствующей области* в ФЛРУ-модели заключается в следующем: мы рассматриваем развивающуюся со временем область пространства, границы которой совпадают с *временными линиями* модели (мировыми линиями идеализированных галактик; см. рис. 3.5 в разделе 3.1). Разумеется, если речь идет о черных дырах, а они, как будет рассказано в следующем разделе, вносят ключевой вклад в проблему энтропии, и возникают серьезные отклонения от точной формы ФЛРУ, может быть не столь ясно, что именно представляет собой сопутствующий объем. Однако если рассматривать ситуацию в достаточно крупном масштабе, то такая неопределенность становится сравнительно несущественной.

В дальнейшем будет целесообразно рассмотреть, что происходит в точных ФЛРУ-моделях на очень больших масштабах. Во всех ФЛРУ-моделях, обсуждаемых в этой главе, существует так называемый *горизонт частицы* — этот термин впервые четко определил Вольфганг Риндлер (Wolfgang Rindler) в 1956 году [Rindler, 1956]. Чтобы дать его общеупотребительное определение, рассмотрим точку P в пространстве-времени и исследуем ее световой конус прошлого $\mathcal{K}$. Значительное множество временных линий (см. раздел 3.1) будут пересекать $\mathcal{K}$,

и можно считать, что сегмент $\mathcal{G}(\mathrm{P})$ пространства, заметаемый этими линиями, содержит семейство *наблюдаемых галактик* пространства P. Но некоторые временные линии могут не пересекать $\mathcal{K}$, так как находятся слишком далеко от P. Так у $\mathcal{G}(\mathrm{P})$ образуется граница $\mathcal{H}(\mathrm{P})$, представляющая собой времениподобную гиперповерхность, свойства которой определяются времениподобными линиями. Данная 3-поверхность $\mathcal{H}(\mathrm{P})$ является *горизонтом частицы* в пространстве P (рис. 3.17).

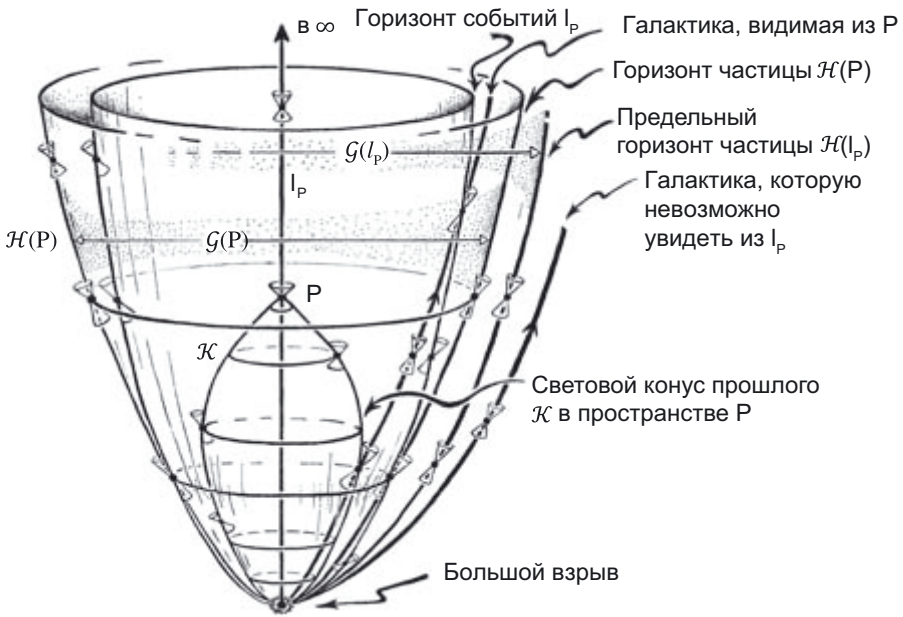


Рис. 3.17. Пространственно-временное представление ФЛРУ-модели: показаны различные типы горизонтов и мировых линий галактик

Для любого конкретного космического времени t срез $\mathcal{G}(\mathrm{P})$ модели Вселенной, сделанный в заданный момент t, будет иметь конечный объем, и максимальное значение энтропии для данной области также будет конечным. Если рассмотреть *всю* временную линию l_P, проходящую через P, то область $\mathcal{G}(\mathrm{P})$ с наблюдаемыми галактиками, скорее всего, будет увеличиваться в будущем вдоль линии l_P. В стандартных ФЛРУ-моделях с $\Lambda > 0$ будет присутствовать ограничивающая «крупнейшая» область $\mathcal{G}(l_\mathrm{P})$, которая окажется пространственно-конечной для любого заранее заданного космического времени t, и можно считать, что в вышеприведенной аргументации требуется рассматривать не более чем ту максимальную энтропию, которая достижима в рамках $\mathcal{G}(l_\mathrm{P})$. В некотором

смысле также полезно рассмотреть и другой тип космологического горизонта (четко определенный Риндлером [Rindler, 1956]), а именно *горизонт событий* бесконечно протяженной в будущее мировой линии типа $l_{\text{р}}$, представляющий собой (расположенную в будущем) границу множества точек, лежащих в прошлом относительно $l_{\text{р}}$. Аналогичный горизонт возникает при коллапсе в черную дыру, однако в данном случае $l_{\text{р}}$ заменяется на мировую линию бесконечно удаленного (и вечного существующего) наблюдателя, не падающего в черную дыру (см. рис. 3.9 в разделе 3.2).

Проблемы, связанные с этими горизонтами, часто выглядят довольно запутанными, как, например, на рис. 3.2 и 3.3 в разделе 3.1 (см. также рис. 3.17, ср. с рис. 3.9 в разделе 3.2). Эти термины можно представить в гораздо более понятном виде, если изобразить их при помощи *конформных диаграмм* [Penrose, 1963, 1964a, 1965b, 1967b; ПкРп.27.12; Carter, 1966]. Такие диаграммы особенно удобны потому, что позволяют представить *бесконечность* как конечную границу пространства-времени. Мы уже сталкивались с этим аспектом конформных представлений на рис. 1.38 *а* и 1.40 в разделе 1.15. Другая особенность таких диаграмм заключается в том, что они проясняют причинные аспекты (то есть горизонты частиц) *сингулярностей* типа Большого взрыва в ФЛРУ-моделях.

В таких изображениях используется конформное преобразование метрического тензора $\mathbf{g}$ (см. разделы 1.1, 1.7 и 1.8) физического пространства-времени $\mathcal{M}$, дающее новую метрику $\hat{\mathbf{g}}$ для конформно связанного пространства-времени $\hat{\mathcal{M}}$, согласно выражению

$$\hat{\mathbf{g}} = \Omega^2 \mathbf{g},$$

где Ω — (обычно положительная) гладко изменяющаяся скалярная величина в пространстве-времени, так что нулевые конусы (и локальное направление времени) не изменяются при замене $\mathbf{g}$ на $\hat{\mathbf{g}}$. В обычных условиях $\hat{\mathcal{M}}$ с гладкой метрикой $\hat{\mathbf{g}}$ действительно дают гладкую границу, где $\Omega = 0$ соответствует *бесконечности* в исходном пространстве-времени $\mathcal{M}$. При $\Omega = 0$ происходит бесконечное «расплющивание» $\mathbf{g}$ в бесконечные области $\mathcal{M}$, чтобы у пространства $\hat{\mathcal{M}}$ получилась конечная граничная область $\mathcal{F}$. Разумеется, такая процедура позволит нам получить гладкую границу $\hat{\mathcal{M}}$ только при подходящих обстоятельствах с понижением метрики (и, возможно, топологии) $\mathcal{M}$, но довольно примечательно, насколько работоспособна эта процедура применительно к пространствам-временам $\mathcal{M}$, особенно интересным с физической точки зрения.

Эту процедуру представления пространственно-временной бесконечности дополняет другая, смежная, согласно которой при подходящих обстоятельствах можно бесконечно «расширять» *сингулярность* в метрике $\mathcal{M}$, чтобы получить область $\mathcal{B}$,

ограничивающую $\mathcal{M}$, которая и представляет эту сингулярность. В космологической модели $\mathcal{M}$ вполне можно получить гладко прилегающую к $\hat{\mathcal{M}}$ граничную область $\mathcal{B}$, которая будет соответствовать Большому взрыву. Возможно, нам еще крупно повезло, что обратный масштабный фактор Ω^{-1} плавно стремится к нулю по мере приближения к $\mathcal{B}$, а именно таким свойством обладают большие взрывы в важнейших рассматриваемых здесь ФЛРУ-космологиях. (Обратите внимание: термин «Большой взрыв» с заглавной буквы означает конкретное сингулярное событие, которое, по-видимому, породило известную нам Вселенную; в то же время, термин «большой взрыв» используется для обозначения исходных сингулярностей в космологических моделях вообще; см. также раздел 4.3.) В толмановских моделях, наполненных излучением, значение Ω^{-1} на границе имеет простой нуль, а во фридмановских пылевых моделях — нуль второго порядка. Если представить и бесконечное будущее, и сингулярное начало $\mathcal{M}$ как гладкие граничные области, прилегающие к конформно связанному $\hat{\mathcal{M}}$, мы получим довольно точное представление об упоминаемых выше типах горизонтов.

При работе с такими конформными представлениями пространства-времени принято придерживаться следующего соглашения: нулевые конусы направлены вверх, и (как правило) их поверхности расположены под уклоном примерно 45° к вертикали. Это показано на рис. 3.18 и 3.19; они, как и рис. 1.43 в разделе 1.15, являются образцами *схематичных* конформных диаграмм, представляющих собой *качественные* модели, на которых мы пытаемся расположить элементы таким образом, чтобы поверхности нулевых конусов располагались под углом около 45° к вертикали. (Также можно представить себе схематичную конформную диаграмму, на которой показана полностью возмущенная модель Вселенной, содержащая множество черных дыр.) При положительной космологической постоянной величина $\mathcal{F}$ оказывается пространственноподобной [Penrose, 1965b; Penrose and Rindler, 1986], и это подразумевает, что область, заключенная в рамках горизонта событий, для любой мировой линии будет пространственно-конечной в любой конкретный момент (космического) времени. Чтобы не противоречить современным наблюдениям, но не включать в описание (обычно признаваемый) *инфляционный* этап на самой ранней стадии развития Вселенной (см. раздел 3.9), точка P на рис. 3.18 должна находиться примерно в начале последней четверти линии l_{p}. Такой вариант складывается, если построить эволюцию Вселенной в будущее согласно эйнштейновским уравнениям с наблюдаемым значением Λ (предположительно постоянным) и если учесть наблюдаемую материю [Tod, 2012; Nelson and Wilson-Ewing, 2011]. Если к тому же учесть инфляционную фазу, то в качественном отношении получится примерно такая же картина, как на рис. 3.18, но точка P будет находиться практически у верхнего конца l_{p}, под самой оконечной точкой Q. В таком случае два конуса прошлого на рисунке почти совпадут (см. также рис. 4.3).

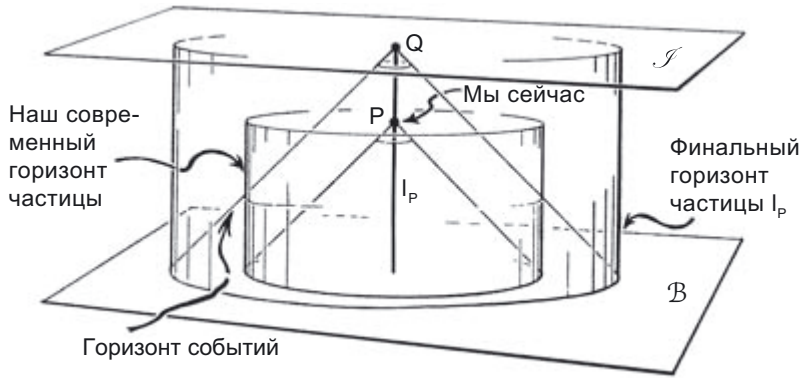


Рис. 3.18. Схематичная конформная диаграмма, на которой показана вся история Вселенной, согласно современной теории, но без инфляционного этапа, который, как принято считать, наступил сразу после Большого взрыва (см. раздел 3.9). Без инфляции наше нынешнее местоположение P находилось бы примерно в начале последней четверти пути на диаграмме (примерно как показано здесь); с инфляцией общая картина была бы похожей в качественном отношении, но точка P была бы практически в самом верху картинки, прямо под Q

На рис. 3.19 приведена сферическая конформная диаграмма, представляющая (не обязательно сферически симметричный) гравитационный коллапс. Показано несколько нулевых конусов, а будущая бесконечность $\mathcal{F}$ считается отсутствующей. На рисунке изображено асимптотически плоское пространство-время с $\Lambda = 0$. При $\Lambda > 0$ будет примерно такая же картина, но с пространственноподобным $\mathcal{F}$, как на рис. 3.18.

Рассматривая пространства-времена, обладающие сферической симметрией (как, например, в ФЛРУ-моделях, показанных на рис. 3.1 и 3.2 в разделе 3.1, или при коллапсе в черную дыру по Оппенгеймеру — Снайдеру, как на рис. 3.9 в разделе 3.2), можно добиться большей точности и компактности, воспользовавшись *строгой* конформной диаграммой (которую в общих чертах описал Брендон Картер в своей докторской диссертации в 1966 году [Carter, 1966]). Это плоская фигура $\mathcal{D}$, представляющая собой область, ограниченную линиями (соответствующими бесконечным областям, сингулярностям или осям симметрии), где считается, что каждая внутренняя точка $\mathcal{D}$ представляет обычную (пространственноподобную) 2-сферу (S^2), так что можно считать, что все пространство-время $\hat{\mathcal{M}}$ заметается при «вращении» области вокруг ее оси (иногда *осей*) симметрии (рис. 3.20). Все *нулевые* направления в $\mathcal{D}$ должны изображаться под углом 45° к вертикали (рис. 3.21). Таким образом можно получить очень хорошее конформное представление искомого пространства-времени $\mathcal{M}$, а затем расширить его до $\hat{\mathcal{M}}$, присоединив к нему его конформную границу.

Рис. 3.19. Схематичная конформная диаграмма, на которой показан коллапс в черную дыру, описанный в разделе 3.9, но сферическая симметрия при этом не предполагается обязательной. Обратите внимание, что (нерегулярная) сингулярность показана пространственноподобной, в соответствии с сильной космической цензурой

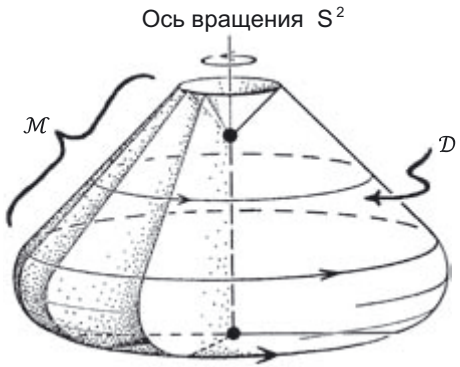


Рис. 3.20. Строгие конформные диаграммы, описывающие сферически симметричные пространства-времена. Плоская область $\mathcal{D}$ вращается вокруг оси S^2, образуя требуемое пространство-время $\mathcal{M}$. Каждая точка на $\mathcal{D}$ представляет («заметает») сферу S^2 в $\mathcal{M}$ с той оговоркой, что любая точка на *оси* (вертикальная граница $\mathcal{D}$, обозначенная штриховой прямой, либо жирная черная точка) соответствует единственной точке в $\mathcal{M}$

Работая с нашими визуализациями, полезно рассуждать в терминах *трехмерного* пространства $\mathcal{M}$, построенного путем вращения $\mathcal{D}$ вокруг вертикальной (то есть временной) оси. Однако необходимо помнить о том, что для получения полного четырехмерного пространства-времени придется вообразить, что это

вращение протекает как *двумерное* сферическое (S^2) действие. Иногда — при рассмотрении моделей, пространственные сечения которых представляют собой *3-сферы* (S^3), — приходится рассматривать случаи с вращением вокруг *двух* осей, а такие случаи порой гораздо сложнее визуализировать! Существуют различные условные обозначения, применяемые при работе со строгими конформными диаграммами; они приведены на рис. 3.22.

На рис. 3.23 *а* 4-пространство Минковского с конформной границей (см. рис. 1.40 в разделе 1.15) показано в виде строгой конформной диаграммы. На рис. 3.23 *б* оно изображено как часть модели эйнштейновской стационарной Вселенной ($S^3 \times \mathbb{R}$) (см. рис. 1.43 в разделе 1.15) в соответствии с рис. 1.43 *б*. Эйнштейнов-

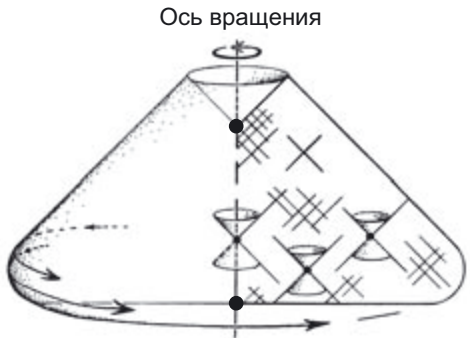


Рис. 3.21. На строгой конформной диаграмме нулевые направления в $\mathcal{D}$ выравниваются под углом 45° к вертикали и, соответственно, являются пересечениями $\mathcal{D}$ с нулевыми конусами в $\mathcal{M}$

Рис. 3.22. Стандартные условные обозначения на строгих конформных диаграммах

ская модель (показанная на рис. 1.42), в свою очередь, изображена как строгая конформная диаграмма на рис. 3.23 *в* (еще раз отметим: здесь используются *две* оси вращения, благодаря которым образуется S^3) или в другом варианте на рис. 3.23 *г*, если мы также хотим включить в нее (конформно сингулярные) граничные точки на прошлой и будущей бесконечностях.

При помощи конформных диаграмм можно представить и многие другие модели Вселенных. На рис. 3.24 я изобразил такие диаграммы для трех фридмановских моделей с $\Lambda = 0$ (в общих чертах показаны ранее, на рис. 3.2 в разделе 3.1), а на рис. 3.25 представлены конформные диаграммы для $\Lambda > 0$ с достаточно большим $\Lambda > 0$ (вместе показаны на рис. 3.3 в разделе 3.1). Строгая конформная диаграмма для де-ситтеровского 4-пространства (вспомните рис. 3.4 *а*) представлена на рис. 3.26 *а*, а на рис. 3.26 *б* мы видим устаревшую стационарную модель Бонди, Голда и Хойла (см. раздел 3.2). На рис. 3.26 *в* и *г* представлены строгие конформные диаграммы для свернутого и развернутого анти-де-ситтеровского пространства — соответственно $\mathcal{A}^4$ и $\Upsilon\mathcal{A}^4$ (ср. с разделом 1.15). На рис. 3.27 приведена строгая конформная диаграмма той схемы, которая была показана ранее на рис. 3.17, и здесь роль различных горизонтов просматривается гораздо более отчетливо, чем прежде.

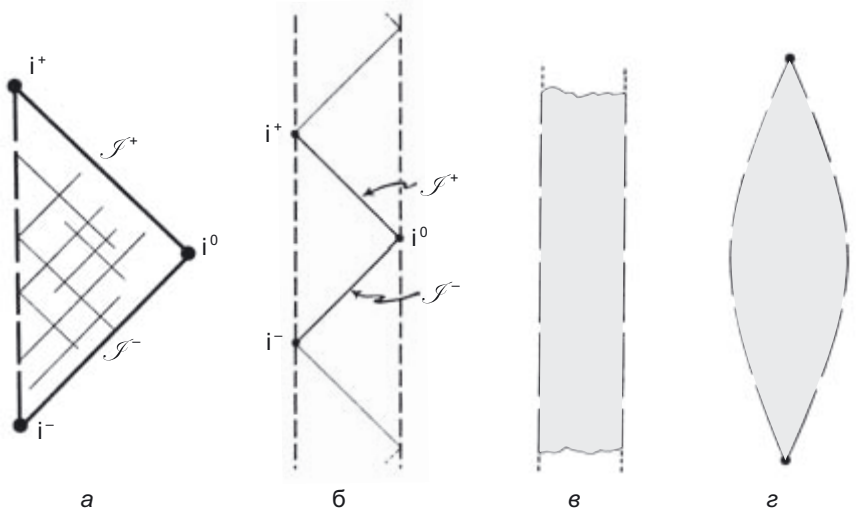


Рис. 3.23. Строгие конформные диаграммы для пространства Минковского и его расширений: *а*) пространство Минковского; *б*) пространство Минковского как часть вертикальной последовательности пространственных диаграмм Минковского, образующих эйнштейновскую Вселенную $\mathcal{E}$; *в*) эйнштейновская Вселенная $\mathcal{E}$ с топологией $\mathbb{R}^1 \times S^3$; *г*) вновь $\mathcal{E}$ — черными кружками обозначены прошлая и будущая бесконечности

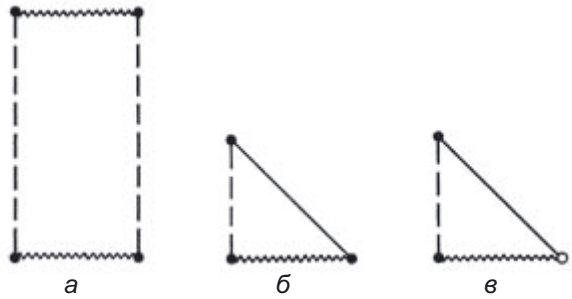


Рис. 3.24. Строгие конформные диаграммы для фридмановских пылевых моделей ($\Lambda = 0$) с рис. 3.2: *а*) $K > 0$; *б*) $K = 0$; *в*) $K < 0$, где пространственная S^2-конформная бесконечность гиперболической геометрии (соответствующая представлениям Бельтрами с рис. 3.1 *в* и 1.38 *б*) обозначена светлым кружком справа внизу

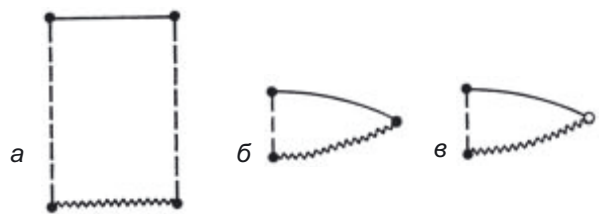


Рис. 3.25. Строгие конформные диаграммы для фридмановских моделей с $\Lambda > 0$, иллюстрирующие пространственноподобную будущую бесконечность $\mathcal{F}$: *а*) $K > 0$ при достаточно большом значении Λ в итоге приводит к экспоненциальному расширению; *б*) $K = 0$; *в*) $K < 0$

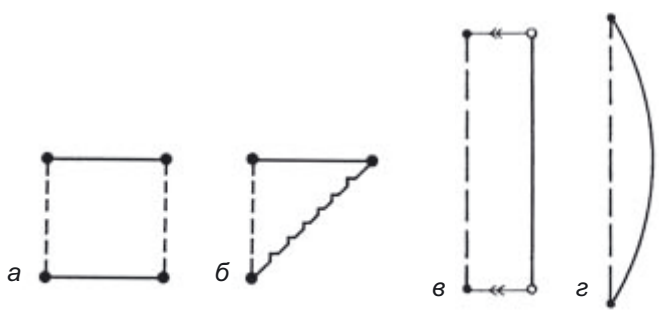


Рис. 3.26. Строгие конформные диаграммы: *а*) полного де-ситтеровского пространства; *б*) части де-ситтеровского пространства — модель описывает стационарную Вселенную (см. рис. 3.4 *в*), *в*) анти-де-ситтеровского пространства $\mathcal{A}^4$, где верхний край отождествляется с нижним, образуя цилиндр; *г*) развернутого анти-де-ситтеровского пространства $\Upsilon\mathcal{A}^4$ (см. раздел 1.15). Диаграммы для $\mathcal{A}^5$ и $\Upsilon\mathcal{A}^5$ не отличаются от приведенных здесь на рис. *в* и *г*, но в них вращение происходит в S^3, а не в S^2

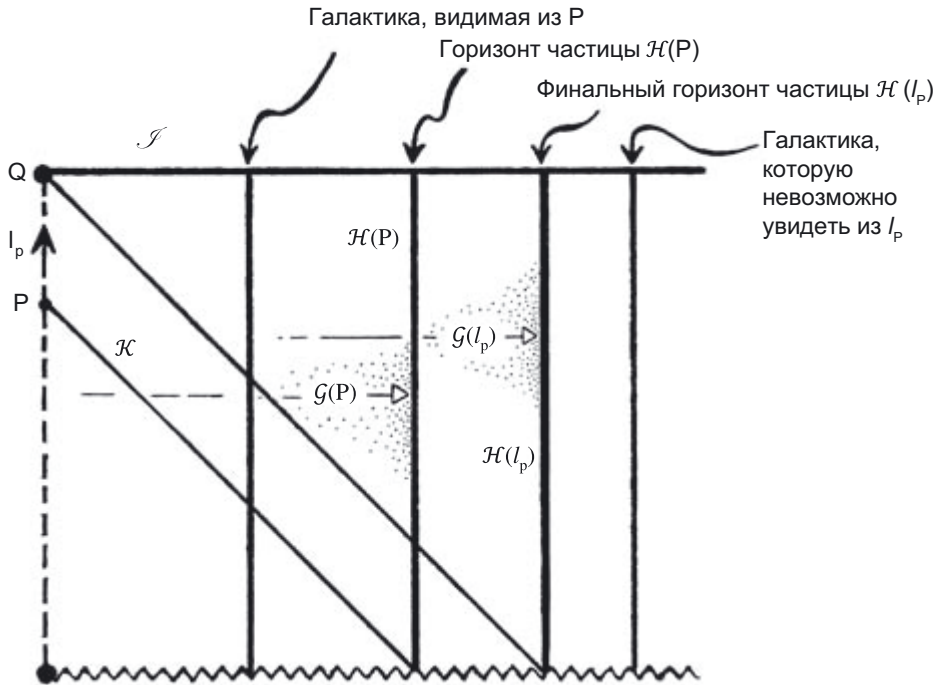


Рис. 3.27. Строгая конформная диаграмма, демонстрирующая те же самые космологические свойства, что проиллюстрированы на рис. 3.17, но на этот раз гораздо нагляднее

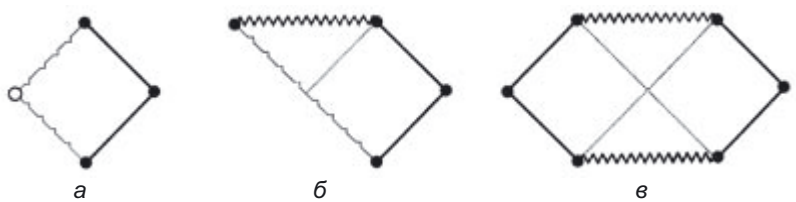


Рис. 3.28. Строгие конформные диаграммы для шварцшильдовской метрики и ее расширений: *а*) исходное пространство-время Шварцшильда; *б*) расширение Эддингтона — Финкельштейна, продолженное за верхний горизонт; *в*) максимальное расширение Синга — Крускала — Секереша

Решение Шварцшильда (см. раздел 3.2) в исходной форме, завершающееся на радиусе Шварцшильда, показано в виде строгой конформной диаграммы на рис. 3.28 *а*. На рис. 3.28 *б* показано его продолжение за горизонт событий, соответствующее метрике, которая известна под названием *форма Эддингтона — Финкельштейна*. Она упоминается в разделе 3.2 и описывает пространство-время в черной дыре. На рис. 3.28 *в* показана *максимально расширенная* форма решения

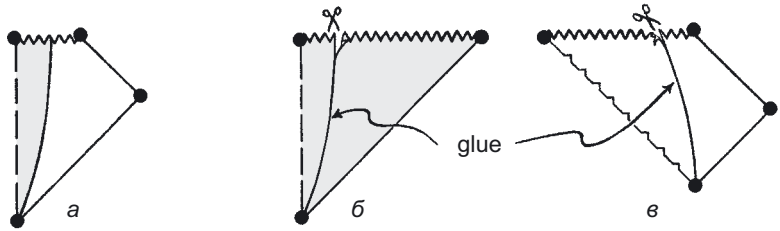


Рис. 3.29. Строгая конформная диаграмма коллапса в черную дыру по Оппенгеймеру и Снайдеру (*а*). Мы получаем ее, склеивая фрагмент (*б*) обращенной во времени фридмановской картины (см. рис. 3.24 *б*) с фрагментом (*в*) картины Эддингтона — Финкельштейна (см. рис. 3.28 *б*). Области, заполненные материей (пылью), показаны серым тоном

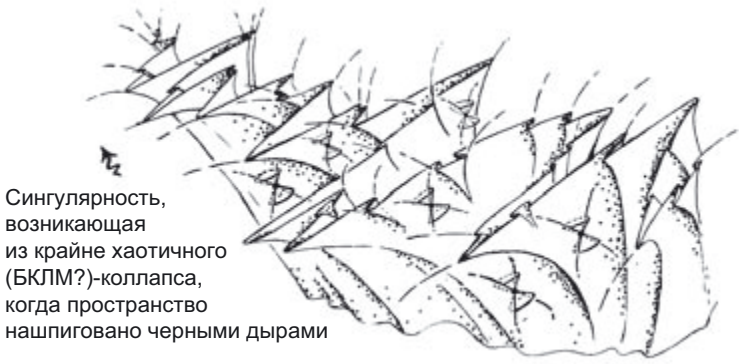


Рис. 3.30. Попытка хотя бы намекнуть на хаотичное поведение пространства-времени, возникающее при приближении к типичной БКЛМ-сингулярности из прошлого

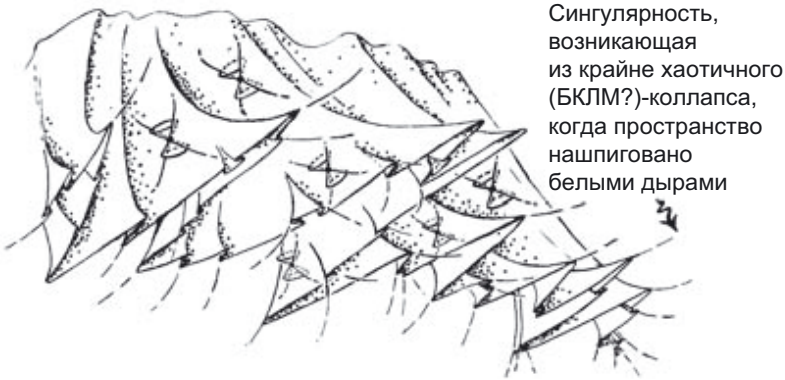


Рис. 3.31. Время, обратное показанному на рис. 3.30: попытка дать подсказку о диком пространстве-времени, возникающем из типичной начальной БКЛМ-сингулярности

Шварцшильда (решение Синга — Крускала), впервые найденная Джоном Лайтоном Сингом [Synge, 1950], а примерно спустя десять лет — и другими учеными (см., в частности, [Kruskal, 1960; Szekeres, 1960]). На рис. 3.29 *а* в виде строгой конформной диаграммы изображен коллапс в черную дыру по Оппенгеймеру и Снайдеру с рис. 3.9, а на рис. 3.29 *б* и *с* показано, как сконструировать пространство-время, сложив фрагменты рис. 3.24 *б* (с обращением времени) и рис. 3.28 *б*.

Мне придется вернуться к той проблеме, которая наиболее всего интересовала нас в разделе 3.4, а именно к парадоксу: хотя, согласно второму закону, Большой взрыв должен был обладать беспрецедентно низкой энтропией, известные нам прямые данные об этом событии, свидетельствующие о несомненно *тепловой* природе РИ, позволяют заключить, что это состояние отличалось *максимальной* энтропией. Как мы убедились в разделе 3.4, ключ к решению этого парадокса кроется в потенциальной пространственной неоднородности, а также в том, каков должен быть коллапс в сингулярность в такой Вселенной, для которой характерна пространственная неоднородность. Эта ситуация изображена на рис. 3.14 *а* и *б*, где слияние множества сингулярностей — черных дыр привело бы к возникновению всеобъемлющей исключительно сложной сингулярности. Сложно проиллюстрировать такое беспорядочное схлопывание на схематичной конформной диаграмме, особенно если мы не списываем со счетов гипотезу сильной космической цензуры (см. раздел 3.4), которая вполне может оказаться верна. Такой коллапс, вероятно, может каким-то образом согласовываться с БКЛМ-процессом, упомянутым в конце раздела 3.2, а на рис. 3.30 я попытался намекнуть, каким образом такой безудержный сингулярный процесс может покатиться к БКЛМ-коллапсу. *Обратив время* (рис. 3.31) в этой исключительно

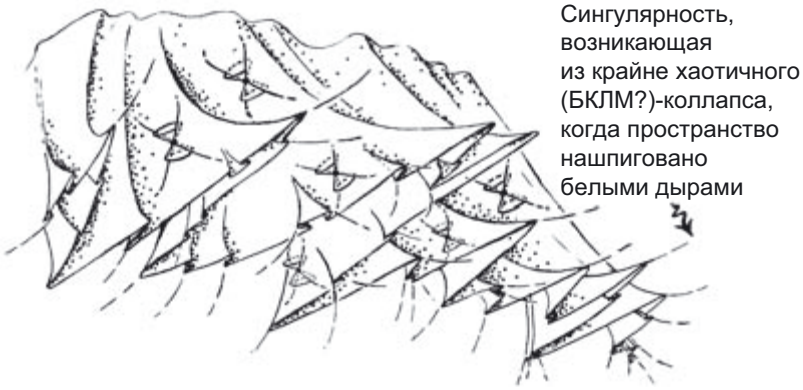


Рис. 3.31. Та же модель, что и на рис. 3.30, но с обращением времени; помогает составить впечатление о том, каким хаотичным образом пространство-время возникает из типичной исходной БКЛМ-сингулярности

высокоэнтропийной ситуации (см. также рис. 3.14 *в*), мы получили бы сингулярность, принципиально отличную по структуре от тех, что возникают в ФЛРУ-модели. Сингулярность такого характера определенно не удастся реалистично описать в терминах *строгой* конформной диаграммы, хотя, по-видимому, ничто не мешает изобразить данную исключительно сложную типичную ситуацию на схематичной конформной диаграмме (см. рис. 3.31). Представляется, что наш конкретный Большой взрыв был по природе именно таков, что его вполне можно смоделировать как ФЛРУ-сингулярность (а также он допускает конформную расширяемость в прошлое, и именно этой проблемой мы вплотную займемся в разделе 4.3), и этот факт чрезвычайно нас ограничивает. Данное ограничение таково, что при нем энтропия может быть лишь исчезающе малой по сравнению с громадными значениями энтропии, допускаемыми в «типичных» пространственно-временных сингулярностях. В следующем разделе мы рассмотрим, насколько обескураживающие ограничения в самом деле присущи сингулярности, которая по типу сильно напоминает модель ФЛРУ.

3.6. Феноменальная точность Большого взрыва

Для того чтобы составить примерное представление о том, насколько невообразимо может возрасти энтропия, если мы позволим себе отступить от однородности ФЛРУ и присовокупить к общей картине гравитацию, давайте поговорим о черных дырах. По-видимому, эти объекты характеризуются максимальной гравитационной энтропией, поэтому неизбежно возникает вопрос: а какую энтропию им присвоить? На самом деле есть чудесная формула для определения энтропии черной дыры S_{bh}, впервые в общем виде полученная Яаковом Бекенштейном [1972, 1973], который воспользовался некоторыми общими и особенно убедительными физическими аргументами; затем ее доработал Стивен Хокинг [1974, 1975, 1976a] (получивший в итоговой формуле точный коэффициент 4 в знаменателе). Хокинг опирался на классические рассуждения с использованием квантовой теории поля в контексте искривленного пространства-времени, описывающие коллапс в черную дыру. Вот эта формула:

$$S_{bh} = \frac{Akc^3}{4\gamma\hbar},$$

где A — площадь горизонта событий черной дыры (вернее, пространственного сечения, проведенного через горизонт; см. рис. 3.9 в разделе 3.2); k, γ и $\hbar$ — соответственно постоянные Больцмана, Ньютона и Планка (последняя — в версии Дирака), а c — скорость света. Следует отметить, что для невращающейся черной дыры массы m выполняется соотношение

$$A = \frac{16\pi\gamma^2}{c^4} m^2,$$

так что

$$S_{bh} = \frac{4\pi m^2 k\gamma}{\hbar c}.$$

Черная дыра также может вращаться, и если ее угловой момент равен am, то получаем (см. [Kerr, 1963; Boyer and Lindquist, 1967; Carter, 1970]):

$$A = \frac{8\pi\gamma^2}{c^4} m(m + \sqrt{m^2 - a^2}),$$

так что:

$$S_{bh} = \frac{2\pi k\gamma}{\hbar c} m(m + \sqrt{m^2 - a^2}).$$

Далее было бы удобно пользоваться так называемыми *естественными единицами* (еще они называются *планковскими* или *абсолютными*) длины, времени, массы и температуры, каждая из которых определяется следующим образом:

$$c = \gamma = \hbar = k = 1,$$

и эти натуральные единицы (приблизительно) пересчитываются в более привычные физические единицы так:

1 метр = $6{,}2 \times 10^{34}\, l_{\text{P}}$

1 секунда = $1{,}9 \times 10^{43}\, t_{\text{P}}$

1 грамм = $4{,}6 \times 10^{4}\, m_{\text{P}}$

1 кельвин = $7{,}1 \times 10^{-33}\, T_p$

Космологическая постоянная = $5{,}6 \times 10^{-122}\, t_{\text{P}}^{-2}$

Теперь все меры превратились просто в безразмерные числа. Тогда вышеприведенные формулы (для невращающейся черной дыры) упрощаются до:

$$S_{bh} = \frac{1}{4} A = 4\pi m^2, \quad A = 16\pi m^2.$$

Значение энтропии в данном случае получается колоссальным, поскольку, как мы полагаем, черные дыры должны возникать при астрофизических процессах (энтропия безразмерна, поэтому не имеет значения, какие единицы выбирать, но удобнее считать, используя естественные единицы). Вероятно, такая колос-

сальность энтропии не покажется удивительной, если задуматься, насколько «необратимы» процессы, связанные с черными дырами. Часто указывают, как огромна энтропия РИ — около 10^8 или 10^9 на каждый барион (см. раздел 1.3) во Вселенной, и эта энтропия гораздо выше, чем при типичных астрофизических процессах. Тем не менее такая энтропия просто меркнет по сравнению с той энтропией, которую следует приписывать черным дырам, в особенности тем наиболее крупным из них, которые располагаются в центрах галактик. Предполагается, что обычная черная дыра, сравнимая по массе со звездой, должна обладать энтропией около 10^{20} на барион. Однако в нашей Галактике есть черная дыра, масса которой равна примерно четырем миллионам солнечных масс, и энтропия ее должна составлять где-то 10^{26} на барион или более. Едва ли можно предположить, что основная *масса* Вселенной в настоящее время сосредоточена в черных дырах; однако мы увидим, как черные дыры могут доминировать по части *энтропии*, если обсудим модель, согласно которой наблюдаемая Вселенная наполнена галактиками, подобными нашему Млечному Пути, в каждой из которых 10^{11} звезд, а в центре черная дыра в 10^6 раз тяжелее Солнца (пожалуй, здесь мы даже занижаем современную среднюю оценку массы, приходящейся на долю черной дыры). Итак, мы получаем общее значение энтропии около 10^{21} на барион, по сравнению с которым энтропия РИ 10^8 или 10^9 просто ничтожна.

Вышеприведенные выражения позволяют заключить, что в больших черных дырах энтропия на барион, вероятно, может быть гораздо выше, возрастая пропорционально массе черной дыры. Следовательно, если говорить о некотором количестве вещества с известной массой, данное вещество достигнет максимальной энтропии, если вся эта масса сосредоточится в одной черной дыре. Если мы сосредоточим в черных дырах всю барионную массу в наблюдаемой части Вселенной (а наблюдаемая часть Вселенной — это область, ограниченная текущим горизонтом частицы), то, считая, что общее количество барионов составляет порядка 10^{80}, получим общую энтропию порядка 10^{123}, тогда как огненный шар, очевидно, породивший РИ, должен был бы обладать энтропией всего лишь ~10^{89}.

В вышеизложенных рассуждениях я до сих пор обходил вниманием тот факт, что на барионную материю, по-видимому, приходится всего около 15 % вещества во Вселенной, а оставшиеся 85 % — это так называемая *темная материя*. (Здесь я также не учитываю темную энергию, то есть Λ, поскольку считаю Λ космологической постоянной, а не некоей «субстанцией», которая могла бы участвовать в гравитационном коллапсе. Проблему энтропии, связанной с Λ, мы обсудим в разделе 3.7.) Вполне можно предположить, что наша гипотетическая черная дыра, вмещающая все содержимое наблюдаемой части Вселенной, должна также включать и массу, приходящуюся на темную материю. В результате максимальное значение энтропии составит скорее около 10^{124} или 10^{125}. Однако в данном

контексте я остановлюсь на более осторожной оценке 10^{123} отчасти потому, что мы, по-видимому, совершенно не представляем, из чего может состоять темная материя. Еще одна причина осторожнее относиться к максимальным значениям заключается в том, что могут существовать и подлинно геометрические проблемы, не позволяющие построить такую подходящую модель расширяющейся Вселенной, в которой всю материю удалось бы «уложить» в одной черной дыре. Возможно, с физической точки зрения реалистичнее представить несколько более мелких черных дыр, разбросанных по всей наблюдаемой части Вселенной. Поэтому в данном случае мы добьемся гораздо более достоверной картины, если воспользуемся коэффициентом 10.

Здесь также следует прояснить еще один момент. Термин *наблюдаемая Вселенная* обычно относится к материи, которая укладывается в световой конус прошлого нашего актуального местоположения P в пространстве-времени, как показано на рис. 3.17 и 3.27. При условии, что мы рассматриваем стандартные космологические модели, это совершенно очевидно, хотя и есть одна небольшая проблема: включать ли в модель события, происходившие до эпохи последнего рассеяния, — ведь они тоже входят в наш световой конус прошлого. Разница здесь несущественная, пока мы не начинаем учитывать подразумеваемый в большинстве случаев *инфляционный* этап эволюции ранней Вселенной, о котором мы поговорим в разделе 3.9. В таком случае кардинально увеличится и расстояние за пределами горизонта частицы, и объем материи, который там бы умещался. Как правило, этот ранний инфляционный период не учитывается в определении термина *горизонт частицы*. Я также буду придерживаться здесь такого принципа.

Необходимо учитывать следующее: рассуждая о том, чем был так необычен Большой взрыв, мы рассматривали в разделе 3.5 модель коллапса, *обращенного во времени*. У нас получалось, что такая гипотетическая модель коллапса оканчивается сингулярностью; сначала образуется множество мелких черных дыр, а затем они сливаются во все более крупные. Пожалуй, достаточно логично предположить, что в конечном итоге такая модель приводит нас к единственной черной дыре, поглотившей основную массу материи (пусть даже и не всю). Здесь нужно отметить, что, смоделировав обращение во времени для такого совершенно беспорядочного коллапса, мы приходим к выводу о том, что максимально-энтропийный Большой взрыв представлял собой не собственно взрыв черной дыры, а обращение коллапса такой дыры во времени — явление, зачастую именуемое «белая дыра». Чтобы представить себе пространство-время, описывающее белую дыру, просто перевернем рис. 3.9 вверх ногами — см. рис. 3.32. Следовательно, строгая конформная диаграмма белой дыры получится такой, как на рис. 3.33, — это перевернутый рис. 3.29 *a*. С математической точки зрения вполне вероятно, что такая конфигурация вполне может быть элементом гораздо

более масштабного большого взрыва, чем описывается в ФЛРУ-модели, и что исходная энтропия в данном случае могла достигать колоссальной величины ~10^{123}, а не тех ничтожных 10^{89}, которыми должен был обладать первозданный пылающий шар, если судить по РИ.

В предыдущем абзаце я строил аргументацию в контексте черной, а не белой дыры, но с точки зрения расчета значений энтропии эти ситуации одинаковы. Как мы помним из раздела 3.3, согласно больцмановскому определению,

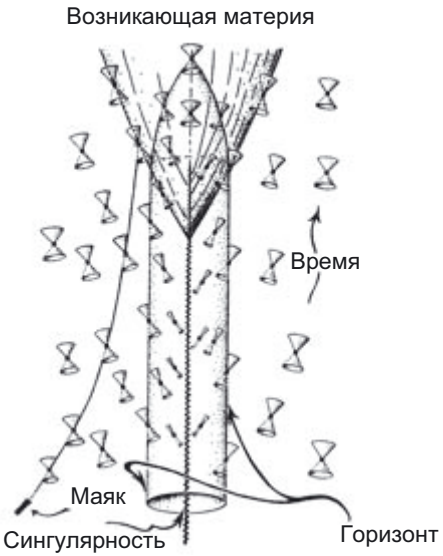


Рис. 3.32. Пространственно-временное представление гипотетической «белой дыры»: та же картина, что на рис. 3.9, но обращенная во времени. До взрыва дыры с образованием материи свет, излученный извне (например, исходящий от изображенного здесь маяка), не может проникнуть через горизонт

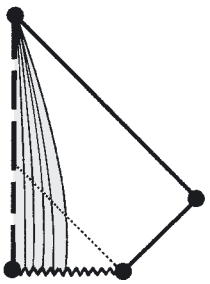


Рис. 3.33. Строгая конформная диаграмма сферически симметричной белой дыры

энтропия зависит от объемов областей с крупнозернистым разбиением в фазовом пространстве. Сама по себе природа фазового пространства не зависит от направления времени (так как при обращении времени значения всех импульсов просто меняются на обратные), а макроскопические критерии для определения областей с крупнозернистым разбиением не должны зависеть от направления времени. Разумеется, напрасно рассчитывать, что в нашей Вселенной найдутся белые дыры — ведь они бы грубо нарушали второй закон. Однако их совершенно допустимо учитывать в вышеизложенных рассуждениях при расчете степени «своеобразия» Большого взрыва, если мы собираемся думать о таких состояниях, которые *действительно* нарушают второй закон.

Итак, мы приходим к выводу, что при значении энтропии как минимум 10^{123} могла существовать пространственно-временная сингулярность, из которой и возникла Вселенная, причем единственное, что требуется, это чтобы такая картина согласовывалась с (симметричными во времени) уравнениями общей теории относительности, с источниками обычной материи и с общим числом барионов в наблюдаемой части Вселенной (около 10^{80}) плюс сопутствующая темная материя. Соответственно, следует допустить существование фазового пространства $\mathcal{P}$ с полным объемом не менее

$$V = e^{10^{123}}$$

(поскольку при $k = 1$ формула Больцмана $S = \log V$ из раздела 3.3 должна допускать $S = 10^{123}$, см. раздел А.1). На самом деле, как показано в разделе А.1, практически не имеет значения (как минимум с точностью до числа 123 в этом выражении), заменим ли мы «e» на 10, поэтому я буду говорить о том, что пространство $\mathcal{P}$ имеет общий объем как минимум

$$10^{10^{123}}.$$

То, что мы на самом деле наблюдаем в нашей Вселенной, — это пылающий шар в эпоху последнего рассеяния, энтропия которого составляла не более 10^{90} (здесь учитываем 10^{80} барионов с энтропией 10^9 на каждый, а также добавляем темную материю). Он должен был занимать область $\mathcal{D}$ с крупнозернистым разбиением значительно меньшего объема:

$$10^{10^{90}}.$$

Насколько меньше этот объем, если выразить его как долю полного фазового пространства $\mathcal{P}$? Очевиден ответ:

$$\frac{10^{10^{90}}}{10^{10^{123}}},$$

что, как показано в разделе А.1, практически неотличимо от

$$\frac{1}{10^{10^{123}}}, \text{ то есть } 10^{-10^{123}},$$

поэтому мы даже не заметим объема $\mathcal{D}$ — столь колоссален общий объем $10^{10^{123}}$, который нам придется закрепить за $\mathcal{P}$. Теперь уже можно постепенно себе представить, с какой исключительной *точностью* создавалась Вселенная в том виде, как мы понимаем ее сегодня. В период между исходной сингулярностью, которую мы можем представить как совсем крошечную область $\mathcal{B}$ с крупнозернистым разбиением в фазовом пространстве $\mathcal{P}$, и эпохой последнего рассеяния могли происходить процессы с существенным возрастанием энтропии (см. рис. 3.34; в подписи указаны величины, характеризующие вклад темной материи). Следовательно, можно ожидать, что возникновение Вселенной было еще более точным процессом, и мерой точности теперь выступает размер региона $\mathcal{B}$ в фазовом пространстве $\mathcal{P}$. Здесь опять имеем число $10^{-10^{123}}$, то есть регион $\mathcal{B}$ был еще меньше, что не так очевидно из-за громадного объема самого пространства $\mathcal{P}$, который составляет $10^{10^{123}}$.

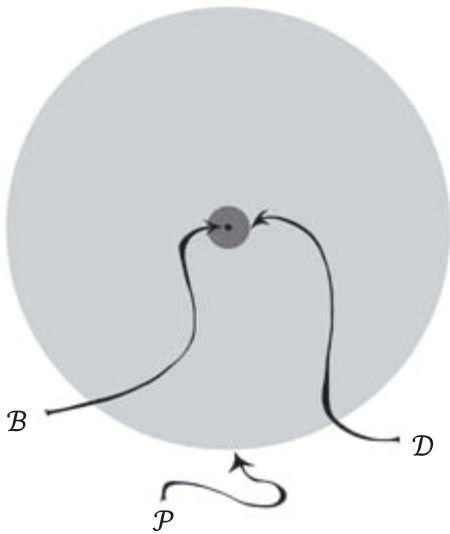


Рис. 3.34. Объем всего фазового пространства $\mathcal{P}$ всей нашей наблюдаемой Вселенной в планковских единицах (или других традиционных единицах) составляет около $10^{10^{124}}$. Область $\mathcal{D}$, соответствующая состояниям эпохи последнего рассеяния, имеет несоизмеримо меньший объем — около $10^{10^{90}}$, а объем региона $\mathcal{B}$ (состояний, представляющих Большой взрыв), был бы еще значительно меньше: каждое составило бы всего $\sim 10^{-10^{124}}$ от полного. Такая схема не дает даже отдаленного представления о несопоставимости этих объемов

3.7. Космологическая энтропия?

Есть еще одна проблема, о которой следует поговорить в контексте того, каков вклад в энтропию некоей («темной») субстанции: как оценить вклад так называемой *темной энергии*, то есть Λ (в моей интерпретации этого термина). Многие физики считают, что из-за наличия Λ нашу постоянно расширяющуюся Вселенную в отдаленном будущем ждет колоссальная энтропия, которая «подключится» на определенном очень позднем (но не указанном) этапе истории Вселенной. Такая точка зрения базируется прежде всего на распространенном и хорошо обоснованном представлении [Gibbons and Hawking, 1977], согласно которому космологические горизонты событий, возникающие в таких моделях, следует интерпретировать точно так же, как горизонты черных дыр. Поскольку рассматриваемый нами горизонт просто колоссален — эта область значительно превосходит по размерам самую крупную из известных к настоящему времени черных дыр (масса которой, по-видимому, составляет около 4×10^{10} солнечных) примерно в 10^{24} раз, мы получим немыслимо огромную «энтропию» S_{cosm}, значение которой составляет приблизительно

$$S_{\text{cosm}} \approx 6{,}7 \times 10^{122}.$$

Данная величина высчитана непосредственно на основе того значения, которым, по современным оценкам, обладает Λ:

$$\Lambda = 5{,}6 \times 10^{-122}$$

по формуле энтропии Бекенштейна — Хокинга (предполагается, что мы допускаем ее использование в таких обстоятельствах), применяемой к области A_{cosm} космологического горизонта событий. Эта область составляет ровно

$$A_{\text{cosm}} = \frac{12\pi}{\Lambda}.$$

Следует отметить, что если верить такой трактовке горизонтов так же, как мы доверяем аргументации Бекенштейна — Хокинга об энтропии черных дыр, то это выражение должно представлять *полное* значение энтропии, а не только вклад «темной энергии». Однако мы обнаруживаем, что площадь горизонта позволяет вывести данное значение S_{cosm} исключительно из значения Λ, совершенно без учета деталей распределения материи или других частных отступлений от точной де-ситтеровской геометрии, показанной на рис. 3.4 (см. [Penrose, 2010], п. B5). Тем не менее, хотя и кажется, что S_{cosm} ($\sim 6 \times 10^{122}$) немного не дотягивает до полного значения энтропии, которое, согласно моим предыдущим рассуждениям, составит (с учетом темной материи) ~ 10^{124}, эта величина, вероятно,

будет неизмеримо выше максимальной полной энтропии порядка 10^{110}, которая потенциально может достигаться в черных дырах с учетом лишь барионной материи, содержащейся в современной наблюдаемой Вселенной, или порядка 10^{112}, если присовокупить к ней еще и темную материю.

Однако необходимо поставить вопрос: чему же соответствует итоговое значение «полной энтропии» S_{cosm} ($\sim 6 \times 10^{122}$). Поскольку оно зависит только от Λ и совершенно не связано с деталями материального состава Вселенной, вполне можно считать, что S_{cosm} — это энтропия, присущая *всей* Вселенной. Но если Вселенная пространственно бесконечна (многие космологи придерживаются такого мнения), то конечное значение энтропии окажется распределено по бесконечному объему пространства и составит лишь исчезающе малый вклад в энтропию рассматриваемой здесь конечной сопутствующей области. При такой интерпретации значение космологической энтропии 6×10^{122} будет соответствовать *нулевой плотности энтропии*, и, следовательно, она должна полностью игнорироваться в рассуждениях о балансе энтропии в нашей динамической Вселенной.

С другой стороны, можно попытаться предположить, что данное значение энтропии относится только к сопутствующему объему, основанному на материальном содержимом наблюдаемой Вселенной, то есть к сопутствующему объему $\mathcal{G}(\mathrm{P})$ внутри нашего горизонта частицы $\mathcal{H}(\mathrm{P})$ (см. раздел 3.5, рис. 3.17 и 3.27), где P — наше нынешнее положение в пространстве-времени. Однако такое предположение не имеет убедительного обоснования, в частности, потому, что «настоящий момент», соответствующий точке P на нашей мировой линии l_{P}, не имеет в данном контексте никакого особого значения. Кажется более уместным оперировать сопутствующим объемом $\mathcal{G}(l_{\mathrm{P}})$, рассмотренным в разделе 3.5, где наша мировая линия l_{P} должна бесконечно продолжаться в будущем. Мера материи, содержащейся в этом регионе, не зависит от момента «времени», в который мы наблюдаем Вселенную. Это вся совокупность материи, которая *когда-либо* будет присутствовать в обозримой Вселенной. В конформной картине (см. рис. 3.18 в разделе 3.5) максимально протяженная l_{P} достигает конформной бесконечности будущего $\mathcal{F}$ (как мы помним, $\mathcal{F}$ — это *пространственноподобная* гиперповерхность при $\Lambda > 0$) в некоторой точке Q, и нас теперь интересует общее количество материи, пересекающей световой конус прошлого $\mathcal{C}_{\mathrm{Q}}$, выходящий из точки Q. В сущности, этот световой конус прошлого и является нашим *космологическим горизонтом событий* и обладает более «абсолютным» характером, чем та материя, которая находится внутри нынешнего горизонта частицы. С течением времени наш горизонт частицы расширяется, и материя в объеме $\mathcal{G}(l_{\mathrm{P}})$ представляет предел такого расширения.

На самом деле (предполагая, что эволюция во времени описывается уравнениями Эйнштейна при наблюдаемом положительном значении Λ, которое мы считаем постоянным) получаем, что общее количество материи, захватываемое $\mathcal{C}_Q$, примерно в 2,5 раза больше, чем умещается в пределах нашего нынешнего горизонта частицы [Tod, 2012; Nelson and Wilson-Ewing, 2011]. Величина максимальной возможной энтропии, которой могла бы достичь вся эта материя, если вся она попадет в одну черную дыру, соответственно более чем впятеро больше максимума, полученного нами для той материи, которая просто укладывается в наш нынешний горизонт частицы. Это более крупное значение составляет $\sim 10^{124}$, а не $\sim 10^{123}$, как мы получили выше. Если учесть при этом и темную материю, то получим $\sim 10^{125}$. Это значение в несколько сотен раз больше S_{cosm}, поэтому, выбрав модель примерно с такой же средней плотностью материи, как в нашей Вселенной, но с достаточным количеством черных дыр, мы, по-видимому, могли бы превысить то предельное значение, которое предположительно устанавливает нам S_{cosm}, а это грубо противоречит второму закону! (Есть еще одна проблема, связанная с тем, что черные дыры в конечном итоге испаряются благодаря механизму, описанному Хокингом, но она не нарушает изложенных здесь рассуждений (см. [Penrose, 2010], раздел 3.5).)

С учетом того, что все эти числа довольно приблизительны, в данном случае по-прежнему кажется вполне правдоподобным, что значение 6×10^{122}, полученное нами для космологической энтропии, на самом деле является «истинной» максимальной энтропией, достижимой для того объема материи, что лежит в пределах $\mathcal{C}_Q$. Однако кроме вышеизложенных существуют и более серьезные причины сомневаться в том, что S_{cosm} — это действительно максимальная энтропия, достижимая в данном сегменте Вселенной, либо что такая «энтропия» вообще имеет физический смысл. Вернемся к исходному аргументу, где предлагалась аналогия между космологическим горизонтом $\mathcal{C}_Q$ и горизонтом событий черной дыры $\mathcal{E}$. Если попытаться при помощи этой аналогии ответить на вопрос, к какой части Вселенной на самом деле может относиться эта космологическая энтропия, то обнаруживается любопытное противоречие. С учетом вышеизложенного мы уже убедились, что эта часть не может быть *всей* Вселенной. Казалось логичным предположить, что часть, о которой идет речь, — это просто область, охватываемая космологическим горизонтом. Однако если сравнить эту космологическую ситуацию с черной дырой, то окажется, что такая интерпретация отнюдь не логична. В случае коллапса в черную дыру энтропия Бекенштейна — Хокинга обычно считается энтропией *черной дыры*, и такая трактовка совершенно оправданна. Но если сравнить эту ситуацию с космологической, а горизонт событий черной дыры $\mathcal{E}$ — с космологическим горизонтом $\mathcal{C}_Q$, как показано на строгих конформных диаграммах (см. рис. 3.35),

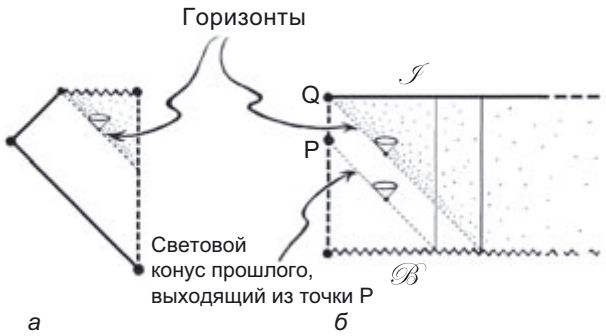


Рис. 3.35. Строгие конформные диаграммы, иллюстрирующие область (обозначена точками), которая соответствует энтропии черной дыры (*а*) и космологии с положительной Λ (*б*). В пространственно-бесконечной Вселенной *плотность* «космологической энтропии» должна быть нулевой.

то окажется, что область пространства-времени *внутри* горизонта черной дыры $\mathcal{C}$ соответствует части Вселенной *вне* космологического горизонта $\mathcal{C}_Q$. Это области, лежащие с «будущей» стороны соответствующих горизонтов, то есть с той стороны, куда направлены нулевые конусы будущего. Как мы убедились выше, в пространственно-бесконечной Вселенной мы в таком случае получим нулевую *плотность* энтропии во всей внешней Вселенной! Опять же, это кажется почти лишенным смысла, если предположить, что на S_{cosm} приходится основной физический вклад в физическую энтропию Вселенной. (Можно выдвигать версии и о том, «где» может находиться энтропия в S_{cosm}, —например, в области пространства-времени, лежащей в условном будущем C_Q, — но, опять же, это кажется почти бессмысленным, так как S_{cosm} совершенно не зависит от любой материи или черных дыр, которые могут попасть в эту область).

В этой аргументации есть момент, который, пожалуй, мог озаботить некоторых читателей, ведь в разделе 3.6 мы ввели концепцию *белой дыры*, а там нулевые конусы направлены вовне, то есть в будущее, прочь от центральной области, — и такая ситуация уже напоминает случай с космологическим горизонтом. Также было указано, что энтропия Бекенштейна — Хокинга должна быть в равной мере применима как к белой, так и к черной дыре, поскольку определение энтропии по Больцману не зависит от направления времени. Следовательно, можно было бы попытаться утверждать, что аналогия между *белой* дырой и космологическим горизонтом, возможно, подтверждает интерпретацию S_{cosm} как реальной физической энтропии. Однако в известной нам Вселенной белые дыры физически невозможны, поскольку грубо противоречат второму закону. В разделе 3.6 они упоминались только как гипотетические объекты. При сравнениях, которые

непосредственно связаны с возрастанием энтропии с течением времени в соответствии со вторым законом, как говорилось двумя абзацами ранее, требуется сравнивать космологический горизонт с горизонтами *черных*, а не белых дыр. Соответственно, та «энтропия», величину которой предположительно дает S_{cosm}, должна была бы относиться к области вне космологического горизонта, а не внутри него, — а это, как указывалось ранее, дало бы нам исчезающее малую плотность энтропии в пространственно-бесконечной Вселенной.

Однако здесь следует обсудить и еще один вопрос, связанный с использованием больцмановской формулы $S = k \log V$ (см. раздел 3.3) в контексте черных дыр. Следует признать, что энтропию S_{bh} для черной дыры (см. раздел 3.6) пока, на мой взгляд, еще не удалось полностью и убедительно отождествить с энтропией больцмановского типа, где четко определен соответствующий объем фазового пространства V. Решить эту задачу пытаются разными способами (см., например, [Strominger and Vafa, 1996; Ashtekar et al., 1998]), но меня по-настоящему не устраивает ни один из них (см. также раздел 1.15 — там описаны идеи, лежащие в основе голографического принципа, который меня также не удовлетворяет). Причины, по которым S_{bh} может максимально серьезно восприниматься в качестве *истинной* меры энтропии черной дыры, отличаются от тех, что до сих пор были связаны с непосредственным использованием формулы Больцмана. Тем не менее эти причины [Bekenstein, 1972, 1973; Hawking, 1974, 1975; Unruh and Wald, 1982] кажутся мне очень вескими и необходимыми для общей непротиворечивости второго закона в квантовом контексте. Хотя в этих случаях формула Больцмана и не используется непосредственно, это никак не подразумевает несогласованности с ней и указывает лишь на характерные трудности, возникающие при попытке вплести непроясненные квантовые феномены фазовых пространств в контекст общей теории относительности (ср. также с разделами 1.15, 2.11 и 4.3).

Читатель, вероятно, уже понимает, что, на мой взгляд, присваивать Вселенной энтропийный вклад $(12\pi/\Lambda)$ исходя из значения Λ — путь с физической точки зрения крайне сомнительный, причем не только по вышеизложенным причинам. Если считать, что S_{cosm} играет какую-либо роль в динамике второго закона и эта роль почему-то проявляется только на очень поздних этапах эволюции нашей Вселенной, при де-ситтеровском экспоненциальном расширении, то нам понадобилась бы какая-то теория о том, «когда» эта энтропия «подключается» к процессу. Де-ситтеровское пространство-время обладает очень высокой степенью симметрии (10-параметрическая группа однородных преобразований Лоренца в пятимерном пространстве; см. раздел 3.1 и, например, [Schrödinger, 1956] и [ПкР], п. 18.2 и 28.4), причем сама суть этого пространства не позволяет естественным образом отметить такое время. Даже если всерьез считать, что

энтропия, задаваемая S_{cosm}, действительно имеет какой-то смысл (например, связана с вакуумными флуктуациями), она, по-видимому, не играет никакой динамической роли во взаимоотношениях с другими формами энтропии. S_{cosm} — это всего лишь *постоянная* величина, какое бы значение мы ни пытались ей присвоить, и она никак не влияет на действие второго закона независимо от того, *собираемся ли мы* учитывать ее как энтропию того или иного рода.

В то же время в случае обычной черной дыры уже выдвигался исходный аргумент Бекенштейна [Bekenstein, 1972, 1973] с описанием мысленных экспериментов, в ходе которых мы медленно погружаем нагретую материю в черную дыру таким образом, чтобы можно было допустить превращение ее тепловой энергии в полезную работу. Оказывается, что, если не присваивать черной дыре такого значения энтропии, которое хотя бы грубо согласовывалось с вышеприведенной формулой S_{bh}, то, в принципе, таким образом мы смогли бы нарушить второй закон. Соответственно, эксперимент показывает, что энтропия Бекенштейна — Хокинга является неотъемлемым элементом, необходимым для общей непротиворечивости второго закона в контексте черных дыр. Такая энтропия четко взаимосвязана с другими формами энтропии и необходима для общей согласованности термодинамики черных дыр. Все это связано с динамикой горизонтов черных дыр и с тем фактом, что горизонты могут увеличиваться при процессах, которые в остальном, по-видимому, могут понижать энтропию, как в случае с погружением нагретой материи в черную дыру с извлечением всего ее массового/энергетического содержимого в виде «полезной» энергии, что нарушает второй закон.

С космологическими горизонтами складывается принципиально иная ситуация. Их положения как таковые очень зависят от наблюдателя, что принципиально отличает их от *абсолютных* горизонтов событий стационарных черных дыр в асимптотически плоском пространстве (см. раздел 3.2). Однако *площадь A* космологического горизонта — это всего лишь *фиксированное* число, определяемое просто значением космологической постоянной Λ в вышеприведенном выражении $12\pi/\Lambda$ и никак не связанное с любыми динамическими процессами, протекающими во Вселенной, например с тем, какое количество массы-энергии проходит сквозь горизонт, или с тем, как распределяется масса, — а все это, несомненно, отражается на *локальной* геометрии горизонта. Эта картина весьма отличается от ситуации с черной дырой, площадь горизонта которой неизбежно увеличивается по мере падения материи за горизонт. Никакие динамические процессы не влияют на значение S_{cosm}, которое, что бы ни случилось, всегда остается равным $12\pi/\Lambda$.

Разумеется, ситуация зависит от того, на самом ли деле постоянна Λ, а не от некоего таинственного неизвестного динамического «поля темной энергии».

Такое «Λ-поле» должно было бы обладать тензором энергии $(8\pi)^{-1}\,\Lambda\mathbf{g}$, так что эйнштейновские уравнения $\mathbf{G} = 8\pi\gamma\mathbf{T} + \Lambda\mathbf{g}$ можно было бы записать в виде

$$\mathbf{G} = 8\pi\gamma(\mathbf{T} + \frac{\Lambda}{8\pi\gamma}\mathbf{g}),$$

как будто в них и нет космологического члена из уточненной теории Эйнштейна, предложенной в 1917 году, а $(8\pi)^{-1}\,\Lambda\mathbf{g}$ считается просто вкладом Λ-поля, добавляемым к тензору энергии-импульса $\mathbf{T}$ всей оставшейся материи, в результате чего получается полный тензор энергии-импульса (в скобках справа). Однако этот дополнительный член совершенно не похож на тот, что обусловлен обычной материей. Наиболее примечательно, что для него характерно гравитационное отталкивание, а также положительная плотность массы-энергии. Более того, если допустить возможность изменения Λ, то возникает множество технических сложностей, наиболее заметная из которых такова: существует серьезная опасность нарушить нулевое энергетическое условие, которое упоминалось в разделе 3.2. Если попытаться представить темную энергию как некую субстанцию либо совокупность субстанций, не взаимодействующих с другими полями, то окажется, что уравнения дифференциальной геометрии (фактически свернутые тождества Бьянки) свидетельствуют о том, что Λ должна быть константой. Однако если допустить отклонение полного тензора энергии-импульса от этой формы, то весьма вероятно, что нулевое энергетическое условие будет нарушено, поскольку оно слабо выполняется, когда тензор энергии-импульса имеет форму $\Lambda\mathbf{g}$.

С проблемой энтропии тесно связан вопрос так называемой космологической температуры T_{cosm}. Если говорить о черной дыре, тот факт, что она должна обладать температурой, связанной с энтропией Бекенштейна — Хокинга для черных дыр (и наоборот), следует из самых базовых принципов термодинамики [Bardeen et al., 1973]. В первых статьях, где Хокинг вывел точную формулу энтропии черных дыр [Hawking, 1974, 1975], он также получил формулу температуры черной дыры, которая, если черная дыра не вращается (и обладает сферической симметрией), дает значение

$$T_{bh} = \frac{1}{8\pi m}$$

в естественных (планковских) единицах. Если говорить о черной дыре такого размера, какая могла бы возникнуть при нормальных астрофизических процессах (где масса m была бы не меньше солнечной), эта температура получается экстремально низкой, а наивысших значений она достигает в случае наименее массивных черных дыр. Даже тогда она получается не намного выше самых низких температур, какие удавалось получить на Земле в лабораторных условиях.

Идея космологической температуры T_{cosm} подсказана аналогией с такой черной дырой, размер горизонта которой был бы равен размеру космологического горизонта $\mathcal{C}_{\text{Q}}$, и в таком случае получаем

$$T_{\text{cosm}} = \frac{1}{2\pi}\sqrt{\frac{\Lambda}{3}}$$

в естественных единицах либо

$$T_{\text{cosm}} \approx 9\times 10^{-62}\,\text{K}$$

в кельвинах.

Это абсурдно низкая температура — она даже ниже, чем хокинговская температура любой черной дыры, которую можно реалистично представить в известной нам Вселенной. Но является ли T_{cosm} реальной температурой в обычном физическом смысле этого слова? По-видимому, среди космологов, всерьез занимающихся этой проблемой, существует единое мнение: да, является.

Существуют различные аргументы, предположительно подкрепляющие такую интерпретацию, и некоторые из них более мотивированы, нежели простая аналогия с черной дырой. Тем не менее все они кажутся мне весьма сомнительными. По-видимому, наиболее математически привлекательным из всех этих аргументов является такой (также используемый в контексте стационарных черных дыр), который связан с *комплексификацией* пространственно-временного четырехмерного многообразия $\mathcal{M}$, которое увеличивает его до *комплексного* четырехмерного многообразия $\mathcal{CM}$ [Gibbons and Perry, 1978]. Комплексификация как таковая применяется к вещественным многообразиям, определяемым достаточно гладкими уравнениями (технически это *аналитические* уравнения), а процедура комплексификации сводится просто к замене всех координат, выраженных вещественными числами, на координаты, выраженные комплексными числами (см. разделы A.5 и A.9), причем сами уравнения остаются совершенно без изменений. Тогда у нас получается комплексное четырехмерное многообразие (которое в то же время является восьмимерным вещественным; см. раздел A.10). Во всех стандартных решениях эйнштейновских уравнений для стационарной черной дыры, как с космологической постоянной Λ, так и без нее, допускаются такие комплексификации. В результате комплексификации получаем пространство с комплексной периодичностью именно на таком масштабе, который посредством тонких термодинамических принципов [Bloch, 1932] порождает температуру, удивительно точно согласующуюся со значением, которое Хокинг ранее получил для черной дыры (вращающейся или нет). Поэтому особенно соблазнительно предположить, что данное значение температуры

следует аналогичным образом присвоить космологическому горизонту. Когда же данное рассуждение применяется к пустому де-ситтеровскому пространству, мы и в самом деле получаем для космологической постоянной Λ температуру $T_{\text{cosm}} = (2\pi)^{-1}(\Lambda/3)^{1/2}$, согласующуюся с вышеуказанным значением.

Однако головоломка возникает с теми решениями эйнштейновских уравнений (с членом Λ), где присутствует и космологический горизонт событий, и горизонт событий черной дыры, поскольку в таком случае процедура дает две комплексные периодичности сразу, и получается несогласованная интерпретация двух разных температур, существующих одновременно. Несогласованность в данном случае не чисто математическая, так как процедуру комплексификации можно выполнять (что немного неаккуратно) в разных местах и разными способами. Соответственно возможна точка зрения, согласно которой одна температура релевантна окрестности черной дыры, а другая — удаленному наблюдателю. Однако в таком случае данному физическому заключению явно недостает убедительности.

Исходный (более физический по сути) аргумент в пользу интерпретации T_{cosm} как реальной температуры связан с применением к де-ситтеровскому пространству-времени квантовой теории поля в искривленном фоновом пространстве-времени [Davies, 1975; Gibbons and Hawking, 1976]. Правда, выяснилось, что он критически зависит от того, какая именно система координат используется для фона КТП [Shankaranarayanan, 2003, см. также Bojowald, 2011]. Такую неоднозначность можно понять в контексте явления известного под именем *эффект Унру* (или «эффект Фуллинга — Дэвиса — Унру»), который предсказывали Стивен Фуллинг, Пол Дэвис и, в частности, Уильям Унру [Fulling, 1973; Davies, 1975; Unruh, 1976] в середине 1970-х годов. Эффект заключается в том, что наблюдатель, движущийся с ускорением, обнаруживает *температуру*, что объясняется квантовой теорией поля. При обычных ускорениях эта температура должна быть ничтожно малой, и она описывается формулой:

$$T_{\text{accn}} = \frac{\hbar a}{2\pi k c},$$

где a — ускорение, или, в натуральных единицах, еще более простой формулой:

$$T_{\text{accn}} = \frac{a}{2\pi}.$$

В случае с черной дырой, если наблюдатель будет подвешен над ней на тросе, прикрепленном к удаленному неподвижному предмету, он должен почувствовать эту (абсурдно малую) температуру хокинговского излучения, значение которой на горизонте событий составляет $T_{bh} = (8\pi m)^{-1}$. Здесь значение a на горизонте вычисляется как «ньютоновское» ускорение $m(2m)^{-2} = (4m)^{-1}$ в радиусе $2m$ от

горизонта. На горизонте ускорение, испытываемое наблюдателем, строго говоря, должно быть *бесконечным*, но при вычислениях учитывается коэффициент замедления времени, который на горизонте событий также равен бесконечности, поэтому и применяется «ньютоновское» значение, используемое здесь.

С другой стороны, наблюдатель, падающий прямо в дыру, ощутит *нулевую* температуру Унру, поскольку в свободном падении наблюдатель не ощущает ускорения (по принципу эквивалентности Галилея — Эйнштейна, согласно которому наблюдатель, свободно падающий в поле силы тяжести, не почувствует никакого ускорения; см. раздел 4.2). Следовательно, хотя и можно трактовать хокинговскую температуру черной дыры как проявление эффекта Унру, мы убеждаемся, что в свободном падении эта температура может быть обнулена. Применяя же идею в космологическом контексте и пытаясь интерпретировать космологическую температуру T_{cosm} аналогичным образом, мы, опять же, приходим к выводу, что наблюдатель в свободном падении *не* обнаружит эту температуру. Это касается любого *сопутствующего наблюдателя* в стандартных космологических моделях, и в частности в де-ситтеровском пространстве. Поэтому заключаем, что сопутствующий наблюдатель не должен испытывать никакого ускорения и, следовательно, не обнаружит температуры Унру. Соответственно с такой точки зрения подобная низкая «температура» T_{cosm} фактически вообще *не* будет ощущаться сопутствующим наблюдателем!

Следовательно, вот и еще одна причина усомниться в соответствующей «энтропии» S_{cosm}, которой якобы отводится некая динамическая роль в контексте второго закона. Сам я отношусь и к T_{cosm}, и к S_{cosm} с известным скепсисом. Я не хочу сказать, что не признаю за T_{cosm} вообще никакого физического смысла. Полагаю, это вполне может быть некая критическая минимальная температура и, возможно, она играет некоторую роль в контексте идей, изложенных в разделе 4.3.

3.8. Энергия вакуума

В предыдущих главах я говорил, что современные космологи соотносят *темную энергию* (не путать, разумеется, с совершенно другой сущностью, *темной материей*) с эйнштейновской *космологической постоянной* Λ, открытой в 1917 году. Они имеют полное право так поступать, поскольку это согласуется со всеми имеющимися результатами наблюдений. Изначально Эйнштейн использовал этот член в своих уравнениях $\mathbf{G} = 8\pi\gamma\mathbf{T} + \Lambda\mathbf{g}$ (см. раздел 1.1), однако, как оказалось, по неверной причине. Он предложил такую модификацию уравнений с целью получить *стационарную* Вселенную в виде замкнутой трехмерной сферы ($\mathcal{E}$ в разделе 1.15), но спустя всего десять лет Эдвин Хаббл убедительно

продемонстрировал, что Вселенная *расширяется*. Впоследствии Эйнштейн считал введение Λ своей величайшей ошибкой, возможно, потому, что именно из-за нее сам упустил возможность *спрогнозировать* такое расширение! Как пишет космолог Георгий Гамов [1979], Эйнштейн в разговоре с ним как-то заметил: «Космологическая постоянная была величайшей ошибкой моей жизни». Однако сегодня мы понимаем, какая ирония судьбы заключается в таком отношении Эйнштейна к Λ как к ошибке, учитывая, сколь фундаментальная роль отводится Λ в современной космологии. Достаточно вспомнить, что Нобелевская премия по физике за 2011 год была присуждена Солу Перлмуттеру, Брайану П. Шмидту и Адаму Г. Риссу [Perlmutter et al., 1998, 1999; Riess et al., 1998] с формулировкой «За открытие ускоренного расширения Вселенной посредством наблюдения дальних сверхновых». Это расширение наиболее логично объясняется влиянием эйнштейновской Λ.

Тем не менее нельзя полностью исключать возможность того, что космологическое ускорение может быть обусловлено и какой-то другой причиной. Среди физиков, независимо от того, считают ли они величину Λ постоянной (как бы то ни было, ее, пожалуй, все равно можно интерпретировать как эйнштейновскую космологическую постоянную), распространено убеждение, что наличие в эйнштейновском уравнении $G = 8\pi\gamma T + \Lambda g$ члена Λ, вернее, тензора Λg, где g — метрический тензор, объясняется так называемой *энергией вакуума*, которой должно быть пронизано все пустое пространство. Причина, по которой физики полагают, что вакуум обладает ненулевой (положительной) энергией и, следовательно, массой (согласно уравнению Эйнштейна $E = mc^2$), связана с основополагающими соображениями квантовой механики и квантовой теории поля (КТП; см. разделы 1.3–1.5).

В КТП принято разлагать поле на колебательные *моды* (см. разел А.11), каждая из которых обладает собственной определенной энергией. Среди этих различных колебательных мод (каждая мода колеблется с конкретной частотой согласно формуле Планка $E = h\nu$) будет и такая, которая обладает *минимальной* энергией, но оказывается, что эта энергия *ненулевая*. Она называется «энергия нулевых колебаний». Следовательно, даже в вакууме *потенциальное* наличие любого поля должно проявляться в виде минимального уровня энергии. При различных вариантах колебаний получаем разные энергетические минимумы, а их общая сумма (для всех разнообразных полей) и является *энергией вакуума*, то есть энергией, содержащейся в самом вакууме.

Вне гравитационного контекста было бы нормально считать, что такую фоновую энергию вакуума вполне можно игнорировать, поскольку она дает всего лишь универсальную постоянную величину, которую можно вычесть из сум-

мы всех энергетических вкладов, и только *разница* между общим и фоновым энергетическим значением имеет какое-либо значение для (негравитационных) физических явлений. Однако если добавить гравитацию, то все сразу меняется, поскольку эта энергия должна обладать массой ($E = mc^2$), а масса — источник гравитации. На локальном уровне, возможно, все это нормально, если фоновое значение энергии достаточно мало. Пусть такое фоновое гравитационное поле может существенно влиять на космологические расчеты, но оно не должно играть какой-либо заметной роли в локальной физике, поскольку гравитационные воздействия исключительно слабые. Но если суммировать все различные энергии нулевых состояний, то получается неудобный ответ: *бесконечность*, поскольку такое суммирование разных колебательных мод представляет собой *расходящийся ряд* — такие ряды рассмотрены в разделе А.10. Как быть с такой явно катастрофической ситуацией?

Зачастую удается доказать, что расходящийся ряд вида $1 - 4 + 16 - 64 + 256 - \ldots$ имеет конечную «сумму» (в данном случае $1/5$, как показано в разделе А.10), и такой ответ нельзя получить, просто сложив все члены, но данное доказательство можно математически обосновать различными способами, что наиболее важно — прибегнув к *аналитическому продолжению*, кратко упоминаемому в разделе А.10. Аналогичная аргументация позволяет доказать еще более удивительное утверждение: $1 + 2 + 3 + 4 + 5 + 6 + \ldots = -\frac{1}{12}$. При помощи таких средств (и связанных с ними процедур) физики, занимающиеся КТП, часто получают *конечные* значения для подобных расходящихся рядов, а значит, и для вычислений, которые в ином случае дали бы бесполезный ответ ∞. Любопытно, что вторая сумма (всех натуральных чисел) даже играет роль при определении 26-мерности пространства, требуемой в исходной бозонной теории струн (см. раздел 1.6). В данном случае важна как раз *сигнатура* пространства-времени, то есть разность между числом пространственных и временных измерений, а именно $24 = 25 - 1$, и это число 24 соотносится с числом 12 в знаменателе суммы.

Подобные процедуры также применяются при попытках найти конечное решение для проблемы энергии вакуума. Следует отметить, что многие считают энергию вакуума реальным физическим явлением, поскольку ее можно экспериментально наблюдать при так называемом *эффекте Казимира*. Эффект проявляется в виде силы, действующей между двумя параллельными пластинками, изготовленными из проводящего материала, но не имеющими электрического заряда. Если держать две пластинки очень близко друг к другу, но все-таки не допускать контакта, то между ними возникает сила притяжения, хорошо согласующаяся с расчетами, выполненными в 1948 году голландским физиком Хендриком Казимиром. Казимир опирался именно на те эффекты энергии

вакуума, что были описаны ранее. Многократно и успешно ставились эксперименты, подтверждающие и этот эффект, а также его обобщения, которые сделали советский физик Евгений Михайлович Лифшиц и его студенты [Lamoreaux, 1997] (речь именно о том Лифшице, что упоминался в разделах 3.1 и 3.2 в связи с сингулярностями в общей теории относительности).

Правда, никакие из этих эффектов не требуют знать точное значение энергии вакуума, поскольку эффект возникает просто из-за *разницы* фоновой энергии в соответствии с вышеизложенными комментариями. Более того, выдающийся американский специалист по математической физике Роберт Л. Джаффе [Jaffe, 2005] указывал, что силу Казимира можно получить и при помощи стандартных методов КТП, без всякой привязки к энергии вакуума (пусть и более сложным способом). Соответственно, безотносительно проблемы расходимости, возникающей при попытке вычислить точное значение энергии вакуума, экспериментально подтвержденное наличие эффекта Казимира еще *не* гарантирует существования энергии вакуума. Это противоречит распространенному мнению о том, что физическая реальность энергии вакуума надежно доказана.

Тем не менее следует всерьез воспринимать весьма значительную вероятность того, что энергия вакуума вызывает гравитационный эффект (то есть выступает в качестве источника гравитационного поля). Если реальна энергия вакуума, вызывающая гравитацию, то она определенно не может обладать бесконечной плотностью. Если для ее вычисления нужно суммировать все моды колебаний такого рода, что были описаны ранее в этом разделе, то необходимо каким-то способом «отрегулировать» бесконечное значение, к которому, казалось бы, неизбежно приводит нас суммирование мод. Одной из таких процедур является аналитическое продолжение, о котором шла речь ранее. Аналитическое продолжение — устоявшийся мощный математический прием, позволяющий приводить кажущиеся бесконечности к конечным результатам. Однако я уверен, что в актуальном здесь физическом контексте не помешала бы известная осторожность в формулировках.

Давайте вспомним процедуру, которая будет вкратце упомянута в разделе A.10. Аналитическое продолжение связано с так называемыми голоморфными функциями комплексной переменной z, причем термин *голоморфный* означает *гладкий* в контексте комплексных чисел (см. раздел A.10). Существует примечательная теорема (см. раздел A.10), согласно которой любая функция f, голоморфная в некоторых окрестностях начала координат 0 на плоскости Весселя, может быть выражена при помощи степенного ряда:

$$f(z) = a_0 + a_1 z + a_2 z^2 + a_3 z^3 + a_4 z^4 + \cdots,$$

где a_0, a_1, a_2, ... — комплексные константы. Если такой ряд сходится для некоторого ненулевого значения z, то он сойдется и для другого значения z, лежащего ближе к началу координат 0 в плоскости Весселя. В этой плоскости будет существовать фиксированный круг с центром в точке 0, радиус которого $\rho\ (>0)$ будет называться *радиусом сходимости* ряда: ряд сходится при $|z| < \rho$ и расходится при $|z| > \rho$. Если ряд расходится при всех ненулевых z, то говорят, что $\rho = 0$, но этот радиус может быть и бесконечным ($\rho = \infty$). В таком случае функция, описываемая рядом, охватывает *всю* плоскость Весселя и называется *целой функцией*. Методом аналитического продолжения о ней больше ничего не узнать.

Однако если ρ — некое конечное положительное значение, то появляется возможность расширить функцию f при помощи аналитического продолжения. Например, такая ситуация возникает с рядом

$$1 - x^2 + x^4 - x^6 + x^8 - ...,$$

рассматриваемым в разделе A.10, а также в качестве примера в разделе B.11. При комплексификации этот ряд допускает замену вещественной переменной x на комплексную переменную z и суммируется в явную функцию $f(z) = 1/(1 + z^2)$ с радиусом сходимости $\rho = 1$. Однако ряд расходится для $|z| > 1$, хотя сумма этого ряда, а именно $f(z) = 1/(1 + z^2)$, определена во всей плоскости Весселя, за исключением двух *сингулярных* точек $z = \pm i$, когда $1 + z^2$ обнуляется (давая $f = \infty$) (см. рис. A.38 в разделе A.10). Если подставить сюда значение $z = 2$, то получается ответ $1-4+16-64+256-...=\frac{1}{5}$.

В более общей ситуации мы можем и не найти явного выражения для суммы ряда, однако, возможно, нам все-таки удастся продолжить функцию f за пределы круга сходимости, сохранив при этом голоморфную природу f. Один из вариантов решения этой задачи (обычно не самый практичный) — признать, что поскольку f голоморфна везде в пределах своего круга сходимости, можно выбрать любую *другую* точку Q в пределах круга сходимости (комплексное число Q с $|Q| < \rho$), так что f и в этой точке должна быть голоморфной, а далее использовать разложение f в степенной ряд через Q, то есть выражение вида

$$f(z) = a_0 + a_1(z-Q) + a_2(z-Q)^2 + a_3(z-Q)^3 + \cdots,$$

и с его помощью представлять f. Данный случай можно расценивать как стандартное разложение в степенной ряд в окрестностях начала координат $w = 0$ в плоскости Весселя для комплексного числа $w = z - Q$, и эта новая точка начала координат будет совпадать с точкой $z = Q$ в z-плоскости, так что центр круга сходимости теперь будет располагаться в точке Q в плоскости Весселя z. В таком

случае можно распространить определение функции на более обширную область, а затем повторять процедуру для все более крупных областей. Это показано на рис. 3.36 для конкретной функции $f(z) = 1/(1 + z^2)$, и на третьем этапе процедура позволяет продолжить функцию на другую сторону сингулярности в точке $z = \mathrm{i}$ (а именно $z = 6\mathrm{i}/5$), причем центры (значения Q) будут последовательно оказываться в точках 0, $3(1 + \mathrm{i})/5$ и $3(1 + 2\mathrm{i})/5$.

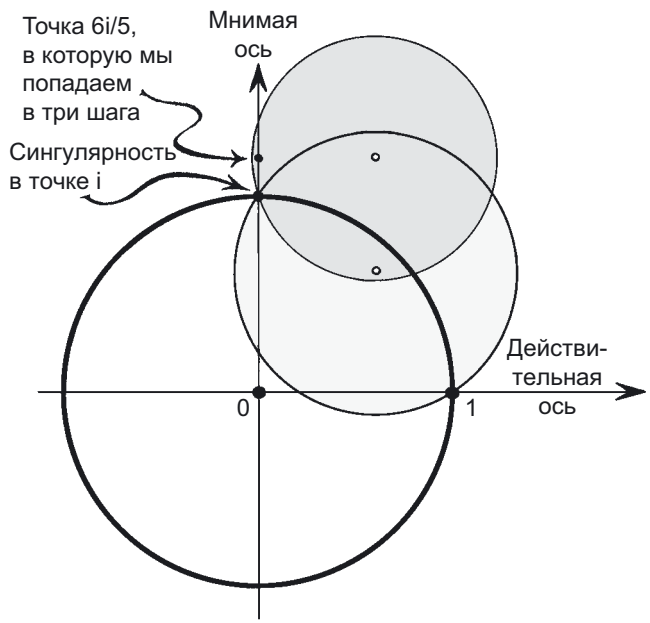


Рис. 3.36. Иллюстрация аналитического продолжения. Круг сходимости для степенного ряда для $f(z) = 1/(1 + z^2)$ — это единичная окружность; ряд расходится за ее пределами, в частности в точке $z = 6\mathrm{i}/5$. Если переместить центр сначала в $z = 3(1 + \mathrm{i})/5$, а затем в $3(1 + 2\mathrm{i})/5$ (малые круги), где круги сходимости определяются сингулярностью в точке $z = \mathrm{i}$, то степенной ряд можно продолжить до точки $z = 6\mathrm{i}/5$

Для функций, содержащих так называемые *ветвящиеся сингулярности*, находим, что аналитическое продолжение приводит нас к неоднозначным ответам, зависящим от того, каким путем мы двинемся по такой ветвящейся сингулярности при продолжении функции. Простейшие примеры такого ветвления возникают с дробными степенями, например $(1 - z)^{-1/2}$ (что в данном случае давало бы разный знак в зависимости от того, каким путем мы направимся по ветвящейся сингулярности в точке $z = 1$) или $\log(1 + z)$. Так мы получим ветвление в точке $z = -1$, что приведет нас к дополнительным неоднозначностям целочисленных

кратных $2\pi i$, зависящим от того, сколько раз мы обошли вокруг ветвящейся сингулярности в точке $z = -1$. Не считая неоднозначностей, возникающих из-за такого ветвления (а это непростая тема), аналитическое продолжение в известном смысле всегда *уникально*.

Эта уникальность, однако, — довольно тонкий момент, и удобнее всего понимать ее в контексте римановых поверхностей, о которых пойдет речь в разделе А.10. В принципе, процедура призвана «выявить» все ветвления путем замены плоскости Весселя ее многоуровневой версией, которую затем можно повторно интерпретировать как некую риманову поверхность, на которой каждое из многих продолжений функции f получает единственное значение. Тогда аналитически продолженная f становится совершенно уникальной на данной римановой поверхности [Miranda, 1995]; кратко познакомиться с этими идеями можно в книге ПкР, разделы 8.1–8.3. Тем не менее надо еще привыкнуть к таким странным неоднозначным «суммам», особенно к тем, что могут возникать из совершенно обычных на первый взгляд (не всегда расходящихся) рядов. Например, в случае с функцией $(1 - z)^{-1/2}$, рассмотренной ранее, значение $z = 2$ дает удивительно неоднозначную расходящуюся сумму:

$$1+\frac{1}{1}+\frac{1\times 3}{1\times 2}+\frac{1\times 3\times 5}{1\times 2\times 3}+\frac{1\times 3\times 5\times 7}{1\times 2\times 3\times 4}+\frac{1\times 3\times 5\times 7\times 9}{1\times 2\times 3\times 4\times 5}+\ldots=\pm i,$$

а при $\log(1 + z)$ значение $z = 1$ дает *сходящуюся*, но в соответствии с вышеуказанной процедурой гораздо *более* неоднозначную сумму:

$$1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\frac{1}{5}-\frac{1}{6}+\ldots=\log 2+2n\pi i,$$

где на месте n может стоять вообще любое целое число. Если всерьез воспринимать такие ответы при решении реальной физической задачи, мы опасно приблизимся к чистой *фантазии*, и чтобы такая теория приобрела физическое правдоподобие, для нее потребуется четкое теоретическое обоснование.

Тонкости такого рода подсказывают, что нужно крайне осторожно пользоваться аналитическим продолжением, если нам требуются физические ответы на вопросы, требующие правильного суммирования расходящихся рядов. Тем не менее только что обозначенные проблемы обычно вполне понятны физикам-теоретикам, работающим с такими вопросами (в частности, теоретикам-струнникам и специалистам в смежных областях). Однако не на этих проблемах я хочу заострить здесь внимание. Мой вопрос скорее связан с индивидуальной верой в конкретные коэффициенты a_0, a_1, a_2, a_3 в некоторых рядах $a_0 + a_1z + a_2z^2 + a_3z^3 + \ldots$. Если эти числа — плод фундаментальной теории, которая должна в точности их определять, то вышеописанные процедуры вполне могут иметь

физический смысл в подходящих ситуациях. Однако если это результат расчетов, содержащих приближения, неопределенности или сложные зависимости от внешних условий, то нужно гораздо осторожнее относиться к результатам, получаемым путем аналитического продолжения (как мне кажется, это касается и любых других операций, связанных с суммированием безнадежно расходящихся бесконечных рядов).

В качестве иллюстрации давайте вернемся к ряду $1 - z^2 + z^4 - z^6 + z^8 \ldots$ ($= 1 + 0z - z^2 + 0z^3 + z^4 + 0z^5 - \ldots$) и также вспомним, что он имеет радиус сходимости $\rho = 1$. Если предположить, что мы случайным образом варьируем коэффициенты $(1, 0, -1, 0, 1, 0, -1, 0, 1, \ldots)$ этого ряда — но лишь самую малость, чтобы ряд попрежнему сходился в пределах единичной окружности (обеспечить это можно, ограничив величину вариации коэффициентов некоторым числом), — то у нас практически гарантированно получится ряд для голоморфной функции, имеющей естественную границу на единичной окружности [Littlewood and Offord, 1948; Eremenko and Ostrovskii, 2007] (на рис. 3.37 показана конкретная функция именно такого рода). Здесь мы имеем дело со следующим свойством единичной окружности: возмущенная функция становится *настолько* сингулярной, что невозможно никакое аналитическое продолжение этой функции вдоль окружности. Следовательно, совершенно ясно, что из всех голоморфных функций, определяемых в (открытом) единичном круге (то есть $|z| < 1$), те функции, которые можно голоморфно продолжить *хотя бы где-нибудь* в этом круге, составляют исчезающе малую долю. Таким образом, строго говоря, нам еще очень повезет, если удастся суммировать расходящиеся ряды методом аналитического продолжения при возмущении этих рядов в обычных физических ситуациях.

Это не говорит о том, что такие процедуры регуляризации бесконечностей обязательно будут бессмысленны в общих физических ситуациях. Энергия вакуума допускает, что существует определяемый расходящимися рядами «фон», который можно результативно просуммировать до конечного значения при помощи процедур аналитического продолжения; сюда также можно добавить сильно *сходящийся* путь, который можно просуммировать отдельно обычным способом. Например, если приплюсовать к нашему вышеуказанному ряду для $f(z) = 1/(1 + z^2)$, определяемому коэффициентами $(1, 0, -1, 0, 1, 0, -1, 0, 1, \ldots)$ другую «небольшую» часть, определяемую коэффициентами $(\varepsilon_0, \varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \ldots)$, для которых ряд $\varepsilon_0 + \varepsilon_1 z + \varepsilon_2 z^2 + \varepsilon_3 z^3 + \varepsilon_4 z^4 + \ldots$ суммируется в *целую* функцию ($\rho = \infty$), то аналитическое расширение за пределы единичного круга вполне может работать так, как работало и без возмущения, и аргументацию с аналитическим продолжением можно развивать как раньше. Однако все эти комментарии приведены здесь для того, чтобы предупредить читателя, как много подводных камней и неочевидных проблем встречается при применении таких процедур

для суммирования расходящихся выражений. В частности, в некоторых ситуациях такой метод оказывается применим, но физические выводы на основе его результатов следует делать с большой осторожностью.

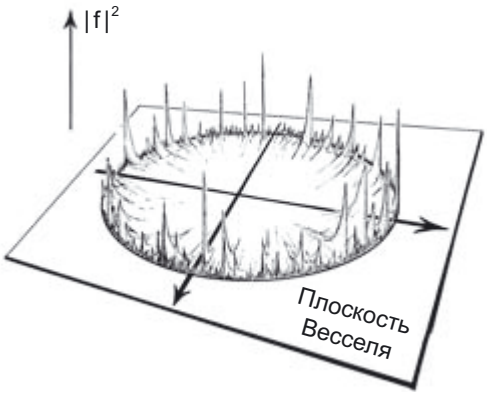


Рис. 3.37. Схематичная иллюстрация своеобразного препятствия для голоморфного продолжения: *естественная граница* голоморфной функции *f*. В любой точке единичной окружности продолжить функцию *f* за пределы этой окружности невозможно, пусть даже *f* голоморфна повсюду внутри этого круга (показан график функции $|f|$)

Теперь давайте вернемся к самой проблеме энергии вакуума и попробуем определить, можно ли трактовать эйнштейновскую Λ как меру энергии, содержащейся в пустом пространстве-времени. В частности, с учетом того множества полей, которые потенциально вносят энергетический вклад в величину Λ, такое вычисление сразу представляется невыполнимой задачей. Тем не менее о некоторых вещах явно можно утверждать с достаточной уверенностью, почти (или вообще) не вдаваясь в детали. Прежде всего, исходя из *локальной лоренц-инвариантности* — в сущности, это тезис, согласно которому не должно быть никаких «предпочтительных» направлений в пространстве-времени, — можно строго утверждать, что тензор энергии-импульса вакуума $\mathbf{T}_{\text{vac}}$ должен быть пропорционален метрическому тензору $\mathbf{g}$:

$$\mathbf{T}_{\text{vac}} = \lambda \mathbf{g},$$

где λ — некая константа. Есть надежда, что найдется аргумент, который покажет, что $\mathbf{T}_{\text{vac}}$ действительно вносит определенный вклад в эйнштейновские уравнения, соответствующий *наблюдаемому* значению космологической постоянной:

$$\lambda = \frac{\Lambda}{8\pi}$$

в естественных единицах. Таким образом, мы получим правильный вклад в правой части уравнений Эйнштейна, ведь $\mathbf{G} = 8\pi\gamma\mathbf{T} + \Lambda\mathbf{g}$ в таком случае (в естественных единицах) теперь можно записать в виде:

$$\mathbf{G} = 8\pi(\mathbf{T} + \mathbf{T}_{\text{vac}}).$$

Однако исходя из соображений КТП, можно прийти, в принципе, только к двум выводам: либо $\lambda = \infty$, и это наиболее «честный», но почти бесполезный ответ, если, конечно, не прибегать к математическим уловкам вроде тех, что описаны ранее; либо $\lambda = 0$, и такая точка зрения была наиболее популярна, пока из наблюдений не стало ясно, что космологическая постоянная должна иметь положительное значение или должно присутствовать что-то еще порядка *единицы*, выраженной в естественных величинах (плюс, возможно, несколько простых степеней π вдобавок). Естественно, этот последний ответ всех бы устроил, если бы не один факт: по данным наблюдений,

$$\Lambda \approx 6 \times 10^{-122}$$

в естественных единицах, а это нестыковка поистине фантастического масштаба!

На мой взгляд, это заставляет всерьез усомниться в интерпретации Λ как меры энергии вакуума. Тем не менее большинство физиков со странным упорством цепляются за такую интерпретацию. Разумеется, если Λ не является энергией вакуума, требуется какое-то теоретическое объяснение, почему член Λ в уравнениях Эйнштейна положителен (особенно с учетом факта, отмеченного в разделе 1.15: специалисты по теории струн заявляют, что теоретически более убедительна отрицательная Λ). В любом случае, Λ — диковинная энергия, поскольку вызывает гравитационное *отталкивание*, несмотря на то что ее значение по сути *положительно*. Все дело в очень любопытном устройстве тензора энергии-импульса, то есть $\lambda\mathbf{g}$, который отнюдь не похож на тензор энергии-импульса какого-либо иного известного (или всерьез рассматриваемого) физического поля. Об этом уже упоминалось ранее, в разделе 3.7.

Если читатель озадачился тем, каким образом положительная Λ может воздействовать на кривизну пространства как положительная масса, но при этом порождать силу отталкивания, лежащую в основе ускоряющегося расширения Вселенной, отсылаю его к концу раздела 3.8. $\Lambda\mathbf{g}$ не является тензором энергии-импульса в физическом смысле, несмотря на термин «темная энергия». В частности, в $\Lambda\mathbf{g}$ присутствуют три отрицательные компоненты давления, обеспечивающие отталкивание, и один положительный компонент плотности, обеспечивающий кривизну пространства.

Более того, именно таким своеобразием тензора энергии-импульса объясняется и любопытный «парадокс»: часто озвучивается тезис, согласно которому «более 68 % материального содержимого Вселенной приходится на неизвестную *темную энергию*». В отличие от всех остальных форм массы-энергии, Λ вызывает гравитационное *отталкивание*, а не притяжение, поэтому в данном отношении она действует совершенно противоположным образом, нежели обычная материя. Более того, если Λ не изменяется со временем (если речь *действительно* идет об эйнштейновской космологической *постоянной* Λ), то значение 68 % будет постепенно увеличиваться по мере расширения Вселенной, а средние *плотности* всех «обычных» форм материи (в том числе темной материи) со временем неизбежно уменьшатся до *совершенно незаметных* значений!

Последний аспект поднимает еще одну проблему, которая изрядно беспокоит космологов: только на нынешнем этапе истории Вселенной (понятие «нынешний» трактуется здесь очень широко, а именно в период 10^9–10^{12} лет после Большого взрыва) «плотность энергии», обеспечиваемая Λ, сравнима с плотностью обычной материи (в том числе темной, что бы она собой ни представляла). На гораздо более ранних этапах истории Вселенной (скажем, $<10^9$ лет) вклад Λ был незначителен, а в далеком будущем (допустим, $>10^{12}$ лет) энергия Λ будет полностью доминировать над всеми остальными формами. Что это — примечательное совпадение, нечто очень странное, требующее объяснения? По-видимому, многие космологи так считают, а некоторые предпочли бы, чтобы «Λ» была каким-то эволюционирующим полем, которое часто именуется *квинтэссенцией*. Я вернусь к этим проблемам в следующих двух разделах, но сам считаю, что совершенно ошибочно воспринимать темную энергию как некую материальную субстанцию либо даже как энергию вакуума. Вполне возможно, требуется разгадать какую-то тайну, связанную с истинным смыслом Λ (см., например, раздел 3.10), но при этом необходимо помнить, что эйнштейновский член Λ — практически единственная модификация его исходных уравнений ($\mathbf{G} = 8\pi\mathbf{T}$), не требующая радикального изменения самых основ его величественной теории. Не вижу причин, почему бы природе не воспользоваться этой шикарной возможностью!

3.9. Инфляционная космология

Давайте обсудим некоторые причины, по которым большинство космологов с таким рвением отстаивают, казалось бы, фантастическую гипотезу *инфляционной космологии*. Что это за гипотеза? Это удивительная картина, которую независимо друг от друга впервые сформулировали около 1980 года советский ученый Алексей Старобинский (хотя в несколько ином контексте) и американ-

ский исследователь Алан Гут. Согласно их идеям, наша Вселенная практически сразу после Большого взрыва — в исчезающе краткий период от 10^{-36} до 10^{-32} секунд *после* этого мгновенного события — пустилась в *экспоненциальное расширение*, именуемое *инфляцией*. Этот эффект мог объясняться воздействием колоссальной космологической постоянной Λ_{infl}, неизмеримо превосходившей современное значение Λ — примерно в 10^{100}, согласно выражению:

$$\Lambda_{infl} \approx 10^{100} \Lambda$$

(есть и много других вариантов инфляционной теории, дающих несколько иные значения). Следует отметить, что такая Λ_{infl} все равно крайне мала (порядка 10^{-21}) по сравнению со значением ~10^{121} Λ, которое выводится из расчетов, связанных с энергией вакуума. В популярной форме эти проблемы изложены в книге Гута [1997]; более научные трактовки содержаться в работах [Blau and Guth, 1987; Liddle and Lyth, 2000 и Muckhanov, 2005].

Прежде чем рассмотреть причины, по которым большинство космологов сегодня всерьез воспринимают эту поразительную гипотезу (инфляция сегодня фигурирует во всех серьезных описаниях современной космологии, как научно-технических, так и популярных), еще раз предупреждаю читателя: в названии этой главы, да и в названии всей книги есть слово «фантазия», которое я употребил потому, что собирался описывать как раз эту гипотезу. В разделе 3.11 мы узнаем, что сегодня в космологии обсуждаются и *другие* идеи, которые можно считать гораздо более *фантастичными*, чем инфляционная. Но что особенно интересно в инфляционной теории — так это то, что она пользуется практически всеобщим признанием в кругах космологов!

Правда, следует отметить, что инфляция — это не единая общепринятая гипотеза. Существует множество разных картин мира, укладывающихся в общую канву инфляции, и часто ставятся эксперименты, призванные как раз отличить одну такую картину от другой. В частности, открытие *B*-мод в поляризации реликтового микроволнового излучения , о чем во всеуслышание объявила команда телескопа BICEP2 в марте 2014 года [Ade et al., 2014], считается сильным доказательством в пользу широкого класса инфляционных моделей, даже неопровержимым свидетельством; оно было сделано в ходе исследования, в основном призванного различить разные варианты инфляции. Считается, что наличие таких *B*-мод — признак существования первичных гравитационных волн, которые были предсказаны в нескольких версиях инфляционной теории (см. также конец раздела 4.3 — там рассматривается альтернативная причина, по которой могут возникать *B*-моды). На момент написания книги интерпретация этих сигналов остается крайне противоречивой, предлагаются и другие варианты объяснения. Тем не менее сегодня практически никто из космологов

не сомневается в том, что фантастическая в целом идея инфляции содержит некое зерно истины, касающееся очень разных этапов расширения Вселенной.

Однако, как я уже объяснял ранее (в предисловии и в разделе 3.1), я *не имею в виду*, что из-за такой фантастичности инфляция не заслуживает серьезного рассмотрения. На самом деле теория выдвигалась в попытке объяснить очень примечательные — даже «фантастические» — *наблюдаемые* свойства нашей реальной Вселенной. Более того, нынешнее всеобщее признание инфляции связано с тем, как хорошо она объясняет, казалось бы, несвязанные и ранее не разгаданные удивительные свойства Вселенной. Кроме того, важно отметить, что если инфляционная гипотеза в нашей Вселенной неверна (чуть позже я приведу аргументы против такой теории), то должно подтвердиться какое-то другое объяснение — возможно, столь же экзотические идеи, которые кажутся чистой фантазией!

Выдвигалось множество различных версий этой теории, и у меня не хватает ни знаний, ни смелости, чтобы описывать в этом разделе какие-либо из них, кроме самой первой, которая сейчас популярна (см., однако, раздел 3.11, где я вкратце упомяну некоторые более экстравагантные версии космической инфляции). В исходной картине источником космической инфляции считалось первичное состояние «ложного вакуума», который изменился путем фазового перехода, подобно превращению жидкости в газ при изменении условий окружающей среды, — и Вселенная квантово-механически *туннелировала*, перейдя в иное вакуумное состояние. В таких разных вакуумах может различаться и значение члена Λ в эйнштейновских уравнениях.

Ранее в этой книге я не затрагивал хорошо известное явление *квантового туннелирования*. Как правило, оно возникает в квантовой системе, где два энергетических минимума **A** и **B** разделены *энергетическим барьером*, и система, исходно обладавшая более высокоэнергетическим состоянием **A**, может спонтанно туннелировать в **B**, не обладая при этом достаточной энергией для преодоления такого барьера. Не буду здесь знакомить читателя с деталями этой процедуры, но, думаю, должен отметить, что применимость этого описания в данном контексте вызывает много вопросов.

Однако в разделе 1.16 я обращал внимание читателей на проблему выбора вакуумного состояния — это неотъемлемая составляющая описания КТП. Можно иметь две различные версии КТП, которые были бы идентичны (то есть в них бы совпадали алгебры операторов рождения и уничтожения и т. д.), если бы не тот факт, что в каждом из случаев указываются разные вакуумы. В разделе 1.16 эта проблема была ключевой, поскольку речь шла об излишнем изобилии различных струнных (или M-) теорий, образующих так называемый ландшафт. Проблема, обозначенная в разделе 1.16, заключалась в том, что каждая КТП в этом

огромном ландшафте должна соответствовать совершенно отдельной вселенной, и физический переход из одной такой вселенной в другую был бы невозможен. Процесс, разворачивающийся при квантовом туннелировании, — совершенно нормальное квантово-механическое явление, однако обычно *не* допускается, что оно могло бы обеспечить переход из состояния в «одной» вселенной в состояние в другой, если в этих вселенных различаются вакуумы.

Тем не менее идея о том, что космическая инфляция могла начаться из-за туннелирования из состояния ложного вакуума (со своими свойствами и со значением энергии вакуума Λ_{infl}) в другой вакуум, обладающий наблюдаемой ныне космологической постоянной Λ, по-прежнему популярна в определенных кругах космологов [Coleman, 1977; Coleman and De Luccia, 1980]. Что касается моих собственных взглядов, я должен напомнить читателю о сложностях, упомянутых в разделе 3.8: я скептически отношусь даже к идее о том, что космологическая постоянная якобы представляет собой проявление энергии вакуума. Разумеется, в контексте фантастических идей, которые могут понадобиться для разгадки происхождения нашей Вселенной с ее многочисленными примечательными свойствами, которые, как кажется, проявлялись на ранних этапах (см., например, раздел 3.4), определенно нельзя игнорировать идеи такого рода, не стесненные рамками традиционной физики. Однако я сторонник особенно осторожного обращения с идеями, которые, насколько это известно, выходят за рамки устоявшихся процедур.

Мы еще вернемся к этой проблеме в разделе 3.11, а в этом разделе я не буду много говорить об изначальной версии, согласно которой инфляция началась из-за туннелирования из одного вакуума в другой, в особенности потому, что эта версия, по-видимому, не пользуется широкой поддержкой, учитывая, как сложен с теоретической точки зрения «изящный выход» из инфляционного этапа (примерно к 10^{-32} с), присутствующий в исходной модели. В этот момент инфляция должна была прекратиться сразу и везде, после чего должен был произойти так называемый *вторичный разогрев*. Чтобы обойти эти сложности, в последующих моделях рассматривалась инфляция по принципу *медленного скатывания*, независимо предложенная Андреем Линде [1982], а также Андреасом Альбрехтом и Полом Стейнхардтом. Большинство моих комментариев касаются идей именно такого рода. При инфляции по принципу медленного скатывания рассматривается скалярное поле φ, именуемое *инфляционным полем* (хотя в некоторых более ранних публикациях φ называлось *полем Хиггса* — неподходящий термин, который сейчас уже не используется). Считалось, что именно оно вызвало инфляционное расширение новорожденной Вселенной.

Термин *медленное скатывание* характеризует особенность *графика* потенциальной функции $V(\varphi)$, описывающей *энергию* φ-поля (рис. 3.38), причем состояние

Вселенной представлено на графике как точка, скатывающаяся по кривой. В различных версиях инфляции «медленного скатывания» (их много) эта потенциальная функция строится искусственно, а не выводится из более общих принципов; значит, при скатывании она должна задавать конкретные свойства, нужные для того, чтобы инфляция работала. Как в традиционной физике частиц, так и, насколько я могу судить, в других областях физики не существует никакого обоснования вида кривой $V(\varphi)$. Та часть кривой, где происходит медленное скатывание, добавляется к графику, для того чтобы Вселенная могла расширяться настолько долго, насколько это необходимо, после чего кривая приближается к минимуму, обеспечивая довольно равномерное затухание инфляции. По мере того как энергетический потенциал $V(\varphi)$ стремится к стабильному минимуму, инфляция завершается. Разные авторы (см., например, [Liddle and Leach, 2003; Antusch and Nolde, 2014; Martin et al., 2013; Byrnes et al., 2008]) предлагают множество различных вариантов $V(\varphi)$, и произвольность этой процедуры, пожалуй, свидетельствует о слабости инфляционной гипотезы.

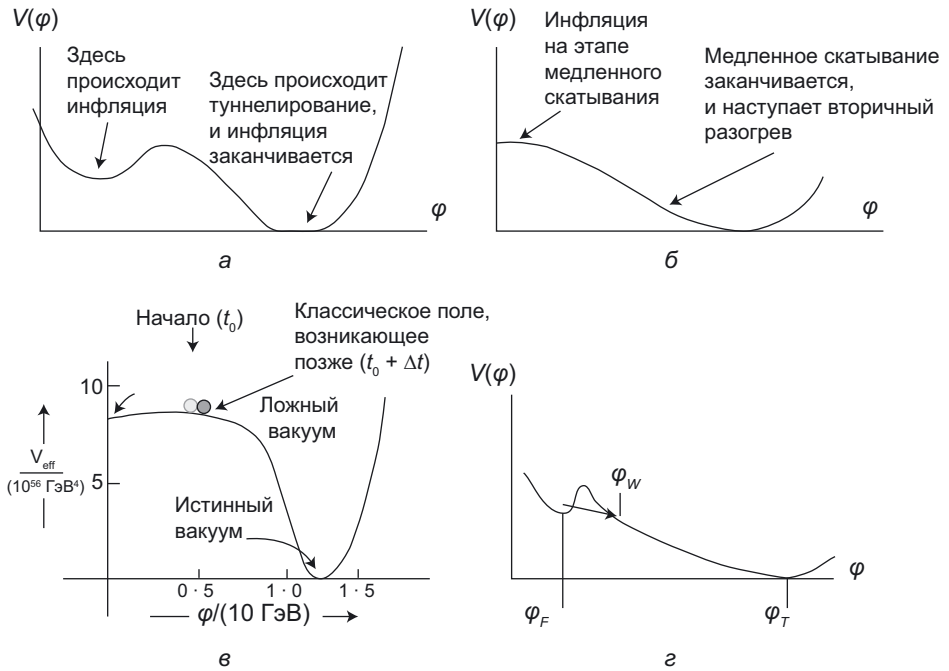


Рис. 3.38. Некоторые из многочисленных вариантов инфляционной потенциальной функции φ, подобранной так, чтобы она отражала все желаемые свойства инфляции. Формы кривой φ значительно различаются, что выдает отсутствие базовой теории, которая бы описывала (инфляционное) φ-поле

Тем не менее такие идеи никогда бы не возникли, не будь на то серьезной мотивации. Поэтому давайте начнем с тех загадок, ради решения которых около 1980 года изначально возникла инфляционная модель. Одной из них было неприятное следствие различных популярных теорий великого объединения (ТВО) из физики частиц. Некоторые ТВО прогнозируют существование магнитных монополей [Wen and Witten, 1985; Langacker and Pi, 1980], то есть *отдельных* самостоятельных северных и южных магнитных полюсов. В традиционной физике (а также на уровне наблюдений — пока, во всяком случае) непарные магнитные полюса нигде не встречаются, а всегда входят в состав *диполей*, то есть двухполюсных обычных магнитов — с северным и южным полюсами. Если сломать магнит, разделив полюсы, то на месте излома образуется новая пара полюсов: новый южный там, где с другой стороны остался северный, а на другой половине — новый северный полюс (рис. 3.39). Фактически известные нам магнитные полюсы — это побочные продукты циркуляции внутренних электрических токов. Отдельные частицы также могут проявлять свойства магнитов (диполей), но никогда не наблюдались в виде монополей, то есть *автономных* северных и южных магнитных полюсов.

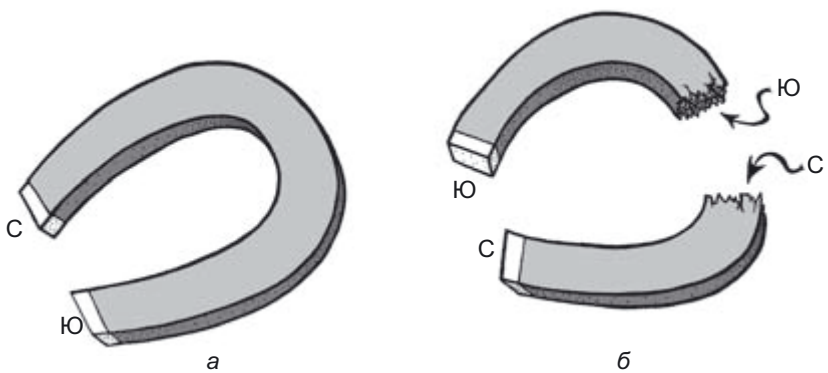


Рис. 3.39. Отдельные магнитные монополи — изолированные северные и южные полюсы — не встречаются в традиционной физике. У обычного магнита с одной стороны северный полюс, с другой — южный. Если сломать магнит, то на противоположных концах половинок возникнут новые полюсы, и магнитный «заряд» каждой из них останется сбалансированным — нулевым

Однако возможность существования таких монополей отстаивали и упорно продолжают отстаивать некоторые физики и, что особенно важно, виднейшие специалисты по теории струн, о чем свидетельствует ремарка, сделанная Джозефом Полчински в 2003 году (см. [Polchinski, 2004]):

> «Если бы мне предложили поспорить о том, что еще предстоит открыть в физике, то ставка на монополи была бы, считай, беспроигрышной».

Согласно таким ТВО, новорожденная Вселенная должна была изобиловать подобными теоретически предсказанными магнитными монополями, но их никто никогда не наблюдал. Нет и прямых эмпирических доказательств того, что они когда-то имелись во Вселенной. Инфляционная идея как раз и была предложена, чтобы уйти от катастрофического противоречия с наблюдениями. Считается, что из-за первичного экспоненциального расширения магнитные монополи, некогда доминировавшие во Вселенной, должны были рассеяться в ней настолько далеко друг от друга, что мы теперь не можем их обнаружить.

Разумеется, такая мотивация сама по себе не показалась бы весомой многим теоретикам (и мне в том числе), поскольку решение могло быть гораздо проще: возможно, несоответствие означает, что все имеющиеся ТВО, описывающие Вселенную, неверны — и неважно, сколь привлекательной считают ту или иную теорию ее сторонники. Тем не менее в данном случае мы приходим к выводу о том, что без учета инфляции многие современные идеи о природе фундаментальной физики, и особенно различные версии теории струн, столкнулись бы с другой серьезной проблемой. Правда, проблема магнитных монополей нечасто фигурирует в современных тезисах о безальтернативности инфляции, обычно основа этих тезисов иная. Давайте далее их и обсудим.

Другие аргументы, на первых порах выдвигавшиеся в качестве ключевой мотивации в пользу космической инфляции, тесно связаны с вопросами, которые я поднимал в разделе 3.6, а именно с исключительной *однородностью* распределения материи в ранней Вселенной. Однако существует принципиальное различие между моей аргументацией и аргументацией инфляционистов. Я заострил внимание преимущественно на загадке второго закона термодинамики (см. разделы 3.3, 3.4 и 3.6) и на том, насколько несимметричным образом объясняется исключительно низкая первичная энтропия, запечатленная в РИ. По-видимому, в тот период из всех степеней свободы уникальным образом выделились лишь *гравитационные*, тогда как другие возможные степени свободы могли быть подавлены (см. конец раздела 3.4). В то же время поборники инфляции уделяют внимание лишь некоторым частным аспектам этой великой загадки, и хотя эти аспекты тесно связаны с вопросами, затронутыми в разделах 3.4 и 3.6, инфляционисты рассматривают их под совершенно другим углом.

Как будет рассказано далее, имеются и другие наблюдательные данные, требующие относиться к идее инфляции со всей серьезностью: для объяснения этих наблюдений необходимы некоторые весьма экзотические предположения об

устройстве реального мира. Но эти данные отсутствовали в исходных обоснованиях инфляционной модели. На самом раннем этапе можно было выделить всего три особенно загадочных космологических факта, известных из наблюдений. Они назывались *проблема горизонта*, *проблема гладкости* и *проблема плоскостности*. Было распространено мнение, и инфляционисты часто выдавали это за свой успех,согласно которому инфляционная модель действительно позволяет решить все три эти проблемы. Но так ли это на самом деле?

Начнем с *проблемы горизонта*. В данном случае нас интересует факт (уже отмеченный в разделе 3.4), что температура РИ, отовсюду льющегося на нас с неба, везде практически одинакова — различия встречаются всего по несколько раз на 10^5, если сделать поправку на доплеровский эффект, обусловленный вращением Земли относительно этого излучения. Что касается однородности, особенно с учетом исключительно *тепловой* природы этого излучения (см. рис. 3.13 в разделе 3.4), вполне возможно, что весь тот огненный шар, который представляла собой Вселенная, возник в результате сильнейшего первичного разогрева, который и привел Вселенную в состояние теплового расширения (с максимальной энтропией).

Рис. 3.40. В стандартной космологической картине без учета инфляции 3-поверхность Большого взрыва $\mathcal{B}$ на конформной диаграмме должна лишь немного предшествовать 3-поверхности эпохи последнего рассеяния $\mathcal{D}$ (как показано на рисунке). Соответственно, два события Q и R на $\mathcal{D}$, которые визуально разделены друг от друга углом чуть более 2° (если наблюдать из той точки, которую занимаем в пространстве-времени мы), никогда не могли вступить в причинно-следственный контакт, поскольку их световые конусы прошлого не пересекаются в прошлом вплоть до самого момента $\mathcal{B}$

Однако в качестве одной из сложностей с этой картиной указывали следующий факт: в соответствии со скоростью расширения в стандартных космологических

моделях Фридмана/Толмана (см. раздел 3.1), существенно отдаленные друг от друга события на 3-поверхности *последнего рассеяния* $\mathcal{D}$ (где фактически и возникало излучение, см. раздел 3.4) лежали бы настолько далеко за пределами горизонтов частицы друг друга, что оказались бы причинно независимыми. Это произошло бы даже с точками P и Q на $\mathcal{D}$, которые с нашей точки зрения отстоят друг от друга на небе всего на 2°, как показано на схематичной конформной диаграмме (рис. 3.40). Согласно космологической картине, такие точки не могли находиться ни в какой причинно-следственной связи, поскольку их световые конусы прошлого оказались бы совершенно разделены еще на момент Большого взрыва (3-поверхность $\mathcal{B}$). Соответственно, они не могли взаимодействовать и в момент разогрева, при котором температуры в точках P и Q могли бы выровняться.

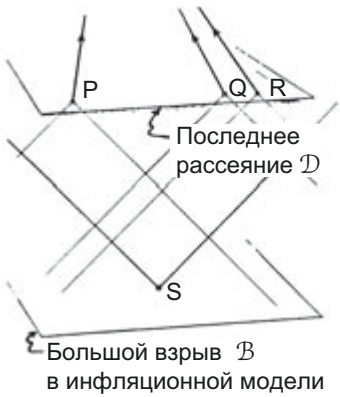


Рис. 3.41. Если в моделях учитывается инфляция (см. рис. 3.40), то 3-поверхность Большого взрыва $\mathcal{B}$ в конформной картине гораздо сильнее смещается вниз (на самом деле гораздо ниже, чем показано здесь). В результате световые конусы прошлого точек P и Q всегда пересекаются еще до достижения $\mathcal{B}$ независимо от того, как велико угловое расстояние между ними из нашей точки наблюдения (см., например, P и R на этой иллюстрации)

Эта проблема озадачивала космологов, но после того как Гут (и Старобинский) предложили свою выдающуюся идею *космической инфляции*, появилась возможная аргументация, казалось бы, позволявшая разрешить этот парадокс. После того как стал учитываться этап первичной инфляции, между 3-поверхностями $\mathcal{B}$ и $\mathcal{D}$ на конформной диаграмме возник существенный разрыв, так что любые две точки P и Q на поверхности последнего рассеяния $\mathcal{D}$, видимые для нас в небе на уровне РИ (даже диаметрально противоположные из нашей точки наблюдения), должны обладать световыми конусами, которые пересекаются на

пути к 3-поверхности Большого взрыва $\mathcal{B}$ (рис. 3.41). Широкий зазор между B и D — в действительности он гораздо больше, чем показано на рис. 3.41, поскольку порядок инфляционного расширения составляет 10^{26}, — является частью де-ситтеровского пространства-времени (см. раздел 3.1). На рис. 3.42 показано «вырезание и склеивание» инфляционной модели, которое, надеюсь, помогает интуитивно понять, почему Большой взрыв именно таким образом сдвигается на конформной диаграмме. Следовательно, при инфляции нашлось бы подходящее время для полного разогрева. При таком решении проблемы горизонта появляется вполне достаточный период для полноценного взаимодействия в пределах всего первичного огненного шара и в рамках нашего горизонта частицы. В этот период достигается тепловое равновесие (при расширении), и поэтому температура РИ повсюду становится практически одинаковой.

Прежде чем перейти к контраргументам, которые я считаю фундаментальными, было бы полезно поговорить о второй проблеме, которую якобы решает инфляционная теория, а именно о *проблеме гладкости*. Эта проблема связана с более или менее однородным распределением материи и структуры пространства-времени, которое, по-видимому, наблюдается повсюду во Вселенной (наличие войдов и т. д. считается относительно незначительными отклонениями от этой однородности). В данном случае утверждается, что (примерно) 10^{26}-кратное экспоненциальное расширение само по себе могло разгладить любые заметные неровности, которые могли присутствовать в изначальном (предположительно, очень неупорядоченном) состоянии Вселенной. Идея заключалась в том, что любые неоднородности, которые могли существовать вначале, растянулись с огромным линейным коэффициентом (допустим, $\times 10^{26}$), и в результате пространство стало таким гладким, каким мы его и наблюдаем.

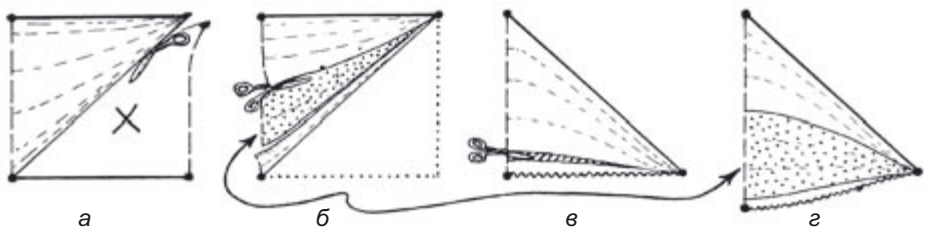


Рис. 3.42. Как построить модель инфляционной космологии в категориях строгих конформных диаграмм: *а*) берем стационарную часть (см. рис. 3.26 *б*) де-ситтеровского пространства (см. рис. 3.26 *а*); *б*) вырезаем из нее очень долгий временной фрагмент (обозначен точками); *в*) убираем из фридмановской модели $K = 0$ очень краткий ранний временной отрезок (допустим, здесь $\Lambda = 0$) (см. рис. 3.26 *б*); *г*) склеиваем с ним отрезок времени, на котором Вселенная является стационарной

Оба этих аргумента — попытки объяснить однородность Вселенной, апеллируя к этапу колоссального расширения в самом начале ее существования. Однако я берусь утверждать, что оба они принципиально ошибочны (см. также [Penrose, 1990]; ПкР, глава 28). Наиболее серьезная причина этого связана с фактом, уже рассмотренным в разделе 3.5, — исключительно низкой энтропией в нашей Вселенной (что очень важно в аспекте имеющегося у нас второго закона термодинамики), которая и выражается как раз в этой однородности. По-видимому, вся идея, лежащая в основе инфляционного аргумента, заключается в том, что наша Вселенная возникает из сколь угодно неупорядоченного исходного состояния (поэтому обладает максимальной энтропией), и из него мы приходим к состоянию исключительно однородному и, следовательно, *низкоэнтропийному в отношении гравитации*, которое отражается в структуре РИ и в значительной однородности, присущей современной Вселенной (см. конец раздела 3.4). Основная проблема — второй закон термодинамики и откуда он взялся. Маловероятно, что он мог возникнуть в результате обычной физической эволюции, определяемой обратимыми во времени динамическими уравнениями, где все начинается со сравнительно *случайной*, то есть *высоко*энтропийной, исходной точки.

Важно отметить, что все динамические процессы, лежащие в основе инфляции, симметричны относительно обращения времени. При этом мы еще не обсуждали уравнения такого типа, что используются при описании инфляции по принципу медленного скатывания. В данном случае требуются различные составляющие, важнейшая из которых — скалярное инфляционное поле φ, удовлетворяющее уравнениям, составленным именно так, чтобы инфляция работала. Некоторые процедуры вызывают вопросы; таковы, например, фазовые переходы (среди которых «кипение», упоминавшееся выше) — они кажутся асимметричными относительно обращения времени, но являются макроскопическими, вызывающими увеличение энтропии, и зависят от симметричных относительно обращения времени субмикроскопических процедур. Соответственно, асимметричность относительно направления времени — это проявление второго закона, а не объяснение его. Инфляционный аргумент формулируется именно таким образом, чтобы интуитивно допускалась возможность, что общий процесс снижения энтропии динамически достижим, — однако второй закон утверждает, что это не так!

Перейдем ко второму из вышеприведенных инфляционных аргументов, согласно которому такая гладкая Вселенная, как наша, неизбежно должна была возникнуть в результате инфляции. Предположим, что действительно верно следующее: если инфляционные процессы начинаются с общего исходного состояния, то к моменту завершения инфляции практически неизбежно возникает совершенно гладкая расширяющаяся Вселенная. Эта концепция принципиально противоречит второму закону. Возможно множество состояний, которые, если

попросту *сложатся* на более позднем этапе, не *будут* гладкими (а в противном случае не требовалось бы никакой инфляции, чтобы от них избавиться). Давайте обратим направление времени от такого макроскопического состояния — но учтем однородно возмущаемые субмикроскопические ингредиенты (с уравнениями, по-прежнему допускающими возможность инфляции с φ-полем и т. д.) — и допустим обращенную во времени динамическую эволюцию. Вероятно, мы к чему-то придем, но теперь энтропия возрастает в направлении *коллапса*. Таким образом, в итоге мы должны получить некое очень сложное высокоэнтропийное состояние из слившихся черных дыр, похожее отнюдь не на ФЛРУ-модель, а на ситуации, изображенные на рис. 3.14 *а* и *б* (см. раздел 3.3). Если вновь обратить время и прийти к исходной картине — примерно как на рис. 3.14 *в*, то окажется, что φ не в силах разгладить такую (гораздо более правдоподобную) исходную ситуацию. На самом деле инфляционные расчеты практически неизбежно делаются в контексте ФЛРУ, что приводит к полностью неверным выводам, так как абсолютное большинство исходных состояний *не* относится к ФЛРУ, в чем мы могли убедиться в разделах 3.3 и 3.4, и нет никаких причин полагать, что в них действительно начнется инфляция.

Что насчет первого аргумента — о том, что почти полная изотропность температуры РИ объясняется инфляцией, из-за которой различные точки на небе, соответствующем поверхности последнего рассеяния, оказываются в причинно-следственном контакте друг с другом? Опять же, проблема такова, что нам приходится выстраивать сложившуюся ситуацию исходя из «общего» (и, следовательно, более высокоэнтропийного) начального состояния. В таком случае нам вообще не помогает «причинно-следственный контакт» между этими точками. Возможно, такой контакт действительно обеспечивает разогрев, ну и что? Нам ведь требуется объяснить, *почему* и каким образом энтропия стала настолько низкой. Разогрев *увеличивает* энтропию (температуры, которые ранее были разными, выравниваются, и это лишь одно из проявлений второго закона). Поэтому, допуская разогрев на этой стадии, мы на практике еще сильнее *снижаем* энтропию, существовавшую в прошлом, и проблема возникновения такого уникального исходного состояния Вселенной только усугубляется!

Этот аспект инфляции никоим образом не касается стоящей перед нами проблемы. Как мне кажется, наблюдаемая изотропия температуры РИ — это *вторичный* эффект в контексте более крупной проблемы, а именно крайне низкого значения первичной энтропии, благодаря чему сингулярность Большого взрыва и получилась столь однородной. Температурная изотропия сама по себе не вносит серьезного вклада в проблему малой величины энтропии во Вселенной в целом (в чем мы убедились в разделе 3.5). Это просто отражение гораздо более важной изотропии тонкой *пространственной геометрии* в новорожденной Вселенной

(то есть в исходной сингулярности). Все дело в том, что гравитационные степени свободы в ранней Вселенной были невозбужденными, о чем свидетельствует полное отсутствие сингулярных белых дыр в составе Большого взрыва, — это и есть основная проблема, с которой приходится столкнуться. По какой-то глубокой причине, которая совершенно не затрагивается инфляционной гипотезой, исходная сингулярность в самом деле была исключительно однородной и *поэтому* породила очень однородный и равномерно развивающийся первозданный огненный шар, о свойствах которого свидетельствует РИ. Этот же фактор вполне может объяснять однородность температуры реликтового микроволнового фона в небе — разогрев тут ни при чем.

Скептикам, не верящим в инфляцию, таким как я, просто повезло, что в стандартных неинфляционных космологиях не допускается возможности разогрева, который охватывал бы все небо, где наблюдается РИ. Такое допущение попросту уничтожало бы всю информацию об истинной природе исходного состояния $\mathcal{B}$, поэтому если на тотальный разогрев в масштабах всего реликтового микроволнового фона действительно не было времени, то это напрямую связано с геометрией $\mathcal{B}$. В разделе 4.3 я изложу собственное мнение о важности этих проблем.

Третий вопрос, который якобы решается в инфляционной модели, — это так называемая *проблема плоскостности*. Здесь я вынужден признать, что инфляционная теория, по-видимому, оказалась по-настоящему успешной в построении прогнозов, связанных с наблюдениями, каковы бы ни были теоретические достоинства (или недостатки) этой модели. Когда впервые (в 1980-е годы) был выдвинут аргумент плоскостности, казалось, что наблюдения четко свидетельствуют о следующем: все материальное наполнение Вселенной (включая темную материю) не может составлять более трети от величины, которая бы обеспечивала плоскую форму Вселенной ($K = 0$). Это воспринималось как свидетельство в пользу Вселенной с *отрицательной* кривизной ($K < 0$), тогда как одно из необходимых следствий инфляционной модели заключается в плоскостности пространства. Тем не менее некоторые убежденные инфляционисты заявляли, что когда измерения станут точнее, будет открыта дополнительная материя и подтвердится гипотеза $K = 0$. Ситуация изменилась в 1998 году, когда появились эмпирические доказательства в пользу положительности Λ (см. разделы 1.1, 3.1, 3.7 и 3.8). По этой причине общая плотность материи возросла, причем именно настолько, чтобы подтвердился теоретический вывод о $K = 0$. Разумеется, это можно трактовать как некоторое эмпирическое подтверждение одного из основных прогнозов, который еще недавно так активно продвигали многие сторонники инфляционной теории.

Аргумент в пользу плоскостности принципиально похож на тот, при помощи которого решалась проблема гладкости. В данном случае рассуждения таковы: если бы до наступления инфляционной фазы Вселенная обладала значитель-

ной кривизной, то колоссальное растягивание, которому поспособствовала инфляция (с линейным коэффициентом 10^{26} или около того, в зависимости от варианта инфляционной теории, о которой идет речь), породило бы геометрию, где пространственная кривизна неотличима от $K = 0$. Тем не менее этот аргумент меня вновь решительно не устраивает по той же причине, что и аргумент о гладкости. Если бы «сейчас» мы обнаружили Вселенную, структура которой принципиально отличалась бы от нашей, то есть была бы крайне иррегулярной или условно ровной, но с $K \neq 0$, мы могли бы отмотать эту ситуацию обратно во времени (в соответствии с темпорально-симметричными уравнениями, допускающими такое значение φ, при котором может начаться инфляция) и посмотреть, как выглядела исходная сингулярность. Следовательно, при развитии из прошлого в будущее такая исходная сингулярность не дала бы нам условно гладкую Вселенную с $K = 0$.

Существует смежный аргумент, связанный с «тонкой настройкой», который иногда приводится как убедительный довод в пользу инфляции. Он касается соотношения ρ/ρ_c, где ρ — это локальная плотность материи, а ρ_c — критическая плотность, при которой возникла бы пространственно-плоская Вселенная. Аргумент заключается в том, что в новоиспеченной Вселенной значение ρ/ρ_c должно было стремиться к единице (возможно, до 100 десятичных знаков), поскольку в противном случае во Вселенной не сложилось бы значение ρ/ρ_c, наблюдаемое сегодня, которое по-прежнему очень близко к единице (примерно до трех десятичных знаков). Соответственно, перед нами встает проблема: объяснить такую исключительную близость ρ/ρ_c к единице на раннем этапе расширения Вселенной. Теоретики-инфляционисты утверждают, что эта первичная близость ρ/ρ_c к единице возникла в результате еще более раннего инфляционного этапа, на котором нивелировались все отклонения ρ/ρ_c от единицы, даже если они присутствовали при Большом взрыве как таковом. Вот так нужные нам условия и возникают сразу после окончания инфляционного этапа. Однако по причине, озвученной в разделе 3.6, возникает реальный вопрос: в самом ли деле инфляция нужна для такого разглаживания?

Тем не менее я вижу здесь проблему: отрицающий инфляцию должен предложить альтернативный теоретический аргумент. (Я изложу мою альтернативную точку зрения в разделе 4.3.) Вторая проблема — прекращение инфляционного этапа (упоминавшаяся выше проблема «изящного выхода» из инфляционного этапа). Такой выход должен был произойти повсюду практически одновременно — лишь при этом условии мы получаем однородное во всем пространстве значение ρ как раз в тот «момент», когда инфляция прекращается. Также просматриваются большие проблемы с требованиями теории относительности, если попытаться увязать их с таким требованием одновременности.

Так или иначе, вопрос «фиксации» значения ρ — лишь очень небольшая часть проблемы, поскольку неявно подразумевается, что должно быть всего одно число ρ, обладающее особым значением. Это незначительная часть большой проблемы, связанной с *пространственной однородностью* такой плотности, и здесь мы уже подходим к тому, что ранняя Вселенная, предположительно, очень напоминала ФЛРУ-модель. Как отмечалось в разделе 3.6, реальная проблема связана именно с этой пространственной однородностью и с крайней незначительностью вклада гравитации в энтропию. Ранее я уже говорил о том, что инфляция даже не подступается к решению этих вопросов.

Однако какие бы недочеты ни присутствовали в таком базовом обосновании инфляции, есть еще по меньшей мере два эмпирических факта, серьезно свидетельствующих в пользу этой теории. Один из них — наблюдаемые корреляции в крошечных отклонениях от однородности РИ, охватывающие очень широкие секторы неба и убедительно показывающие, что действительно существуют причинные-следственные связи между сильно удаленными друг от друга точками реликтового микроволнового фона (например, P и Q на рис. 3.40). Этот важный факт действительно не согласуется с космологией Большого взрыва по Фридману/Толману, но полностью согласуется с инфляцией (см. рис. 3.41). Если инфляционная модель ошибочна, то эти корреляции требуется объяснить исходя из какого-то другого принципа, явно связав их с процессами, происходившими до Большого взрыва! Подобные модели будут рассматриваться в разделах 3.11 и 4.3.

Еще один важный эмпирический факт, свидетельствующий в пользу инфляции, связан с природой крошечных отклонений от температурной однородности РИ по всему небу (так называемые *температурные флуктуации*). Наблюдения показывают, что они очень близки к *масштабной инвариантности* (то есть не различаются на разных масштабах). Такие факты были независимо открыты Эдвардом Р. Харрисоном и Яковом Борисовичем Зельдовичем [Zel'dovich, 1972; Harrison, 1970] за много лет до появления инфляционных идей, но последующие наблюдения РИ [Liddle and Lyth, 2000; Lyth and Liddle, 2009; Mukhanov, 2005] значительно расширили диапазон, в котором наблюдается такая инвариантность. Экспоненциальный (и поэтому самоподобный) характер инфляционного расширения в целом это объясняет, причем в инфляционной модели самые ранние зачатки иррегулярности считаются квантовыми флуктуациями φ-поля, которые каким-то образом по мере развития расширения становятся классическими. (Это один из наиболее слабых аспектов теоретической аргументации, поскольку стандартный аппарат квантовой механики не содержит никаких доводов в пользу такого квантово-классического перехода (см. [Perez et al., 2006]).) Инфляция претендует на объяснение не просто такой почти полной масштабной инвариантности, но и минимальных отклонений от нее, зависящих

от так называемого *спектрального параметра*. Эти флуктуации дают ключевой исходный вклад для вычисления *спектра мощности* РИ (и рассчитываются на основе гармонического анализа РИ по всей небесной сфере; см. раздел А.11). На рис. 3.43 показано примечательнейшее соответствие между наблюдаемыми свойствами РИ (по данным космической обсерватории «Планк», запущенной в 2009 году) и теоретическими расчетами (как минимум при больших значениях ℓ — это значения k из раздела А.11). Правда, следует иметь в виду, что в числовом отношении вклад инфляции в эти расчеты минимален (в принципе, всего два числа), и основные очертания кривой определяются другими факторами из стандартной космологии, физики частиц и гидромеханики, которые действовали в период между прекращением инфляции и последним рассеянием. Этот длительный период (около 380 000 лет) неинфляционной космологии можно уподобить промежутку между 3-поверхностями $\mathcal{B}$ и $\mathcal{D}$ на рис. 3.40, но теперь $\mathcal{B}$ соответствует моменту прекращения инфляции, а не Большому взрыву. Физика этого периода вполне понятна, и вклад инфляции в эту физику минимален [Peebles, 1980; Börner, 1988].

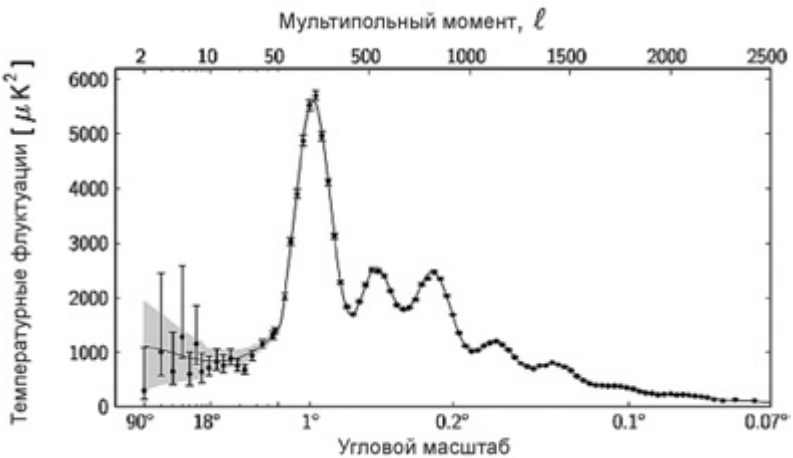


Рис. 3.43. Спектр мощности РИ, измеренный космической обсерваторией «Планк». По оси ординат отсчитываются температурные флуктуации, а по оси абсцисс (как написано над рисунком) — общий параметр сферических гармоник l (аналогичен k из раздела А.11)

С этими несомненно впечатляющими успехами контрастируют некоторые загадочные аномалии инфляционной теории, впрочем, они несколько озадачивают и безотносительно к инфляции. Таков, например, следующий факт: температурные корреляции РИ между сильно удаленными точками, по-видимому,

ограничены углом около 60° (с нашей точки наблюдения), хотя, согласно инфляционной аргументации, никаких ограничений таких корреляций быть не должно. Более того, в крупномасштабном распределении масс наблюдаются некоторые иррегулярности — например, существуют войды, уже упоминавшиеся в разделе 3.5. А на самых крупных масштабах [Starkman et al., 2012; Gurzadyan and Penrose, 2013] встречаются такие примеры асимметрии и неоднородности, которые, по-видимому, отвергают привычную инфляционную картину, причем считается, что первичный источник плотностных флуктуаций имеет *случайное* квантовое происхождение. Такие проблемы требуют серьезного объяснения, и они плохо стыкуются с общепринятыми инфляционными идеями. Я вернусь к этим вопросам в разделе 4.3.

Здесь стоит отметить, каким особенным способом выполняется такой анализ, а именно рассказать о *гармоническом анализе* (см. раздел A.11) всего космического микроволнового фона. Здесь нас будет интересовать практически только лишь спектр мощности (то есть вклад в общую интенсивность РИ, связанный со всеми модами каждого отдельного значения ℓ). Несмотря на то что эта процедура, несомненно, увенчалась удивительным успехом — обнаружено почти полное совпадение между теорией и наблюдениями, как показано на рис. 3.43 (для значений ℓ более 30), следует отметить, что у анализа такого типа есть определенные ограничения. Возможно, именно из-за них наши интересы развивались в одних направлениях, а не в других.

Для начала следует отметить, что, просто сконцентрировавшись на спектре мощности, мы будем игнорировать все больше и больше доступной информации по мере увеличения ℓ. Остановимся на этом подробнее. В разделе A.11 рассматриваются величины $Y_{\ell m}(\theta, \phi)$, называемые *сферическими гармониками*, представляющие собой различные моды, на которые можно разложить температурные паттерны, прослеживаемые в космическом микроволновом фоне. Если зафиксировать ℓ (пусть это будет неотрицательное целое число, $\ell = 0, 1, 2, 3, ...$), то у целого числа m может быть любое из $2\ell + 1$ альтернативных значений $-\ell, -\ell+1, -\ell+2, -\ell+3, ..., \ell-2, \ell-1, \ell$. Для каждой такой пары (ℓ, m) сферическая гармоника $Y_{\ell m}(\theta, \phi)$ — это конкретная функция на сфере (в данном случае имеется в виду *небесная сфера*, представленная полярными координатами θ, ϕ; см. раздел A.11). Для конкретного максимума ℓ, который мы обозначим L, общее число различных m составит L^2, а это гораздо больше количества различных ℓ, равного всего лишь L. Максимальное значение ℓ для спектра мощности, показанного на рис. 3.43, который был получен космической обсерваторией «Планк», составляет $L = 2500$, что в итоге дает около 6 250 000 разных значений, характеризующих распределение температуры в космическом микроволновом фоне. Если бы мы использовали всю информацию о РИ вплоть до такой точности, то

получили бы $L^2 = 6\,250\,000$ значений. Следовательно, полученный спектр мощности учитывает лишь $1/L = 1/2500$ всей доступной информации!

Так или иначе, несмотря на явный успех, достигнутый при сверке теории с наблюдениями при помощью спектра мощности, есть и другие способы плодотворного анализа РИ. Разложение космического микроволнового фона на моды с применением сферических гармоник — это такой анализ, который можно применять, например, к модам вибрации резинового шара. Можно утверждать, что такая модель в некотором роде аналогична Большому взрыву, но в данном случае вполне приемлемы и другие аналогии. Рассмотрим земное небо. Вряд ли есть какая-то польза в разложении его на сферические гармоники! Сложно себе представить, как вообще могла бы возникнуть астрономия, если бы мы анализировали ночное небо на материале его спектра мощности. Было бы довольно сложно обнаружить Луну как локальный объект, не говоря уж о звездах и галактиках, и для нас осталась бы загадкой природа периодических изменений, вызываемых сменой фаз Луны, тогда как просто посмотрев на нее, мы бы без труда обнаружили, что эти изменения связаны с изменением внешнего вида Луны. Я считаю, что в случае РИ мы так сильно полагаемся на анализ гармоник, поскольку руководствуемся заранее сложившимися представлениями о Большом взрыве. Есть и альтернативы, об одной из которых речь пойдет в разделе 4.3.

3.10. Антропный принцип

Представляется, что по крайней мере некоторые инфляционисты (см., например, [Guth, 2007]) уже осознали, что инфляция сама по себе не может объяснить то исключительно низкоэнтропийное в гравитационном плане и изотропное состояние, которое наблюдалось в ранней Вселенной, и что для такой однородности Вселенной требовалось бы нечто большее, нежели просто динамическая возможность возникновения инфляции. Даже если инфляция верно обрисовывает часть эволюционной истории Вселенной, требуется нечто большее, в частности условие, благодаря которому исходная сингулярность оказалась так близка к типу ФЛРУ. Если бы мы старались придерживаться центрального пункта исходной инфляционистской философии, а именно считали бы, что исходный момент Вселенной должен был являться, в сущности, *случайным*, то есть без принципиальной тонкой настройки, которая обеспечивала бы низкую энтропию, то нам потребовалось бы либо серьезно нарушить второй закон, либо найти какой-нибудь иной критерий отбора возможного раннего состояния Вселенной. Один из активно обсуждаемых возможных критериев такого типа — *антропный принцип* [Dicke, 1961; Carter, 1983; Barrow and Tipler, 1986; Rees, 2000], вкратце упомянутый в конце раздела 1.15.

Антропный принцип базируется на следующей идее: какова бы ни была природа Вселенной либо той части Вселенной, которую мы наблюдаем вокруг, подчиняющейся любым динамическим законам, которые, казалось бы, ею управляют, это должна быть природа, явственно благоприятствующая нашему существованию. Ведь определенно, если бы было иначе, то и мы оказались бы не здесь, а где-то в другом месте (например, на другой планете), или в другом времени (жили бы в другую эпоху), или даже в совершенно иной Вселенной. Разумеется, «мы» в данном случае — не обязательно люди или даже какие-либо иные создания, с которыми когда-либо доводилось встречаться людям, но все-таки некие наделенные разумом существа, способные воспринимать и рассуждать. Обычно в таком случае употребляется термин *разумная жизнь*.

Следовательно, как нередко утверждают, для наблюдаемой нами Вселенной было необходимо, чтобы в ней сложились очень специфические исходные условия, допускающие возможность возникновения разумной жизни. Совершенно произвольное исходное состояние, наподобие того, что изображено на рис. 3.14 *в* (см. раздел 3.3), вполне можно считать абсолютно непригодным для развития разумной жизни. Прежде всего, оно не приводит к развитию низкоэнтропийных высокоорганизованных ситуаций, которые кажутся абсолютно необходимыми для возникновения некоей разумной жизни, способной обрабатывать информацию, — чего, по-видимому, и требует антропный принцип. Следовательно, можно принять точку зрения, согласно которой антропный аргумент действительно налагает серьезные ограничения на геометрию Большого взрыва, если считать антропный принцип элементом Вселенной, которая может быть населена и, следовательно, постигнута разумными существами.

Однако насколько вероятно, что такого антропного требования будет достаточно, чтобы ограничить круг возможных вариантов геометрии $\mathcal{B}$ (то есть Большого взрыва) — настолько, что, пожалуй, остальное мог бы довершить инфляционный процесс? Нередко роль инфляции рассматривается именно в контексте антропного принципа [Linde, 2004]. Соответственно, требуется вообразить, что исходная 3-поверхность $\mathcal{B}$ на самом деле (была!) сложной и хаотичной, как на рис. 3.14 *в* (см. раздел 3.3), но поскольку $\mathcal{B}$ обладает бесконечной протяженностью, то на ней просто волей случая должны были найтись странные участки, настолько ровные, что инфляция могла на них возобладать. Аргументация такова, что именно такие места могли экспоненциально расширяться, в духе инфляции, и в итоге образовать во Вселенной участки, пригодные для жизни. Несмотря на то, как сложно подвести под такую картину аргументы, хотя бы условно претендующие на научную строгость, я полагаю, что можно привести по-настоящему сильные доводы против таких возможностей.

Для того чтобы выдвинуть сколь-нибудь серьезный аргумент на эту тему, потребуется придерживаться строгой версии «космической цензуры» (см. раздел 3.4), фактически подразумевающей, что $\mathcal{B}$ может рассматриваться как *пространственноподобная* (3-) поверхность (см. рис. 1.21 в разделе 1.7), так что различные участки $\mathcal{B}$ будут причинно независимы друг от друга. Однако поверхность $\mathcal{B}$ вполне может быть не слишком ровной. Тем не менее оказывается, что в каждой точке на $\mathcal{B}$ возникнет «световой конус», наполовину находящийся в будущем (см. [Penrose, 1998a]). («Точки» сингулярной границы $\mathcal{B}$ точно определяются в категориях несингулярной части пространства-времени; они указываются как *граничные неразложимые множества будущего* (см. также [Geroch et al., 1972]).)

Аргумент из раздела 3.6 предполагает, что, независимо от эффекта инфляции, доля общего «объема» $\mathcal{B}$ (в известном смысле), в которой могло бы начаться расширение наподобие того, которое мы наблюдаем вокруг — вплоть до нашего горизонта частицы, не должна превышать $10^{-10^{124}}$, поскольку нам нужна область $\mathfrak{R}$ в $\mathcal{B}$, столь уникальная, что ее энтропия оказалась бы достаточно низкой, чтобы соответствовать выкладкам, приведенным в конце раздела 3.6. (Просто для определенности: здесь я учитываю вклад темной материи (см. раздел 3.6), но в контексте изложенной ниже аргументации это несущественно). Эти расчеты просто зависят от формулы Бекенштейна — Хокинга для вычисления энтропии черных дыр, а также от оценки общей массы, вовлеченной в процесс, но нечувствительны к эффектам инфляции, с той оговоркой, что любой участвующий в инфляции процесс, вызывающий рост энтропии, просто уменьшает число подходящих нам областей в $\mathcal{B}$, то есть *уменьшает* число $10^{-10^{124}}$. Если протяженность $\mathcal{B}$ бесконечна, то, несмотря на всю маловероятность, где-то в $\mathcal{B}$ непременно найдется такая область $\mathfrak{R}$ — исключительно ровная, с предельно низкой энтропией. Далее, в соответствии с инфляционной гипотезой, эта область $\mathfrak{R}$ увеличится до размеров всей Вселенной, по природе напоминающей нашу (рис. 3.44 *а*), и в такой экспоненциально увеличивающейся за счет инфляции области сможет возникнуть разумная жизнь, причем только в такой области. Согласно такой картине проблема энтропии будет решена, по крайней мере это заявлено.

Однако в самом ли деле можно решить проблему таким образом? Поразительнейшее свойство, связанное с низкой энтропией в нашей Вселенной, заключается в том, что такая низкая энтропия — не локальное явление, которое бы наблюдалось лишь поблизости от нас. Нет, основные структуры Вселенной: планеты, звезды, галактики, галактические скопления — по-видимому, изобилуют во всей наблюдаемой части Вселенной (насколько мы можем судить) и при этом везде довольно похожи друг на друга. В частности, второй закон термодинамики одинаково работает везде, куда бы мы ни взглянули, во всей пространственно-необъятной Вселенной точно так же, как и рядом с нами. Мы

видим, что изначально материя была распределена довольно однородно и что часто она собирается в виде звезд, галактик и черных дыр. Мы видим значительную вариативность температуры (от жарких звезд до холодного вакуума), возникшую в результате гравитационного комкования. Именно благодаря этому эффекту мы получаем низкоэнтропийную звездную энергию, незаменимую для возникновения жизни, причем (предположительно) спорадически может появляться и разумная жизнь (см. в конце раздела 3.4).

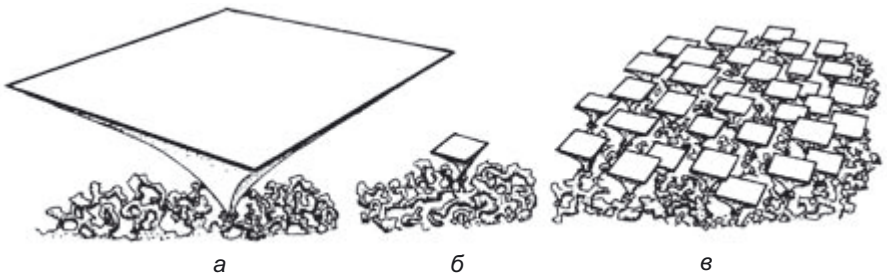


Рис. 3.44. Представления инфляционистов о том, как очень-очень редко возникает достаточно гладкая область, в которой может начаться инфляция — и возникает вселенная, подобная нашей, в которой действует второй закон термодинамики, благоприятствующий возникновению разумной жизни (*а*); с гораздо меньшей вероятностью, но гораздо проще «надуть» более компактную область, но разумных жизней, естественно, возникнет меньше (*б*); чтобы довести численность разумных жизней до уровня, который достигается в крупной области, было бы проще «надуть» множество мелких регионов, а не один крупный (*в*)

Однако разумная жизнь здесь, на Земле, нуждается лишь в малой толике этой гравитационной низкоэнтропийной энергии. Трудно понять, почему для нас должно быть жизненно важно, чтобы подобные условия существовали, например, в туманности Андромеды, хотя для нее было бы неплохо предусмотреть некоторые мягкие ограничения — скажем, чтобы оттуда на нас не могло попасть никакого опасного излучения. В данном случае важнее, что даже далеко во Вселенной повсюду наблюдается обширное подобие условий окружающей среды — условия, точно такие же, к каким мы привыкли в нашей локальной области; и, по-видимому, так повсюду, как бы далеко мы ни заглядывали. Если бы нам действительно просто требовались условия, благоприятствующие появлению разумной жизни именно здесь, то значение $\sim 10^{-10^{124}}$, характеризующее степень невероятности возникновения именно таких естественных условий, в которых мы очутились, смехотворно меньше, чем гораздо более существенный показатель, в котором мы нуждаемся «для себя». Наше существование никак

не зависит от благоприятных условий в Туманности Андромеды; тем более мы не нуждаемся в том, чтобы такие условия существовали в дальних регионах скопления Волос Вероники, равно как и в других невероятно отдаленных частях наблюдаемой Вселенной. Именно из-за относительно низкой энтропии этих крайне далеких регионов и получается такое ничтожно малое значение $+10^{-10^{124}}$, которое невообразимо меньше, чем любой мыслимый показатель, необходимый для появления разумной жизни здесь, на Земле.

Для того чтобы яснее представить этот момент, давайте вообразим, что мы не видим во Вселенной таких огромных пространств, условия в которых похожи на свойства окружающей среды в непосредственной близости от нас, а видим только до некоторого предела, допустим, вдесятеро меньше, чем сейчас. Возможно, горизонт частицы у нас у́же либо в более дальних областях состояние Вселенной ничуть не напоминает знакомые нам гравитационно-низкоэнтропийные условия. В таком случае массосодержание в наших расчетах с $10^{10^{124}}$ уменьшается с коэффициентом 10^3 и максимальная энтропия черных дыр снижается до $(10^3)^2 = 10^6$. Тогда от 10^{124} остается всего 10^{118}, поэтому теперь мы получаем абсурдно меньшую невероятность (абсурдно большую вероятность) $10^{-10^{118}}$ найти такую область, скажем $\mathfrak{Q}$, в пределах $\mathcal{B}$, которая в результате инфляции превратится как раз в вышеописанную Вселенную (см. рис. 3.44 *б*).

Можно попытаться утверждать, что эта сравнительно ограниченная область Вселенной, раздувающаяся из $\mathfrak{Q}$, вместит не так много разумных жизней, поэтому получившаяся область, где вероятность их возникновения повысилась, все равно не так благоприятна для их появления по сравнению с наблюдаемой нами большой Вселенной. Но такой аргумент неубедителен, так как, хотя мы теперь и имеем всего $\frac{1}{1000}$ таких жизней в более компактной пригодной для жизни области, можно довести это число до реальной величины, какой бы она ни была в нашей Вселенной, просто рассмотрев 1000 компактных расширяющихся вселенных (рис. 3.44(в)), увеличив невероятность

$$10^{-10^{118}} \times 10^{-10^{118}} \times 10^{-10^{118}} \times \ldots \times 10^{-10^{118}}$$

в 1000 раз, то есть $(10^{-10^{118}})^{10^3} = 10^{-10^{121}}$, что и близко несопоставимо с невероятностью порядка $10^{-10^{124}}$, — а именно такое значение мы вычислили для нашей конкретной Вселенной. Соответственно, если говорить о невероятностях, гораздо «легче» сделать множество небольших пригодных для жизни локальных вселенных (то есть 1000 $\mathfrak{Q}$-подобных суб-областей $\mathcal{B}$), чем сделать всего одну такую большую область (из $\mathfrak{R}$). Антропная аргументация нам в данном случае вообще не помогает!

Некоторые инфляционисты могли бы сказать, что картина, обрисованная мною выше, не показывает должным образом, как должна развиваться локальная область, в которой происходит инфляция, и в реальности следует говорить скорее об инфляционном *пузыре*. Интуитивно можно подумать о том, что «граница» раздувающегося пузыря должна, грубо говоря, представлять собой сопутствующую 2-поверхность, как на рис. 3.45, причем экспоненциальное увеличение масштаба, происходящее при инфляции, будет представлено просто как экспоненциально возрастающий конформный множитель Ω, где Ω связывает метрику диаграммы с частью расширяющейся Вселенной, представленной на рисунке. Но сторонники подобной «инфляции по принципу пузыря», предположительно затрагивающей лишь часть целой Вселенной, не всегда могут четко пояснить, как провести границу между расширяющейся и нерасширяющейся частями Вселенной.

Зачастую в словесных описаниях предполагается, что граница расширяющейся области (например, новый «ложный вакуум») будет распространяться вовне со скоростью света, захлестывая по пути окружающее пространство-время. По-видимому, это приводит нас к картине типа представленной на рис. 3.45 *б*, но можно ожидать, что случайные воздействия областей $\mathcal{B}$, находящихся вне $\mathfrak{R}$, радикально исказят инфляционную картину. Более того, в свете вышеизложенных аргументов это не сулит нам ничего хорошего, поскольку теперь все временные линии в раздувающейся области выходят фактически из единственной точки на $\mathcal{B}$, а 3-поверхности постоянного (инфляционного) времени теперь представлены в виде гиперболических 3-поверхностей, обозначенных на схеме тонкими линиями. У этих 3-поверхностей будет *бесконечный* объем — тогда в пределах раздувающейся области возникнет пространственно-*бесконечная* (гиперболическая) Вселенная. Теперь область инфляции описывает бесконечную Вселенную, которая, предположительно, в целом похожа на нашу, так что показатель невероятности из $10^{-10^{124}}$ превращается в $10^{-10^{\infty}}$, что вряд ли можно назвать прогрессом! Так или иначе, даже если не считать эту область пространственно-бесконечной, а полагать, что у нее есть некая граница, все равно остаются без ответов серьезные вопросы, в частности как физически трактовать необъяснимый разрыв на этой границе.

По-видимому, еще остается примерно такая возможность, как показана на рис. 3.45 *в*. Здесь вся инфляционная часть эволюции и последующая история Вселенной — в конечном итоге, вероятно, приводящая к своеобразному неинфляционному де-ситтеровскому экспоненциальному расширению, наблюдаемому сейчас, — изображена в небольшой области, напоминающей по форме усеченную пирамиду. Несмотря на то, какой «маловероятной» кажется такая картинка, в некоторых аспектах она максимально логична, поскольку если трактовать

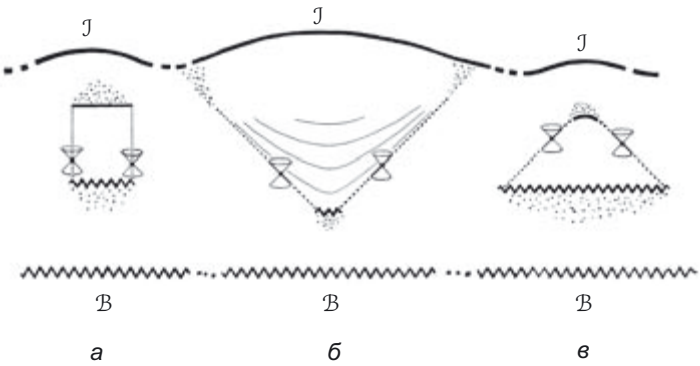


Рис. 3.45. Схематичные конформные диаграммы, иллюстрирующие различные идеи, связанные с инфляционными пузырями: *а*) граница пузыря следует временным линиям; *б*) граница пузыря расширяется со скоростью света; *в*) граница пузыря стремится вовнутрь со скоростью света; тем не менее со временем ее объем может увеличиться. Прерывистая линия сверху обозначает неопределенность, касающуюся того, как будущие бесконечности из этих моделей соотносятся с будущей бесконечностью окружающего пространства

инфляционный этап как *ложный* вакуум, то можно ожидать, что этот этап перейдет в другой. Аргументация в данном случае заключается в том, что его малый явный «размер» должен был компенсироваться колоссальным конформным множителем Ω, преобразующим его геометрию в гигантскую, возможно, дважды экспоненциально расширенную метрическую область (предполагается, что именно это и произошло с наблюдаемой Вселенной)! Подобная картина определенно сопряжена с серьезными сложностями и, подобно описанным ранее, практически не снимает вопросов, касающихся абсурдности показателя невероятности $10^{-10^{124}}$.

При всей кажущейся неподатливости рассматриваемые здесь вопросы, в частности коэффициенты невероятности, напрямую связаны с традиционными представлениями об энтропии, изложенными Больцманом, а также Бекенштейном и Хокингом [Unruh and Wald, 1982], но, по-видимому, теоретики-инфляционисты редко всерьез рассматривают такие выкладки. Я по-прежнему решительно не признаю, что инфляция помогает разобраться с важнейшими парадоксами, обозначенными в разделе 3.6. При этом пусть рассматриваемое нами исходное состояние и обладает чрезвычайно низкой энтропией, такие ограничения касаются лишь гравитационных степеней свободы, и получается Вселенная, по типу очень близкая к модели ФЛРУ. Антропный аргумент ничуть не помогает инфляции разрешить эту путаницу.

На самом деле антропный аргумент еще менее пригоден для решения загадки второго закона, нежели я описал ранее. Несомненно, известная нам жизнь развивается в гармонии с известным нам вторым законом, но антропный аргумент, основанный на факте существования жизни, фактически никак не помогает объяснить *существование* второго закона. Почему?

Жизнь на этой планете возникла в процессе неумолимого эволюционного процесса — естественного отбора, который приводил к возникновению все более и более сложных организмов, длительное существование и развитие которых предполагает низкую энтропию. Более того, жизнь критически зависит от «резервуара» низкой энтропии, то есть от жаркого Солнца на фоне сравнительно тусклого неба, а такая ситуация предполагает, что исходное состояние должно было характеризоваться исключительно низкой (гравитационной) энтропией (см. раздел 3.4). Важно понимать, что все это согласуется со вторым законом термодинамики. Общая энтропия действительно постоянно возрастает, несмотря на то что существуют выдающиеся образцы самоорганизации — экзотические растения, изящные животные и т. д., порожденные естественным отбором. Поэтому, возможно, так соблазнителен *антропный* вывод, согласно которому жизнь сама по себе как-то объясняет наличие второго закона и из-за самого нашего существования этот закон становится «неизбежно антропным».

В известном нам мире жизнь действительно возникла именно так. Мы к этому привыкли. Однако если говорить о требовании низкой энтропии, на самом ли деле это «путь наименьшего сопротивления», то есть «самый вероятный» вариант возникновения наблюдаемого мира? Скорее всего, *нет*! Можно весьма приблизительно оценить, с какой вероятностью земная жизнь в ее нынешнем виде, со всеми тончайшими молекулярными и атомными процессами и состояниями, могла бы возникнуть просто в результате случайных взаимодействий частиц, прилетающих из космоса, скажем, за шестидневный срок. Чтобы это произошло спонтанно, требуется невероятность порядка $10^{-10^{60}}$, то есть таким образом разумная жизнь возникла бы гораздо «легче», чем эволюционным путем! Нечто подобное совершенно очевидно, если отталкиваться от сути второго закона как такового. Сравнительно низкоэнтропийные более ранние состояния Вселенной, с которых началась предыстория человечества (причем энтропия этих состояний была ниже просто в соответствии со вторым законом), должны были являться гораздо менее вероятными (в описываемом смысле), чем нынешняя ситуация. Это просто второй закон в действии. Итак (в контексте невероятностей), любое состояние должно «проще» возникать обычной волей случая, поскольку оно проистекает из более раннего состояния, где энтропия ниже, если *то*, более раннее состояние, также сложилось случайно! Такую аргументацию можно продолжать вплоть до Большого взрыва. Если нас интересует вероятностный антропный

аргумент, как рассмотренный выше (связанный с субобластями $\mathfrak{R}$ и $\mathfrak{Q}$ пространства $\mathfrak{B}$), то чем позже мы позиционируем момент сотворения, тем «легче» он наступит! Естественно, должна допускаться и иная причина для возникновения абсурдно низкоэнтропийного исходного состояния (Большого взрыва), нежели простая случайность. Абсурдно несбалансированная природа этого исходного состояния (в котором явно были полностью подавлены гравитационные степени свободы) должна была возникнуть по совершенно иной и гораздо более глубокой причине. Антропный аргумент помогает нам разобраться в этих проблемах ничуть не более, чем инфляция. (Такой парадокс иногда именуется «больцмановский мозг». К этой проблеме мы вернемся в разделе 3.11).

С другой стороны (можно утверждать, что то же самое касается и инфляции), антропный аргумент действительно помогает объяснить некоторые другие важные проблемы, связанные с некоторыми глубинными свойствами физики элементарных частиц. Самый ранний пример такого рода, с которым мне довелось познакомиться, был приведен на лекции известного астрофизика и космолога Фреда Хойла (он упоминался в разделе 3.2, когда мы обсуждали космологическую модель стационарной Вселенной). Я побывал на его лекции в Кембридже, еще будучи молодым научным сотрудником в колледже Святого Иоанна. Лекция называлась, если мне не изменяет память, «Религия как наука», и мне кажется, что состоялась она в университетской церкви осенью 1957 года. В этой лекции Хойл коснулся деликатной проблемы: могут ли физические законы быть тонко настроены именно так, чтобы они благоприятствовали существованию жизни.

Всего несколькими годами ранее, в 1953 году, Хойл сделал примечательнейший прогноз о том, что у углерода должен быть ранее явно не учтенный энергетический уровень (около 7,68 МэВ), при котором углерод (и, следовательно, многие другие элементы тяжелее углерода) может образовываться в звездах — красных гигантах, которые превращаются в сверхновые, взрываются и при этом рассеивают углерод в пространстве. С некоторым трудом Хойл убедил физика-ядерщика Уильяма Фаулера из Калифорнийского технологического института проверить, на самом ли деле существует этот энергетический уровень. Как только Фаулер наконец согласился и взялся за дело, он немедленно обнаружил, что прогноз Хойла верен! Сегодня эмпирически выяснено, что точное значение этого уровня немного ниже хойловского прогноза (около 7,65 МэВ), но вполне укладывается в обозначенные пределы. (Странно, что Хойл не разделил в 1983 году Нобелевскую премию по физике, присужденную Фаулеру и Чандрасекару). Любопытно, что хотя этот факт ныне абсолютно подтверждается наблюдениями, его *теоретическое* обоснование с точки зрения современной теоретической ядерной физики по-прежнему проблематично [Jenkins and Kirsebom, 2013]. В своей лекции 1957 года Хойл отметил, что если бы энергетический уровень углерода

и энергетический уровень кислорода (выясненный ранее) не сочетались друг с другом столь точно, то ни кислород, ни углерод и близко не смогли бы достичь того соотношения, что требуется для возникновения жизни.

Этот экстраординарный и успешный прогноз энергетического уровня углерода, сделанный Хойлом, часто приводится как пример предсказания, сделанного на основе антропного принципа, — действительно, это пока единственный несомненный успех данного принципа [Barrow and Tipler, 1986; Rees, 2000]. Однако другие авторы [Kraagh, 2010] считают, что исходно прогноз Хойла не основывался на антропных идеях. На мой взгляд, спорить об этом достаточно безнадежно. Понятно, что у Хойла были серьезные основания сделать такой прогноз, поскольку углерод на Земле встречается в значительных количествах (естественно!), и он должен был откуда-то взяться. Нет необходимости апеллировать и к тому, что эти пропорции, как оказалось, благоприятствовали биологической эволюции на Земле и, следовательно, развитию разумной жизни. Чрезмерное внимание к биологической подоплеке этого аргумента — это даже преуменьшение его силы. *Да*, углерод здесь встречается в изобилии, и, согласно нашим современным физическим представлениям, сложно представить, каким еще образом он мог бы образоваться, кроме как синтезироваться в звездах (красных гигантах). Тем не менее важность этой проблемы в связи с возникновением жизни определенно весьма повлияла на стремление Хойла докопаться до *сути*: почему на Земле оказалось такое изобилие углерода.

Мне понятно, что на тот момент Хойл также всерьез интересовался «антропными рассуждениями». В 1950 году, когда я учился на старших курсах математического факультета в Университетском колледже Лондона, я прослушал серию вдохновляющих радиолекций Хойла на тему «Природа Вселенной». Явственно припоминаю, что в одной из них он поднял вопрос о благоприятности земной природы для существования жизни и сказал, что некоторые усматривают «провидение» в том, сколь идеально условия на этой планете во многих отношениях подходят для биологической эволюции. На это Хойл отвечал, что в противном случае «нас бы здесь не было; мы были бы где-то в другом месте». Удивительно «антропная» формулировка[1] Хойла особенно меня поразила, хотя и следует помнить, что сам термин *антропный* был лишь много позже предложен Брендоном Картером [1983] — именно он гораздо точнее формализовал эту идею.

На самом деле ту версию антропного аргумента, к которой Хойл апеллировал в своей радиолекции, Картер именовал *слабым* антропным принципом. Она

[1] Я точно помню, что в своей радиолекции Хойл употребил слово «провидение», но транскрипта этих лекций [Hoyle, 1950] найти не могу.

сводится к (почти тавтологичной) проблеме, упомянутой в начале этого раздела, а именно: нам требуется оказаться в благоприятной точке нашей пространственно-временной Вселенной в пространстве или во времени. В то же время *сильный* антропный принцип Картера — это проблема такого плана: могут ли быть «тонко настроены» законы природы или числовые значения констант (например, соотношение масс протона и электрона), от которых зависит работа этих законов, именно таким образом, чтобы это способствовало возникновению разумной жизни, Хойловский прогноз того, что у углерода должен быть энергетический уровень 7,68 МэВ, как раз может считаться примером сильного антропного принципа.

Еще один важный пример антропных рассуждений, на сей раз тяготеющий к слабой версии, связан с *гипотезой больших чисел*, предложенной Дираком [Dirac, 1937, 1938]. Поль Дирак исследовал некоторые фигурирующие в физике безразмерные числа, то есть не зависящие от используемых единиц измерения. Некоторые из них — не слишком большие, и их вполне можно объяснить при помощи какой-нибудь математической формулы (включающей, например, комбинации π, $\sqrt{2}$ и т. д.). Примеры такого рода — обратная постоянная тонкой структуры

$$\frac{\hbar c}{e^2} = 137{,}0359990...$$

(где e — заряд электрона) или отношение массы протона m_p к массе электрона m_e:

$$\frac{m_p}{m_e} = 1836{,}152672...,$$

хотя точных математических формул для получения ни первого, ни второго числа мы не знаем.

Правда, Дирак утверждал, что другие безразмерные числа в физике элементарных частиц столь велики (или малы), что формул для их расчета. скорее всего, не существует. Один из таких примеров — отношение сил электрического и гравитационного притяжения между электроном и протоном в атоме водорода. Это очень большое число (не зависящее от расстояния между электроном и протоном, поскольку здесь действует закон обратных квадратов), равное приблизительно

$$2{,}26874 \times 10^{39} = 2268740000000000000000000000000000000000,$$

причем, естественно, никто и не полагает, что большинство из этих цифр — действительно нули! Дирак указывал на это, пользуясь *естественной единицей*

времени, которая зависит (скажем) от массы протона m_p или массы электрона m_e, то есть соответствующих величин T_{prot} или T_{elect}, описываемых формулами

$$T_{\text{prot}} = \frac{\hbar}{m_p c^2} = 7{,}01 \times 10^{-25} \text{ секунд},$$

$$T_{\text{elect}} = \frac{\hbar}{m_e c^2} = 1{,}29 \times 10^{-21} \text{ секунд}.$$

В таком случае находим, что возраст Вселенной равен приблизительно $1{,}38 \times 10^{10}$ лет, или $4{,}35 \times 10^{17}$ секунд, то есть примерно $6{,}21 \times 10^{41}$ в протонных единицах или $3{,}37 \times 10^{38}$ в электронных.

Эти колоссальные безразмерные числа (в некоторой степени зависящие от того, какую частицу мы возьмем в качестве естественных часов) удивительно близки к отношению гравитационных и электромагнитных сил.

Дирак полагал, что у сходства этих очень больших чисел должна быть глубокая причина, касающаяся и некоторых других чисел, о которых мы вскоре поговорим. Поэтому он утверждал, что, согласно его *гипотезе больших чисел*, должно существовать (пока неизвестное) глубокое физическое объяснение такому близкому сходству этих чисел, которые по порядку величины либо почти не отличаются друг от друга (например, отношение масс протона и электрона ~1836), либо отличаются всего лишь степенями этого очень большого числа. Примеры таких степеней — величины m_{p} или m_{e}, выраженные в *планковских* (то есть абсолютных) единицах:

$$m_{\text{p}} = 7{,}685 \times 10^{-20}$$
$$m_{\text{e}} = 4{,}185 \times 10^{-23},$$

близкие к обратному значению квадратного корня из тех чисел, которые мы только что обсуждали. Можно подумать, что все эти числа — относительно небольшие множители простых степеней большого числа N, порядок которого составляет около

$$N \approx 10^{20}.$$

Далее выясняется, что все обычные элементарные частицы (протон, нейтрон, пи-мезон и т. д.) в планковских единицах имеют массу $\sim N^{-1}$. Отношение гравитационных и электромагнитных сил для обычных частиц равно $\sim N^2$. Возраст Вселенной в единицах времени обычных частиц равен $\sim N^2$, а в планковских единицах $\sim N^3$. Общая масса Вселенной в пределах нашего нынешнего (или последнего) горизонта частицы в планковских единицах также равна $\sim N^3$, а ко-

личество элементарных частиц, обладающих массой, в этой части пространства равно ~N^4. Более того, значение космологической постоянной Λ в планковских единицах примерно равно ~N^6.

Тогда как большинство этих величин, например отношение электрических и гравитационных сил или массы частиц в планковских единицах, кажутся постоянными (как минимум почти постоянными), жестко вшитыми в динамические законы Вселенной, сам возраст Вселенной, возникшей после Большого взрыва, *не может* быть константой, поскольку со временем он, очевидно, должен увеличиваться! Соответственно, рассуждал Дирак, и число N не может быть константой, равно как не могут быть константами и все эти большие (или, соответственно, малые) числа; они должны изменяться со скоростью, зависящей от степени N, релевантной в каждом случае. Таким образом, Дирак надеялся, что фундаментальное физико-математическое объяснение «немыслимо» большого числа N вообще не потребуется, так как N — это просто *дата*!

Разумеется, это было красивое и хитроумное предположение, и когда Дирак его выдвинул, оно согласовывалось с наблюдениями. В принципе, в соответствии с гипотезой Дирака, сила гравитации со временем должна была медленно ослабевать, тогда как планковские единицы, зависящие от гравитационной постоянной γ, принимаемой за единицу, также со временем должны изменяться. Однако, к несчастью для этой гипотезы, более точные последующие измерения [Teller, 1948; Hellings et al., 1983; Wesson, 1980; Bisnovatyi-Kogan, 2006] показали, что значение γ *не меняется* — по меньшей мере с той скоростью, какую предполагает эта гипотеза. Более того, в таком случае мы сталкиваемся с еще одним парадоксом: нынешняя дата в планковских единицах очень близка к N^3, а это число определяется физическими законами, явно не изменяющимися со временем.

Именно здесь нам на помощь приходит слабый антропный принцип. Роберт Дикке в 1957 году и Брендон Картер в 1983-м (более подробно) отмечали [Dicke, 1961; Carter, 1983], что если рассмотреть все основные процессы, от которых зависит жизненный цикл обычной звезды главной последовательности, например нашего Солнца, где действительно задействованы силы электрических и гравитационных взаимодействий между электронами и протонами, то можно вычислить примерную длительность жизненного цикла такой звезды. Оказывается, она равна N^2 или около того, поэтому разумные существа, жизнь которых зависит от такой звезды, нуждающиеся в стабильном и надежном потоке излучения от нее, способные глядеть во Вселенную и условно достоверно оценивать ее возраст, скорее всего, заметят удивительное совпадение: этот возраст действительно составит примерно N^2 в единицах времени обычных частиц, или N^3 в абсолютных единицах.

Это классический пример использования *слабого* антропного принципа, позволяющий решить проблему, которая изначально казалась огромной загадкой. Однако таких примеров удручающе мало (на самом деле других я и не знаю). Естественно, данный аргумент предполагает, что такие разумные существа будут условно похожи на тех, которые нам известны: их эволюция должна протекать в стабильной и подходящей планетной системе, расположенной вокруг стабильной звезды главной последовательности. Более того, в известной нам Вселенной, где, пожалуй, существуют различные «хойловские» совпадения, обеспечивающие синтез химических элементов в нужном количестве, образование этих элементов зависит от, казалось бы, «удачно сложившегося» расположения энергетических уровней. В таком случае возможен вопрос: а могла ли вообще возникнуть жизнь, если бы все эти на первый взгляд просто счастливо сложившиеся тонкие свойства физических законов хотя бы чуть-чуть отличались от известных нам — или, возможно, были бы совершенно другими? Здесь мы вступаем на территорию *сильного* антропного принципа, о котором и поговорим далее.

Сильный антропный принцип иногда формулируется в квазирелигиозном виде — словно физические законы по воле провидения были тонко настроены при создании Вселенной именно таким образом, чтобы в ней могла возникнуть (разумная) жизнь. В принципе, тот же самый аргумент можно немного переформулировать: предположить, что должно существовать неисчислимое множество параллельных вселенных, в каждой из которых есть своя совокупность значений физических констант либо даже действует иной набор (предположительно математических) законов, управляющих физическими явлениями. Идею сильного антропного принципа можно сформулировать так: все эти различные вселенные могут каким-то образом сосуществовать, и большинство из них безжизненны в том смысле, что в них нет никаких сознающих (разумных) существ. Только в тех вселенных, где такие существа могут возникнуть, фиксируются необходимые совпадения — открывать их и изумляться им остается самим этим существам.

Мне не нравится, сколь часто физики-теоретики апеллируют к таким аргументам, чтобы скомпенсировать недостаточный прогностический потенциал имеющихся у них теорий. Мы уже сталкивались с этим, когда рассуждали о ландшафте в разделе 1.16. Да, сначала на теорию струн и ее более поздние ответвления возлагались надежды — предполагалось, что эти теории приведут нас к какой-то уникальности и позволят математически объяснить различные величины, измеренные в физических экспериментах. Но теоретики-струнники поспешили укрыться за сильным антропным принципом и попытались таким образом сузить всевозможные альтернативы до абсолютного минимума. На мой взгляд, это очень грустная и убогая участь для теории.

Более того, мы почти не представляем себе, каковы именно требования для возникновения (разумной) жизни. Часто они формулируются в категориях, важных для человекоподобных существ: землеподобная планета, наличие жидкой воды, кислорода, углеродных соединений и т. д. — или же просто как простейшие химические условия. Необходимо учитывать, что с нашей, человеческой, точки зрения можно очень предвзято и ограниченно оценивать круг таких возможностей. Мы наблюдаем вокруг разумную жизнь и склонны забывать о том, как мало знаем об истинных потребностях жизни или об исходных условиях, нужных для ее возникновения. В принципе, научная фантастика нередко напоминает нам, как слабо мы разбираемся в требованиях, от которых может критически зависеть развитие разума. В качестве примеров можно привести «Черное облако» Фреда Хойла и «Яйцо дракона» Роберта Форварда, а также продолжение этого романа «Звездотрясение» [Hoyle, 1957; Forward, 1980, 1985]. Оба этих автора пишут увлекательно, их романы полны оригинальных и научно обоснованных идей. Хойл описывает высокоразвитое разумное галактическое облако. Форвард с удивительными подробностями рассказывает, как жизнь могла развиться на поверхности нейтронной звезды и существовать на значительно более высоких скоростях, нежели наша. Однако эти выдуманные человеком формы жизни все равно в основном связаны именно с такими структурами, которые мы уже обнаружили во Вселенной.

В качестве заключительного комментария: на самом деле нельзя сказать, что условия в нашей Вселенной *действительно* благоприятны для разумной жизни. Здесь, на планете Земля, разумная жизнь имеется, но у нас нет прямых доказательств того, что она существует где-то еще, поэтому разумную жизнь можно считать в лучшем случае крайне редким явлением. Можно задаться вопросом, до какой степени именно наша Вселенная благоприятна для сознающих существ!

3.11. Некоторые более фантастические космологии

Вновь напомню читателю, что эпитет *фантастический* употребляется здесь отнюдь не в уничижительном смысле. Как я подчеркивал ранее, особенно в разделах 3.1 и 3.5, наша Вселенная в различных аспектах сама по себе довольно фантастична, поэтому кажется, что и для ее постижения требуются фантастические идеи. Многие из них уже удалось напрямую проследить в РИ, которое остается не только самым непосредственным свидетельством Большого взрыва, но и проливает свет на некоторые любопытные свойства Большого взрыва как такового. Оказывается, в нем поразительным образом сочетались две противоположности: почти полная произвольность (на что указывает тепловой спектр РИ) и крайняя упорядоченность, невероятность которой просто экстремальна:

$10^{-10^{123}}$ (о чем свидетельствует однородность космического микроволнового фона во всем небе). Основная сложность с моделями, предложенными учеными, до сих пор заключается не столько в безумности этих моделей (хотя некоторые действительно в известной степени сумасшедшие), сколько в том, что они отнюдь не так безумны, чтобы с их помощью удалось объяснить сразу два этих экстраординарных эмпирических факта. Большинство теоретиков, по-видимому, вообще не понимают степени или даже экстравагантности этих фактов, характерных для юной Вселенной, хотя другие загадочные свойства, отразившиеся в РИ, подробно изучаются.

В последние годы я тоже напрактиковался в этом деле и соорудил собственную довольно безумную космологическую модель — именно в попытке найти объяснение этим специфическим аспектам Большого взрыва. Однако в уже прочитанных вами главах я намеренно старался не навязывать читателю мною собственноручно изобретенную схему. Тем не менее позволю себе очень кратко описать эту «безумную» модель в предпоследнем разделе (см. раздел 4.3). Моя трактовка тех гипотез, на которые я ссылаюсь в текущем разделе, будет еще короче, поскольку я полагаю, что будет и неуместно, и слишком сложно разбирать их здесь во всех деталях. Все дело в многообразии и разбросе — а зачастую и в сущей неправдоподобности — множества рассматриваемых схем.

Здесь следует привлечь внимание к широкому классу экстраординарных гипотез, поскольку они очень часто обсуждаются в науке и, по-видимому, воспринимаются столь серьезно, словно значительная часть заинтересованных людей расценивает их как научно *признанные* идеи! Я имею в виду теории, согласно которым наша Вселенная — лишь одна из огромного числа параллельных вселенных. В принципе, есть две, может быть, три различных направления аргументации, которые подталкивают нас к таким убеждениям.

Одно из них связано с ключевой проблемой, обсуждаемой в главе 2, которая вплотную рассматривается в разделе 2.13. Речь идет о такой интерпретации формализма квантовой механики, которая приводит нас к так называемой эвереттовской, или многомировой, интерпретации квантовой механики. В таком случае можно логически рассуждать, придерживаясь мнения, что унитарная эволюция **U** применяется ко всей Вселенной целиком, без какой-либо физически реальной редукции состояния (**R**). Соответственно, как и в случае с котом Шрёдингера, описанном в разделе 2.13, *одновременно* рассматриваются две альтернативные цепочки событий, происходящих с котом: кот проникает и через дверцу A, и через дверцу B, но это происходит в параллельных мирах. Поскольку эта точка зрения предполагает, что такие бифуркации происходят постоянно, приходится признать, что одновременно сосуществует абсолютно невообразимое

множество таких миров. Как я рассказывал в конце раздела 2.13, сам я всерьез не воспринимаю такие представления, как адекватное описание физической реальности, однако понимаю, почему этой точки зрения придерживаются многие, кто непоколебимо верит в абсолютную истинность квантового формализма.

Правда, здесь я собираюсь описывать не параллельные вселенные. В данном случае нас интересует иная линия рассуждений (хотя кому-то может показаться, что в некотором смысле две картины идентичны или по меньшей мере взаимосвязаны). Эта линия рассуждений была приведена в разделе 3.10 как одна из интерпретаций *сильного антропного принципа*, где я говорил о том, что, действительно, параллельно с нашей Вселенной могут существовать и другие вселенные, которые, однако, никак не могут сообщаться с нашей. Естественные числовые константы, выраженные в безразмерном виде (и даже законы природы), могут отличаться от вселенной к вселенной, и те параметры, которые присущи непосредственно наблюдаемой нами Вселенной, особенно благоприятны для жизни. Аргументация заключается в следующем: мы сможем понять якобы «предумышленно» установленные благоприятные числовые значения естественных констант, если вообразим истории вселенных, не слишком отличающихся от нашей, которые *действительно* существуют параллельно с нами, но имеют иные совокупности значений, выраженных в этих безразмерных числах. Только те вселенные, где значения естественных констант благоприятны для жизни, могут быть населены наделенными сознанием разумными существами. Далее утверждается, что поскольку мы — *как раз* такие существа, нашу совокупность констант следует считать благоприятной для жизни.

Ко второй точке зрения близка еще одна, которая, пожалуй, имеет более прямую физическую мотивацию и развилась на основе идей инфляционной космологии. Как вы, вероятно, помните, в начале раздела 3.9 мы говорили о том, что исходная трактовка инфляции была такой: почти сразу после возникновения Вселенной, примерно через 10^{-36} с после Большого взрыва, сложилось исходное состояние Вселенной («ложный вакуум»), где значение космологической постоянной Λ кардинально отличалось от нынешнего (весьма приблизительно в ~10^{100} раз), после чего Вселенная «туннелировала» в известный нам вакуум, и это произошло в конце периода инфляции, в момент 10^{-32} с. Однако обратите внимание на мои предостережения по поводу такого туннелирования, высказанные в начале раздела 3.9. Как мы помним, согласно «соображениям Дирака», изложенным в разделе 3.10, все большие безразмерные числа должны быть степенями некоего особенно большого числа N (только теперь мы говорим не о «возрасте Вселенной», а о «средней длительности жизненного цикла звезды из главной последовательности»). Тогда мы, в частности, находим, что космологическая постоянная $\Lambda \approx N^{-6}$ в планковских единицах. Согласно такой точке зрения, имеем инфляционную

космологическую постоянную $\Lambda_{\text{infl}} \approx 10^{100}\Lambda$, позволяющую предположить, что инфляционное значение N_{infl} должно очень приблизительно составлять

$$N_{\text{infl}} \approx 2000,$$

поскольку в таком случае $N_{\text{infl}}^{-6} \approx (2\times10^3)^{-6} \approx 10^{-20} = 10^{100}\times10^{-120} \approx 10^{100}\Lambda \approx \Lambda_{\text{infl}}$, как и требовалось, и фактические значения безразмерных чисел на инфляционном этапе должны соответствующим образом измениться, став такими, как сейчас. Это согласовывалось бы с гипотезой больших чисел Дирака, а также учитывало бы антропный аргумент Дикке — Картера, касающийся возраста Вселенной. Предположительно, значение N_{infl}, равное 2×10^3, было бы неблагоприятным для развития разумной жизни на инфляционном этапе — но хотя бы подумать об этом стоит!

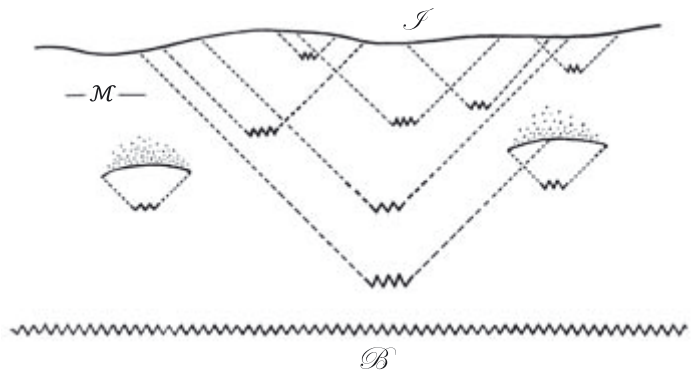


Рис. 3.46. Схематичное конформное представление вечной инфляции: считается, что, пусть и исключительно редко, происходят такие события, при которых лопаются локальные инфляционные пузыри (в данном случае преимущественно в соответствии с рис. 3.45 *б*)

Существуют различные расширенные версии инфляционных идей, наиболее влиятельные из которых называются *вечная инфляция* [Guth, 2007; Hartle et al., 2011], *хаотическая инфляция* [Linde, 1983] или *вечная хаотическая инфляция* [Linde, 1986] (Статья, в которой объясняются эти термины, написана Александром Виленкиным в 2004 году.) По-видимому, общая идея всех этих гипотез заключается в том, что инфляция может начинаться в различных точках пространства-времени, и (очень редко) такие события приводят к тому, что некоторая пространственная область начинает доминировать над всем смежным пространством, — это выглядит как стремительное экспоненциальное расширение в пространстве. Такие области часто именуются *пузырями* (см. раздел 3.10).

Возможно, наблюдаемая нами область Вселенной — как раз такой пузырь. В некоторых версиях такой гипотезы предполагается, что у такого процесса, в сущности, нет начала, и принято считать, что не будет и конца. Подобный процесс обосновывают исходя из следующего предположения инфляционной теории: хотя вероятность туннелирования из одного вакуума в другой крайне мала, в экспоненциально расширяющейся вселенной неопределенного размера такие события должны обязательно происходить время от времени (моделируется на примере де-ситтеровского пространства; см. раздел 3.1). Конформные диаграммы, изображающие такой процесс, иногда изображаются как на рис. 3.46 (в сущности, за основу взят рис. 3.45 *б* из раздела 3.10). Иногда предполагается, что такие расширяющиеся пузыри могут пересекаться, причем не вполне понятно, в каком виде это должно наблюдаться. Также сложно описать такие явления на языке геометрии, но иногда встречаются заявления, что подобные вещи уже зафиксированы [Feeney et al., 2011a, b].

Нередко считается, что подобные схемы по некоторым свойствам напоминают гипотезы о параллельных вселенных, поскольку в разных расширяющихся пузырях должны различаться значения космологической постоянной Λ (в некоторых пузырях Λ будет отрицательной). Согласно рассмотренной ранее гипотезе больших чисел Дирака (с уточнениями Дикке и Картера) вполне можно ожидать и изменений в значениях безразмерных чисел, поэтому в некоторых пузырях могут складываться благоприятные для жизни условия, а в других — нет. Опять возвращаемся к антропным соображениям, рассмотренным в разделе 3.10. Однако читатель, вероятно, уже понял, с каким скепсисом я отношусь к гипотезам, завязанным на антропные аргументы такого рода!

Одна из общепризнанных сложностей, связанных с этими «пузырьковыми» инфляционными моделями, — проблема, известная под названием «больцмановский мозг». (По-видимому, имя Больцмана ассоциируется с этой идеей благодаря короткой статье [Boltzmann, 1895], в которой Больцман рассуждал, мог ли второй закон термодинамики возникнуть в результате исключительно маловероятной случайной флуктуации. Однако сам Больцман не утверждал, что верит в эту идею, а приписал ее «старому ассистенту доктору Шютцу». На самом деле эти базовые соображения я уже излагал в разделе 3.10, пытаясь продемонстрировать, что антропный принцип абсолютно бесполезен при попытках объяснить, почему второй закон термодинамики именно такой. Однако аналогичная аргументация зачастую представляет серьезную проблему при обсуждении таких гипотез, как вечная инфляция.)

Сложность заключается в следующем. Допустим, инфляционные проблемы в явном виде требуют, чтобы в пространстве-времени с исключительно малой

вероятностью встречалась область $\mathfrak{R}$ — возможно, это всего лишь часть 3-поверхности Большого взрыва $\mathcal{B}$, однако такая область может локализоваться и где-то в *глубинах* пространства-времени, и как раз такой вариант предполагает гипотеза вечной инфляции. В данном случае $\mathfrak{R}$ подобна *семени*, запускающему инфляционный этап, в результате которого, предположительно, и возникла наблюдаемая нами Вселенная. Абсурдность антропного объяснения, почему мы обязательно должны появиться в таком пузыре, понятна сразу же, если только подумать, насколько проще (в категориях невероятностей; см. раздел 3.9) было бы просто произвести путем случайных столкновений частиц готовую и уже населенную Солнечную систему либо даже всего несколько сознающих *разумов*, именуемых *больцмановским мозгом*. Поэтому проблема заключается в следующем: почему мы возникли не *таким образом*, а в результате гораздо более маловероятного Большого взрыва, спустя $1{,}4 \times 10^{10}$ томительных лет ненужной эволюции? Мне кажется, что такой парадокс просто указывает на тщетность попыток объяснить с таких антропных позиций, почему в нашей конкретной Вселенной потребовалась столь низкая энтропия — и, на мой взгляд, явно указывает на некорректность идеи о Вселенной-пузыре. Как я уже говорил в разделе 3.10, все это свидетельствует о бессильности антропного аргумента в качестве объяснения наблюдаемой нами Вселенной именно с таким вторым законом термодинамики, который действует именно так. Требуется совершенно иное объяснение тому, почему Большой взрыв обладал столь примечательной структурой (которой, по-видимому, он в самом деле обладал, см. раздел 4.3). Если для жизнеспособности идей вечной или хаотической инфляции действительно требуется такой антропный аргумент, то я назвал бы эти идеи попросту *нежизнеспособными*.

Завершая эту главу, я хотел бы остановиться еще на двух космологических гипотезах, не столь экстравагантных, как рассмотренные выше, но по-своему интригующе фантастических. Обе они — как минимум в исходных формулировках — базируются на определенных идеях из струнной теории высших измерений. Поэтому в свете аргументов, изложенных мною в главе 1, вполне может сложиться впечатление, что я отнюдь не симпатизирую таким гипотезам. Тем не менее, как уже неоднократно отмечалось в этой главе, я считаю, что совершенно необходима теория (инфляционная или неинфляционная), которая давала бы нам очень специфическую исходную геометрию (в данном случае выраженную в виде структуры на 3-поверхности $\mathcal{B}$). На самом деле вполне логично обратиться к идеям теории струн, чтобы почерпнуть вдохновение для создания особой геометрии, позволяющей освободиться от пут классической теории относительности, особенно в контексте такой физики, которая бы действовала на $\mathcal{B}$. Более того, я искренне считаю, что такие идеи довольно важны для обеих гипотез, которые собираюсь описывать далее, пусть ни одна из них меня, в сущности, и не устраивает. Обе эти модели описывают события, происходившие до

Большого взрыва, но отличаются друг от друга в различных аспектах. Первую модель предложил Габриэле Венециано, впоследствии подробно ее разработавший вместе с Гасперини [Veneziano, 1991, 1998; Gasperini and Veneziano, 1993, 2003; Buonanno et al., 1998a, b]. Вторая модель — это экпиротическая/циклическая космология Стейнхардта, Турока и их коллег [Khoury et al., 2001, 2002b; Steinhardt and Turok, 2002, 2007].

Возможен вопрос: в чем польза распространения космологических моделей Вселенной за пределы Большого взрыва, особенно с учетом *теорем о сингулярности* (см. раздел 3.2), согласно которым, при условии сохранения классических уравнений Эйнштейна (с разумными допущениями, например со стандартным предположением о локальной положительности энергии у обычной материи), несингулярное продолжение такой модели за пределы Большого взрыва не представляется возможным? Более того, не существует общепризнанной гипотезы, в рамках которой можно было бы, прибегнув к квантовой гравитации, построить такое продолжение в обычных обстоятельствах, хотя есть ряд интересных идей такого рода. Так, в работах [Ashtekar et al., 2006 и Bojowald, 2007] предлагаются варианты такого продолжения в контексте теории петлевой квантовой гравитации. Однако если *не признавать* стандартную инфляционную космологию (что мне представляется рациональным в свете проблем, изложенных в разделах 3.9 и 3.10), то явно становится необходимо всерьез допустить возможность того, что 3-поверхности $\mathcal{B}$ нашего Большого взрыва вполне могла *предшествовать* некая «более древняя» область пространства-времени.

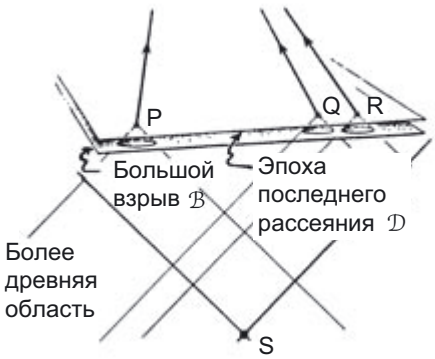


Рис. 3.47. При условии, что инфляции не было, корреляции между источниками РИ могут происходить за пределами горизонта классической космологии, если допустить, что некая область существовала до Большого взрыва. На этой схематичной конформной диаграмме событие S, произошедшее до Большого взрыва, могло бы быть причиной корреляции для событий Q, R и даже P на значительном угловом расстоянии от них

Почему так? Как упоминалось выше, в разделе 3.9, и было проиллюстрировано на рис. 3.40, в стандартных космологических моделях (Фридмана/Толмана) не должны прослеживаться корреляции в рисунке РИ между точками, угловое расстояние между которыми превышает 2°. Тем не менее сегодня такие корреляции уже убедительно доказаны при помощи наблюдений, причем угол между этими точками достигает 60°. В стандартной инфляционной модели эта проблема решается так: чрезвычайно увеличивается «конформное расстояние» между 3-поверхностями $\mathcal{B}$ и $\mathcal{D}$ (*поверхность последнего рассеяния*) (см. рис. 3.41 в разделе 3.9). Однако если *до* $\mathcal{B}$ существовало достаточно много пространства-времени, то в действительности возможны такие корреляции, возникающие под действием процессов, протекавших до Большого взрыва, как показано на рис. 3.47. Следовательно, если абстрагироваться от инфляционной модели, то имеющиеся наблюдения действительно допускают возможность каких-то явлений до Большого взрыва!

В гипотезе Гасперини — Венециано присутствует оригинальная идея, согласно которой инфляция *как таковая* произошла до Большого взрыва — что является прекрасным примером изменений правил игры по ходу матча! У этих авторов были свои причины сдвинуть таким образом момент инфляции во времени, поскольку они руководствовались струнно-теоретическими соображениями о степенях свободы; речь идет о так называемом *дилатонном поле*. Этот феномен тесно связан с величиной Ω, фигурирующей в конформных преобразованиях метрики $(\hat{\mathbf{g}} = \Omega^2 \mathbf{g})$, рассмотренных нами в разделе 3.5. В этой работе такие преобразования характеризуются как переходы от одной конформной системы отсчета к другой. В теории струн с высшими измерениями есть и такая проблема: существуют как «внутренние» измерения (они крошечные, свернутые и незаметные), так и «внешние», которые могут по-разному проявлять себя при преобразованиях. Однако независимо от этих конкретных причин, связанных с обсуждением конформных преобразований, определенно существует одна интересная возможность (крайне важная и в той схеме, которую я опишу в разделе 4.3). Например, в модели Венециано может возникать такая геометрическая странность: мы предполагаем, что инфляция, разгоняемая дилатонным полем, должна начинаться при предшествующем Большому взрыву *коллапсе*, но трактовка этого этапа будет различаться в зависимости от избранной нами конформной системы отсчета. Инфляционное *сжатие*, описанное в одной системе отсчета, в другой может выглядеть как расширение. В рамках данной модели делается серьезная попытка описать исключительно невероятную структуру Большого взрыва (исходную 3-поверхность $\mathcal{B}$), и существуют аргументы, позволяющие вывести наблюдаемую почти полную масштабную инвариантность температурных флуктуаций РИ, не прибегая к привычной инфляционной картине.

Экпиротическая[1] гипотеза Пола Стейнхардта, Нила Турока и их коллег, подобно теории струн, вводит пятое пространственное измерение, соединяющее две копии четырехмерного пространства-времени, и эти копии именуются *бранами* (предположительно они напоминают D-браны или *миры на бранах*, рассмотренные в разделе 1.15, но не описываются таким образом в статьях — авторы оперируют терминами *браны М-теории* и *браны орбиобразий*). Идея заключается в том, что непосредственно перед отскоком, происходящим в такой модели и переводящим Большое сжатие в Большой взрыв, расстояние между двумя этими бранами стремительно уменьшается, в момент отскока становится нулевым, а непосредственно после этого события вновь увеличивается. Считается, что структура такой пятимерной геометрии должна оставаться несингулярной, с внутренне связными уравнениями, несмотря на сингулярную природу проецируемого четырехмерного пространства-времени. Пусть даже инфляция в обычном понимании здесь отсутствует, выдвигаются аргументы, действительно позволяющие продемонстрировать почти полную масштабную инвариантность, характерную для температурных флуктуаций РИ [Khoury et al., 2002a].

Уместен вопрос: а как избежать проблемы, проиллюстрированной на рис. 3.48 и описанной в разделе 3.9? Речь о хаотичном схлопывании при гравитационном коллапсе с постоянно возрастающей энтропией (см. рис. 3.14 *а* и *б* в разделе 3.48), которое должно каким-то образом превратиться в Большой взрыв с низкой гравитационной энтропией. В данном случае идея такова: перед последним коллапсом, приводящим к Большому сжатию, был еще этап до отскока, на который пришлось де-ситтеровское экспоненциальное расширение (наблюдаемое нами расширение под действием Λ), которое, согласно этой модели, должно было продолжаться около 10^{12} лет. За этот период такое расширение должно было привести к полному рассасыванию черных дыр и всех прочих высокоэнтропийных остатков (правда, следует отметить, что за время такого расширения черные дыры далеко не успели бы испариться путем хокинговского излучения — на это потребовался бы гораздо более длительный срок, порядка 10^{100} лет; см. раздел 4.3). В таком случае «рассасывалась» бы *плотность* энтропии относительно сопутствующего объема; общая энтропия сопутствующего объема уменьшиться не может — это противоречит второму закону. Это касается и последующего этапа, коллапса, который должен произойти спустя 10^{12} лет, поэтому общая энтропия сопутствующего объема все равно не может уменьшиться. Итак, каким

[1] От древнегреческого слова «экпиросис»; в философии стоиков этим термином именуется периодическое уничтожение космоса в великом пожаре, происходящем в каждый Великий год. После этого космос возрождается, но лишь для того, чтобы вновь сгореть в конце следующего цикла. — *Примеч. пер.*

же образом все это согласуется со вторым законом? Чтобы понять, как все это работает, лучше перейти к *циклической* версии данной теории.

Рис. 3.48. Основная проблема, с которой сталкиваются теории, захватывающие период до Большого взрыва, заключается в том, как происходит «отскок» коллапсирующей Вселенной в состояние расширения, чтобы эта Вселенная хотя бы отдаленно напоминала нашу. Если начальный этап расширения характеризовался очень низкой гравитационной энтропией (то есть Вселенная обладала почти однородной пространственной геометрией), что, по-видимому, справедливо для нашей Вселенной, то как это могло произойти, если на этапе коллапса разворачивались предполагаемые хаотические явления (возможно, из разряда БКЛМ) с очень высокой гравитационной энтропией?

Ранее я просто описывал исходную версию экпиротической модели, которая (в трактовке Венециано) учитывает лишь единственный отскок, при котором сжатие переходит в расширение. Однако Стейнхардт и Турок расширили свою модель до непрекращающейся *последовательности* циклов, каждый из которых начинается с большого взрыва, изначально развивающегося почти в полном соответствии с ФЛРУ Λ-космологией (без раннего инфляционного этапа), но спустя примерно 10^{12} лет (преимущественно экспоненциального) расширения превращается в сжимающуюся модель, завершающуюся большим сжатием. Затем модель претерпевает *экпиротический отскок*, приводящий к новому большому взрыву, и весь процесс повторяется. Получается череда циклов, бесконечная в обоих направлениях. Все нестандартные ФЛРУ-явления (которые просто не соответствуют Λ-уравнениям Эйнштейна) управляются пятым измерением, огра-

ниченным бранами, точно так же как в рассмотренной ранее модели с однократным экпиротическим отскоком. При каждом отскоке расстояние между бранами сокращается до нуля, и этот процесс разворачивается несингулярным образом.

Напрашивается вопрос: как такая циклическая модель избегает противоречий со вторым законом? Насколько я могу судить, проблема решается с двух сторон. Во-первых, хотя рассматриваемые выше сопутствующие объемы и могут быть продолжены *сквозь* отскок, следуя при каждом отскоке вдоль сопутствующих временных линий, они не обязательно должны сохранять *размеры* от цикла к циклу — считается, что они их и не сохраняют. Давайте рассмотрим конкретный срез пространства-времени $\mathcal{S}_1$, соответствующий времени $t = t_0$ в конкретном цикле, а затем выберем в точности соответствующий ему срез $\mathcal{S}_2$ в следующем цикле, опять же соответствующий $t = t_0$ (время в каждом цикле отсчитывается от большого взрыва). Можно проследить выбранную сопутствующую область Q_1 в более раннем срезе времени и проследить ее временные линии, пройдя через отскок, и так до тех пор, пока не достигнем среза $\mathcal{S}_2$. Теперь оказывается, что если ни на йоту не отклоняться от временных линий, то достигнутая нами область Q_2 в $\mathcal{S}_2$ несравнимо больше, так что значение энтропии, которое успело сильно увеличиться по сравнению со значением в предыдущей области Q_1, теперь распределено в гораздо более обширном объеме Q_2, поэтому плотность энтропии может получиться точно такой же, как ранее в Q_1, что, однако, не противоречит второму закону.

Здесь уместно задать вопрос, можно ли обосновать простым увеличением объема совершенно колоссальное увеличение энтропии, которое, предположительно, должно было происходить на протяжении всего цикла, представляющего собой историю нашей Вселенной. Эта проблема связана с другой, упомянутой выше; ведь львиная доля энтропии даже в нынешней Вселенной, не говоря уж о Вселенной далекого будущего, заключена в сверхмассивных черных дырах, расположенных в центрах галактик. За то время, которое, предположительно, продлится общий жизненный цикл нашей Вселенной, то есть около 10^{12} лет (как минимум столько должно продлиться расширение Вселенной), эти черные дыры должны сохраниться, и именно они внесут основной вклад в энтропию Вселенной. Пусть они и будут постепенно рассеиваться и становиться разреженными, при заключительном коллапсе они вновь сольются и во многом определят заключительный этап Большого сжатия. Я не вполне понимаю, почему их можно игнорировать в экпиротических представлениях о переходе от сжатия ко взрыву!

Следует отметить, что выдвигаются и другие гипотезы, авторы которых всерьез пытаются описать особую природу Большого взрыва. Наибольшего внимания заслуживает, как мне кажется, предложенная Хокингом и Хартлом [1983] *ги-*

потеза об отсутствии граничных условий. Эту модель, несмотря на присущую ей хитроумную оригинальность, я не считаю *подлинно* фантастической. Насколько мне известно, ни одна из этих гипотез не объясняет *фантастическое* несоответствие между безумно высокоэнтропийной геометрией сингулярностей черных дыр и крайне своеобразной геометрией Большого взрыва. Нужно нечто еще более фантастическое!

Итак, все эти гипотезы подлинно фантастичные, и они призваны решить серьезные проблемы, связанные с любопытной природой Большого взрыва. Обычно они базируются не только на космологии, но и на таких физических дисциплинах, которые по каким-то причинам стали модными (теория струн, дополнительные измерения и т. д.). В них есть интересные и зачастую серьезно обоснованные идеи, наводящие на размышления. Тем не менее они кажутся мне в чем-то искусственными — по крайней мере, в их нынешней форме — и потому маловероятными; при этом они так и не решают фундаментальных проблем, обозначенных в разделе 3.4 и касающихся основополагающей роли второго закона в контексте *странной сингулярной* природы Большого взрыва.

ГЛАВА 4

Новая физика Вселенной

4.1. Теория твисторов: альтернатива теории струн?

Припоминаю, что после первой из моих принстонских лекций (о моде) ко мне обратился перспективный магистрант, занимавшийся теоретической физикой, который мучительно искал, какой области исследований себя посвятить, — он попросил меня дать ему совет. Кажется, я серьезно поколебал его энтузиазм, ведь он так стремился с головой окунуться в фундаментальную науку и заняться расширением ее границ. Ему, как и многим другим, идеи теории струн казались привлекательными, но его несколько демотивировала моя лекция, ведь я негативно отзывался о том направлении, в котором неудержимо развивалась эта дисциплина. На тот момент я не мог дать ему в достаточной мере положительный или конструктивный совет. Я не решился предложить в качестве достойного направления мою собственную теорию твисторов — не только потому, что, казалось, никто был не в силах взяться за разработку этой темы, но и потому, что амбициозному студенту было бы нелегко добиться реального прогресса, посвятив себя этой сфере, тем более что мой собеседник получил физическое, а не математическое образование. По мере развития теория твисторов стала требовать серьезной математической подготовки, знания таких вещей, которые обычно не входят в программу физического факультета. Более того, около 30 лет эта теория буксовала, не в силах преодолеть, казалось бы, несокрушимое препятствие. Речь идет о *googly-проблеме*, которую я опишу ближе к концу этого раздела.

Этот разговор случился за день или за пару дней до того, как у меня была намечена встреча за обедом с Эдвардом Виттеном — настоящей звездой Принстона, специалистом по математической физике. Я переживал, что Виттену может не понравиться мой скепсис относительно того, в каких направлениях развивается теория струн. На самом же деле Виттен меня удивил, рассказав о своей недавней работе, в которой ему удалось скомбинировать некоторые идеи теории струн

с идеями теории твисторов; сделав это, он добился некоторых выдающихся успехов в исследовании математики сильных взаимодействий. Особенно меня потрясло, что весь этот формализм Виттен разработал как раз для описания процессов, протекающих в *четырехмерном* пространстве-времени. Как вы уже поняли из первой главы, я скептически отношусь к современной теории струн почти исключительно потому, что она явно требует дополнительных пространственно-(временных) измерений. У меня возникают некоторые сложности и с суперсимметрией (которая все-таки присутствовала в виттеновской твисторно-струнной модели), но они вызывают у меня не столь острое неприятие — в любом случае, новые идеи Виттена казались далеко не так сильно зависящими от суперсимметрии, как ортодоксальная теория струн, в которой слишком большое внимание уделяется высшим измерениям пространства-времени.

Выкладки Виттена очень меня заинтересовали, ведь они не просто разворачивались именно в такой размерности пространства-времени, которую я считал правильной, но и были непосредственно применимы к известным базовым процессам из физики частиц. Речь идет о глюонном рассеянии, и такие процессы имеют фундаментальное значение при сильных взаимодействиях (см. раздел 1.3). Глюоны являются переносчиками сильного взаимодействия, точно так же как фотоны — переносчиками электромагнитных сил. Однако фотоны не вступают друг с другом в непосредственное взаимодействие, поскольку взаимодействуют лишь с заряженными (или магнитными) частицами, а не с другими фотонами. В этом заключается суть принципа *линейности* в электромагнитной теории Максвелла (см. разделы 2.7 и 2.13). Но сильные взаимодействия абсолютно *нелинейны* (они удовлетворяют уравнениям Янга — Миллса; см. раздел 1.8), и взаимодействия глюонов друг с другом составляют основу природы таких взаимодействий. Новые идеи Виттена [Witten, 2004], корни которых прослеживаются в некоторых других, более ранних работах (см., например, [Nair, 1988; Parke and Taylor, 1986; Penrose, 1967]), продемонстрировали, как можно радикально упростить считавшиеся тогда стандартными процедуры вычисления глюонного рассеяния (в то время такие расчеты делались при помощи фейнмановских диаграмм, см. раздел 1.5). Оказывается, в таком случае целый фолиант математических расчетов сокращается всего до нескольких строк.

С тех пор многие взяли на вооружение такие разработки, изначально в основном потому, что Виттен пользовался значительным авторитетом в физико-математическом сообществе; теория твисторов после этого также получила новый импульс благодаря активным разработкам: появлялись все более и более мощные методы для расчета амплитуд рассеяния частиц при высоких энергиях, когда массы (массы *покоя*) частиц можно считать несущественными и эти частицы фактически допустимо трактовать как безмассовые. Не все эти методы связаны

с теорией твисторов — в этой сфере сушествуют многочисленные иные научные школы, но общий вывод от этого не меняется: для упомянутых расчетов удалось разработать гораздо более эффективные методы, которые обладают несравнимо бóльшим потенциалом, нежели диаграммы Фейнмана. Несмотря на то, как важна была струнно-теоретическая составляющая для тех идей, что дали начало этим разработкам, эта составляющая как-то иссякла, освободив дорогу более современным разработкам. Тем не менее некоторые элементы струнных концепций (теперь уже в стандартном четырехмерном пространстве-времени) по-прежнему остаются важными.

Правда, следует отметить, что многие такие вычисления выполняются в рамках особого класса теорий, для которых характерны чрезвычайно особенные, предельно упрощенные и не всегда физически реалистичные свойства; в особенности это касается суперсимметричной теории Янга — Миллса для $n = 4$ (см. раздел 1.4). Принято считать, что эти модели подобны очень простым идеализированным ситуациям из классической механики, которые и надо изучить первым делом, например простому гармоническому осциллятору, используемому в обычной квантовой физике; далее предполагается, что более сложные и реалистичные системы удастся понять только после того, как мы основательно изучим простые. Что касается меня, то я определенно понимаю ценность изучения упрощенных моделей, отражающих самую суть процесса и действительно позволяющих сделать важные выводы; при этом я считаю, что аналогия с гармоническим осциллятором весьма ошибочна. Простые гармонические осцилляторы практически повсеместно применяются для описания малых колебаний в недиспергирующих классических системах, однако суперсимметричные поля Янга — Миллса с $n = 4$, по-видимому, не играют никакой соответствующей роли в существующих в природе квантовых полях.

Было бы уместно вкратце ознакомить вас с основами теории твисторов, только для того чтобы вы получили представление об основных идеях и о некоторых деталях, но я просто не смогу достаточно подробно рассказать здесь о разработках в области теории рассеяний, равно как и о самой теории твисторов. Более детальное представление о них можно получить в работах [Penrose, 1967a; Huggett and Tod, 1985; Ward and Wells, 1989; Penrose and Rindler, 1986; Penrose and MacCallum, 1972], а также в книге ПкР (глава 33). Центральная идея заключается в том, что пространство-время само по себе может трактоваться как *вторичное* понятие, построенное на основе более примитивного *твисторного пространства*, дополняемого квантовыми свойствами. Основной руководящий принцип в формализме этой теории таков: она объединяет элементарные понятия квантовой механики с (традиционной четырехмерной) релятивистской физикой пространства-времени, причем объединяющим звеном между двумя

этими предметами оказываются магические свойства комплексных чисел (см. разделы A.9 и A.10).

В квантовой механике действует принцип суперпозиции, позволяющий комбинировать различные состояния при помощи комплексных чисел, называемых амплитудами и имеющих фундаментальное значение в рамках этой теории (см. разделы 1.4 и 2.7). В разделе 2.9, где речь шла о квантовомеханическом понятии спина (особенно полуцелого спина $\frac{1}{2}$), было показано, как тесно эти комплексные числа связаны с геометрией трехмерного пространства, где *риманова сфера* (см. рис. A.43 в разделе A.10 и рис. 2.18 в разделе 2.9) различных возможных отношений пар комплексных амплитуд может отождествляться с различными направлениями в обычном трехмерном пространстве — они будут соответствовать возможным направлениям оси спина у частицы со спином $\frac{1}{2}$. Очевидно, что в релятивистской физике риманова сфера играет совершенно самостоятельную, особую роль, которая, опять же, специфична именно для трехмерного пространства (но вместе с одномерным временем). Оказывается, что в данном случае именно *небесную сферу*, образованную различными направлениями, пролегающими вдоль светового конуса наблюдателя, легко можно отождествить с римановой сферой [Penrose, 1959][1].

В определенном смысле теория твисторов сочетает квантово-механическую роль комплексных чисел с релятивистской, выражая обе эти физические роли на сфере Римана. Итак, мы постепенно начинаем понимать, как магия комплексных чисел могла бы дать нам некую связь для объединения квантового микромира с релятивистскими принципами физики пространства-времени макромира.

Как все это может работать? В качестве исходного представления теории твисторов рассмотрим пространство $\mathbb{PN}$ (далее я объясню, почему выбрана именно такая аббревиатура; буква $\mathbb{P}$ означает *проективное*, именно таким свойством обладают гильбертовы пространства, рассмотренные в разделе 2.8). Каждая отдельная точка $\mathbb{PN}$ *физически* уподобляется целому *лучу света*, который в категориях пространства-времени является прямой *нулевой* линией: это полная история свободно движущейся безмассовой (то есть светоподобной) частицы, например фотона (рис. 4.1). Этот луч света можно будет представить при по

[1] Есть одна тонкость, которая может смутить некоторых читателей: дело в том, что квантово-механическая риманова сфера обладает группой симметрий SU(2), более ограниченной, нежели релятивистская группа симметрий SL(2, C), Правда, вторая группа в полной мере вовлечена в явления, связанные с квантовым спином при повышении или понижении спина со стороны твисторных операторов, и называется *четвертым физическим приближением* Пенроуза [1980] (см. также [Penrose and Rindler, 1986], раздел 6.4).

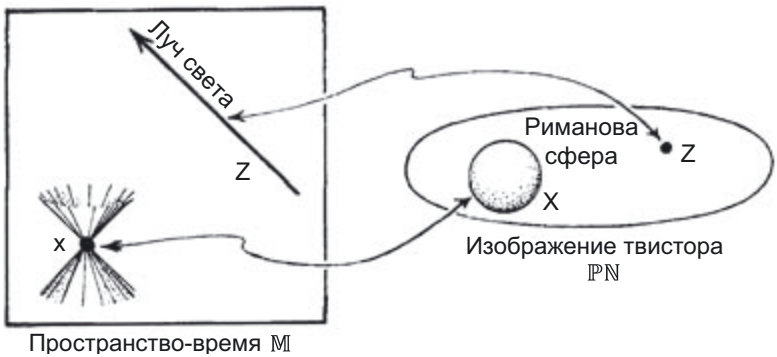


Рис. 4.1. Простейшее твисторное соответствие. Каждая точка **Z** в твисторном пространстве ℙℕ соответствует лучу света (прямая нулевая линия) в пространстве Минковского 𝕄 (возможно, в бесконечности). Каждая точка **x** в 𝕄 соответствует римановой сфере **X** в ℙℕ

мощи традиционной физики пространства-времени, причем будет считаться, что физические процекссы происходят в пространстве Минковского 𝕄, где действует специальная теория относительности (см. раздел 1.7, нотация как в разделе 1.11); но в твисторной картине весь этот луч света геометрически представляется единственной *точкой* в ℙℕ. Напротив, чтобы представить *пространственно-временную точку* (то есть *событие*) **x** в 𝕄 в терминах структур из данного твисторного пространства ℙℕ, достаточно всего лишь рассмотреть семейство всех лучей света в 𝕄, проходящих через **x**, и посмотреть, какую структуру это семейство будет иметь в ℙℕ. Исходя из вышеизложенного, можно заключить, что локус в ℙℕ, соответствующий пространственно-временной точке **x**, — это просто *риманова сфера* (в сущности, небесная сфера **x**) — простейшая риманова поверхность. Поскольку римановы поверхности — это просто комплексные кривые (см. раздел A.10), представляется логичным, что ℙℕ должно представлять собой комплексное многообразие, а такие римановы сферы должны возникать как комплексно-одномерные подмногообразия. Однако на практике такой подход не слишком работоспособен, поскольку пространство ℙℕ *нечетно*мерное (пятимерное), но в вещественном выражении у него должно быть *четное* число измерений — только тогда его можно будет представить в виде комплексного многообразия (см. раздел A.10). Требуется другая размерность! Но весьма примечательно следующее: оказывается, что, если учесть спиральность (то есть спин) и энергию безмассовой частицы, то ℙℕ физически естественным образом расширяется до вещественного шестимерного многообразия ℙ𝕋, которое естественным образом структурируется в виде комплексного 3-многообразия, по сути, комплексного проективного 3-пространства ($\mathbb{CP}^3$), именуемого *проективным твисторным пространством* (рис. 4.2).

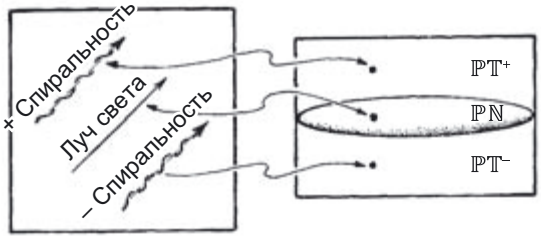


Рис. 4.2. Проективное твисторное пространство $\mathbb{PT}$ состоит из трех частей, где представлены следующие безмассовые частицы: $\mathbb{PT}^+$ — правоориентированные, $\mathbb{PT}^-$ — левоориентированные и $\mathbb{PN}$ — неориентированные

Как именно все это работает? Чтобы понять твисторный формализм, лучше всего рассмотреть комплексное четырехмерное *векторное пространство* $\mathbb{T}$ (см. раздел А.3), иногда именуемое *непроективным твисторным пространством* или просто *твисторным пространством*, причем вышеупомянутое пространство $\mathbb{PT}$ — это его проективная версия. Отношение между $\mathbb{T}$ и $\mathbb{PT}$ точно такое же, как между гильбертовым пространством $\mathcal{H}^n$ и его проективной версией $\mathbb{P}\mathcal{H}^n$, рассмотренной в разделе 2.8 (проиллюстрировано на рис. 2.16 *б* в разделе 2.8); таким образом, можно сказать, что все ненулевые комплексные множители $\lambda\mathbf{Z}$ данного ненулевого твистора $\mathbf{Z}$ (элемента $\mathbb{T}$) дадут нам *один и тот же* проективный твистор (элемент $\mathbb{PT}$). На самом деле твисторное пространство $\mathbb{T}$ по алгебраической структуре весьма напоминает четырехмерное гильбертово пространство, хотя в отношении физической интерпретации оно полностью отличается от гильбертовых пространств из квантовой механики. Вообще, именно *проективное* твисторное пространство $\mathbb{PT}$ пригодится нам при обсуждении геометрических вопросов, а когда речь зайдет об алгебре твисторов, удобнее будет работать с пространством $\mathbb{T}$.

Как и в случае с гильбертовым пространством, для элементов $\mathbb{T}$ определены *скалярное произведение*, *норма* и *ортогональность*, но в данном случае мы обойдемся без нотации вроде <...>, которой пользовались в разделе 2.8, поскольку оказывается, что удобнее будет обозначать скалярное произведение твисторов $\mathbf{Y}$ и $\mathbf{Z}$ следующим образом:

$$\bar{\mathbf{Y}} \cdot \mathbf{Z},$$

где комплексно-сопряженный твистору $\mathbf{Y}$ твистор $\bar{\mathbf{Y}}$ — это элемент *дуального* твисторного пространства $\mathbb{T}^*$, так что норма твистора $\|\mathbf{Z}\|$ описывается выражением

$$\|\mathbf{Z}\| = \bar{\mathbf{Z}} \cdot \mathbf{Z},$$

а ортогональность твисторов $\mathbf{Y}$ и $\mathbf{Z}$ означает, что $\bar{\mathbf{Y}} \cdot \mathbf{Z} = 0$. Правда, твисторное пространство $\mathbb{T}$ *алгебраически* совсем не похоже на гильбертово пространство (а вдобавок применяется с совершенно иной целью, нежели гильбертовы пространства в квантовой механике). Так, норма $\|\mathbf{Z}\|$ *не является* положительно определенной (как было бы в истинном гильбертовом пространстве)[1], а это означает, что для ненулевого твистора $\mathbf{Z}$ можно получить все три альтернативы:

$\|\mathbf{Z}\| > 0$ для *положительного*, или *правоориентированного*, твистора $\mathbf{Z}$, относящегося к пространству $\mathbb{T}^+$;

$\|\mathbf{Z}\| < 0$ для *положительного*, или *левоориентированного*, твистора Z, относящегося к пространству $\mathbb{T}^-$;

$\|\mathbf{Z}\| = 0$ для *нулевого* твистора $\mathbf{Z}$, относящегося к пространству $\mathbb{N}$.

Все твисторное пространство $\mathbb{T}$ представляет собой несвязное объединение трех частей: $\mathbb{T}^+$, $\mathbb{T}^-$ и $\mathbb{N}$, равно как его проективная версия $\mathbb{PT}$ — это дизъюнктное (несвязное) объединение $\mathbb{PT}^+$, $\mathbb{PT}^-$ и $\mathbb{PN}$ (рис. 4.3).

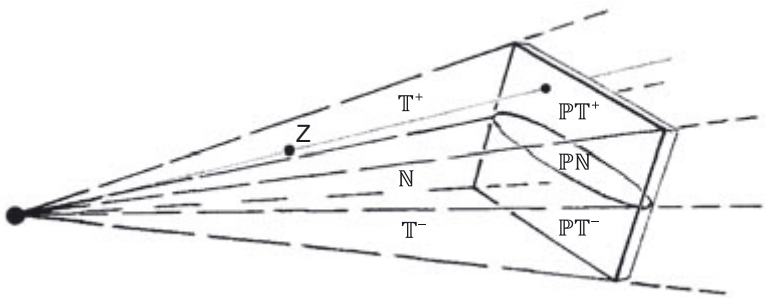


Рис. 4.3. Комплексные линии, проходящие через начало координат непроективного твисторного пространства $\mathbb{T}$, соответствуют точкам в проективном твисторном пространстве $\mathbb{PT}$

---

[1] В разделе 2.8 рассматривалось (конечномерное) гильбертово пространство. Это комплексное векторное пространство, обладающее эрмитовой структурой с положительно определенной сигнатурой (+ + + ... +). Здесь, напротив, требуется сигнатура (+ + + − −), и это означает, что в категориях традиционных комплексных координат (возведенная в квадрат) норма вектора $\mathbf{z} = (z_1, z_2, z_3, z_4)$ представляется как $\|\mathbf{z}\| = z_1\bar{z}_1 + z_2\bar{z}_2 - z_3\bar{z}_3 - z_4\bar{z}_4$. Правда, оказывается, что в стандартной твисторной нотации удобно пользоваться (полностью эквивалентными) твисторными координатами $\mathbf{Z} = (Z^0, Z^1, Z^2, Z^3)$ (не путать со степенями отдельно взятой величины $\mathbf{Z}$), где $\|\mathbf{Z}\| = Z^0\overline{Z^2} + Z^1\overline{Z^3} + Z^2\overline{Z^0} + Z^3\overline{Z^1}$.

Именно *нулевые* твисторы непосредственно связывают лучи света в пространстве-времени, то есть в проективной версии $\mathbb{PN}$ пространства $\mathbb{N}$, и эта проективная версия представляет пространство лучей света в пространстве Минковского $\mathbb{M}$ (в том числе особые «идеализированные» лучи света в бесконечности $\mathcal{F}$, когда $\mathbb{M}$ расширяется до *компактифицированного* пространства Минковского $\mathbb{M}^{\#}$, упоминаемого в разделе 1.15; см. рис. 1.41). В случае с нулевыми твисторами имеем самую непосредственную геометрическую интерпретацию отношения ортогональности $\bar{\mathbf{Y}} \cdot \mathbf{Z} = 0$ (или, что эквивалентно предыдущему выражению, $\bar{\mathbf{Z}} \cdot \mathbf{Y} = 0$). Условие ортогональности просто постулирует, что лучи света, представленные как **Y** и **Z**, *пересекаются* (возможно, в бесконечности).

Как и в обычном гильбертовом пространстве, каждый элемент **Z** в $\mathbb{T}$ имеет *фазу*, которая изменяется при умножении на $e^{i\theta}$ (θ — вещественное). Хотя эта фаза имеет некий геометрический смысл, я здесь его проигнорирую и сосредоточусь на физической интерпретации твистора **Z** *с точностью до* такого фазового умножения. В таком случае обнаружим, что **Z** представляет структуру *импульса* и *момента импульса* свободной *безмассовой частицы* в соответствии с обычными требованиями классической теории относительности (в том числе с учетом определенных предельных ситуаций, где 4-импульс обнуляется и безмассовая частица оказывается в бесконечности). Таким образом, наше описание полностью соответствует физической структуре нашей свободно движущейся безмассовой частицы, которая уже не просто луч света, поскольку в данной интерпретации учитываются как *не*нулевые твисторы, так и нулевые. Оказывается, что твистор в самом деле позволяет корректно определить энергию-импульс и момент импульса безмассовой частицы, так как включает и ее *спин* в направлении движения. Правда, в таком случае мы получаем фактически *нелокальное* описание безмассовой частицы, когда она обладает ненулевым собственным спином, так что мировая линия луча света в данном случае будет определена лишь *приблизительно*.

Необходимо подчеркнуть, что такая нелокальность — это не дефект, который можно было бы списать на необычность описаний твисторов; это (часто упускаемый из виду) аспект *традиционного* описания безмассовой частицы, имеющей спин, если она представлена в категориях собственного импульса и момента (последний иногда понимается как *момент количества движения* вокруг некоторого начала координат; см. также раздел 1.14, рис. 1.36). Хотя конкретные алгебраические описания в теории твисторов отличаются от традиционных, в только что приведенных мною интерпретациях нет ничего нетрадиционного. По меньшей мере на данном этапе теория твисторов дает всего лишь своеобразный формализм. Она не вносит никаких новых предположений об устройстве физического мира (чего не скажешь, например, о теории струн). Правда, она позволяет взглянуть на вещи под необычным углом, предполагая, что, возможно, феномен пространства-вре-

мени было бы полезно рассматривать как некое вторичное свойство физического мира, а геометрию твисторного пространства считать в некотором роде более фундаментальной. Также следует отметить, что аппарат теории твисторов к настоящему времени еще определенно не вызывает какой-либо экзальтации, а его применимость в теории рассеяния частиц очень высоких энергий обусловлена исключительно тем, что твисторный формализм удобен для описания таких процессов, при которых массы покоя частиц можно игнорировать .

Для твистора $\mathbf{Z}$ принято использовать координаты (четыре комплексных числа), где первая пара компонентов (Z^0, Z^1) — это комплексные компоненты величины $\boldsymbol{\omega}$, именуемой *2-спинор* (ср. также с разделом 1.14), а вторую пару (Z^2, Z^3) образуют два компонента величины π, представляющей собой 2-спинор немного иного типа (иного, поскольку он является дуальным, сложносопряженным), поэтому весь[1] твистор целиком можно представить в виде

$$\mathbf{Z} = (\boldsymbol{\omega}, \pi).$$

(Во многих недавних работах используется нотация, где вместо π употребляется λ, а вместо $\boldsymbol{\omega}$ — μ; ею изначально пользовался и я в работе [Penrose, 1967a], где повелся на некоторые неподходящие «общепринятые соглашения», в основном связанные с размещением верхних и нижних индексов. На практике многие до сих пор придерживаются этих неудачных соглашений.) Здесь я не хочу подробно описывать, что представляет собой 2-спинор (иногда именуемый *спинором Вейля*), но если хотите составить некоторое представление об этой идее, вернитесь к разделу 2.9. Два компонента (амплитуды) w и z, отношение которых $z{:}w$ определяет направление спина для частиц со спином $\frac{1}{2}$ (см. рис. 2.18), можно уподобить двум компонентам, определяющим 2-спинор, и это в равной степени касается $\boldsymbol{\omega}$ и π.

Однако чтобы точнее описать 2-спинор с геометрической точки зрения, отсылаю читателя к рис. 4.4, где продемонстрировано, как можно изобразить (ненулевой) 2-спинор в контексте пространства-времени. Строго говоря, изображение 4.4 *б* должно относиться к пространству, касательному пространству-времени в некоторой точке (см. рис. 1.18 *в*), но поскольку наше пространство-время в данном случае является плоским пространством Минковского $\mathbb{M}$, можно считать, что вся эта картина относится к целому $\mathbb{M}$ и выстраивается вокруг некоторого начала координат O. Вплоть до фазового множителя 2-спинор представляет собой направленный в будущее нулевой вектор, именуемый его *флагштоком* (отрезок OF

[1] В стандартной 2-спинорной записи индексов [Penrose and Rindler, 1984] индексы у $\boldsymbol{\omega}$ и $\boldsymbol{\pi}$ ставятся так: ω^A и $\pi_{A'}$.

на рис. 4.4). Можно считать, что *направление* флагштока задается точкой P на абстрактной (римановой) сфере $\mathcal{S}$, состоящей из нулевых направлений будущего (рис. 4.4 *а*). Фаза 2-спинора как таковая (с точностью да знака) представлена вектором $\overrightarrow{PP'}$, касательным к $\mathcal{S}$ в точке P, где P′ — точка в $\mathcal{S}$, соседняя с P. В категориях пространства-времени эта фаза задается как нулевая полуплоскость, ограниченная флагштоком и именуемая *полотнищем флага* (см. рис. 4.4 б). Хотя подробности этой картины для нас не слишком важны, полезно учитывать, что спинор — это совершенно определенный геометрический объект (неоднозначный лишь в том смысле, что построенная нами картина не позволяет отличить конкретный 2-спинор от *отрицательного* ему 2-спинора).

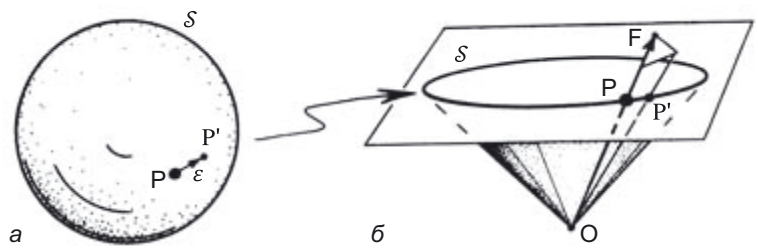


Рис. 4.4. Геометрическая интерпретация 2-спинора, где его можно представить как пространство, касательное к пространственно-временной точке O, либо представить в полном пространстве Минковского $\mathbb{M}$ относительно некоторого начала координат O. Риманова сфера $\mathcal{S}$ (*а*) представляет нулевые направления в будущем «флагштока» OR (*б*). Касательный вектор $\overrightarrow{PP'}$ в точке **P** сферы $\mathcal{S}$ представляет направление «полотнища» вдоль OF, которое описывает фазу 2-спинора с точностью до знака.

Если говорить о твисторе **Z**, то его 2-спинорная часть $\boldsymbol{\pi}$ с точностью до фазы описывает четырехмерный вектор энергии-импульса частицы как тензорное произведение:[1]

$$\mathbf{p} = \boldsymbol{\pi}\bar{\boldsymbol{\pi}}$$

[1] Запись спинорного произведения просто в виде ab (опуская символ между сомножителями) означает *несвернутое* произведение, поэтому произведение $\pi\bar{\pi}$ дает нам вектор (фактически *ковектор*) $p_a = p_{AA'} = \pi_A\bar{\pi}_{A'}$, где каждый 4-пространственный индекс представлен (при применении абстрактного индексного формализма, см. Penrose and Rindler [1984]) как пара спинорных индексов — один со штрихом, другой без. В явном виде выражение для тензора момента импульса M^{ab} в терминах ω^A и $\pi_{A'}$ можно представить в спинорной форме [Penrose and Rindler, 1986] в виде $M^{ab} = M^{AA'BB'} = i\omega\left(^A\bar{\pi}^B\right)\varepsilon^{A'B'} - i\bar{\omega}\left(^{A'}\pi^{B'}\right)\varepsilon^{AB}$, причем круглые скобки означают симметризацию, а символы у «эпсилон» являются кососимметричными.

(см. раздел 1.5), причем черта над символом означает комплексное сопряжение. Если умножить $\boldsymbol{\pi}$ на фазовый коэффициент $e^{i\theta}$ (где θ — вещественное число), то $\bar{\boldsymbol{\pi}}$ будет умножаться на $e^{-i\theta}$, и, следовательно, $\mathbf{p}$ не изменится; действительно, ведь $\mathbf{p}$ — это флагшток 2-спинора $\boldsymbol{\pi}$. Когда $\boldsymbol{\pi}$ известно, дополнительные данные в $\boldsymbol{\omega}$ эквивалентны релятивистскому моменту импульса частицы (см. раздел 1.14) относительно вращения вокруг начала координат, выраженному в терминах (симметризованных) произведений $\boldsymbol{\omega}\bar{\boldsymbol{\pi}}$ и $\boldsymbol{\pi}\bar{\boldsymbol{\omega}}$.

Комплексная сопряженная величина $\bar{\mathbf{Z}}$, которая имеет вид

$$\bar{\mathbf{Z}} = (\bar{\pi},\ \bar{\omega}),$$

это дуальный твистор (то есть элемент $\mathbb{T}^*$), а значит, он отлично подходит для скалярных произведений с твисторами (см. раздел А.4). Следовательно, если $\mathbf{W}$ — любой дуальный твистор $(\boldsymbol{\lambda}, \boldsymbol{\mu})$, то можно образовать его скалярное произведение с $\mathbf{Z}$, которое есть комплексное число

$$\mathbf{W} \cdot \mathbf{Z} = \boldsymbol{\lambda} \cdot \boldsymbol{\omega} + \boldsymbol{\mu} \cdot \boldsymbol{\pi}.$$

Норма $\|\mathbf{Z}\|$ твистора $\mathbf{Z}$ в таком случае является (вещественным) числом

$$\|\mathbf{Z}\| = \bar{\mathbf{Z}} \cdot \mathbf{Z} = \bar{\boldsymbol{\pi}} \cdot \boldsymbol{\omega} + \bar{\boldsymbol{\omega}} \cdot \boldsymbol{\pi} = 2\hbar s.$$

Оказывается, что s — это *спиральность* безмассовой частицы, описываемая $\mathbf{Z}$. Если s положительна, то частица обладает правым спином со значением s; если s отрицательна, то частица обладает левым спином со значением $|s|$. Соответственно, правоориентированный фотон (с круговой поляризацией) будет иметь $s = 1$, а левоориентированный $s = -1$ (см. раздел 2.6). Именно такая картина изображена на рис. 4.2. Правоориентированный и левоориентированный гравитоны будут иметь соответственно $s = 2$ и $s = -2$. Если считать нейтрино и антинейтрино безмассовыми частицами, то они будут иметь соответственно $s = -1$ и $s = +1$.

При $s = 0$ частица является бесспиновой, и в таком случае твистор $\mathbf{Z}$, именуемый *нулевым твистором* ($\bar{\mathbf{Z}} \cdot \mathbf{Z} = 0$), можно геометрически интерпретировать в пространстве Минковского $\mathbb{M}$ (или в его компактифицированной версии $\mathbb{M}^\#$, если допустить $\boldsymbol{\pi} = 0$) как луч света или *прямую нулевую линию* $\mathbf{z}$ (нулевую геодезическую — см. раздел 1.7). Это мировая линия частицы, соответствующая приведенному выше описанию «исходной картины» нулевого твистора, показанного на рис. 4.1. Луч света $\mathbf{z}$ имеет в пространстве-времени направление $\mathbf{p}$, тогда как $\mathbf{p}$ вдобавок дает *энергетический* масштаб $\mathbf{z}$, и такой масштаб также зависит от самого твистора $\mathbf{Z}$. Направление флагштока $\boldsymbol{\omega}$ теперь также получает прямую интерпретацию при условии, что световой конус $\mathbf{z}$ пересекается со световым

конусом, выходящим из начала координат O, в некоторой конечной точке Q — это интерпретация в виде отрезка OQ, то есть радиус-вектора **y**, записываемого как $\omega\bar{\omega}\left(i\bar{\omega}\cdot\pi\right)^{-1}$ (рис. 4.5).

Такое базовое соответствие между пространством Минковского $\mathbb{M}$ и твисторным пространством $\mathbb{PN}$ алгебраически выражается при помощи так называемого *отношения инцидентности* между *нулевым* твистором **Z** и пространственно-временной точкой **x**, которое можно выразить в виде[1]

$$\omega = \mathrm{i}\mathbf{x}\cdot\pi,$$

что в более знакомой нам матричной нотации приобретает вид

$$\begin{pmatrix} Z^0 \\ Z^1 \end{pmatrix} = \frac{i}{\sqrt{2}}\begin{pmatrix} t+z & x+iy \\ x-iy & t-z \end{pmatrix}\begin{pmatrix} Z^2 \\ Z^3 \end{pmatrix},$$

где (t, x, y, z) — стандартные пространственно-временные координаты точки **x** в пространстве Минковского (при $c = 1$). Инцидентность в $\mathbb{M}$ интерпретируется как пространственно-временная точка x, лежащая на нулевой линии z. В контексте $\mathbb{PN}$ интерпретация инцидентности такова, что точка $\mathbb{P}\mathbf{Z}$ лежит на проективной линии **X**, и в исходной картине, изложенной нами ранее, данная линия соответствует римановой сфере, представляющей **x**, а эта риманова сфера соответствует комплексной проективной прямой линии в проективном трехмерном пространстве $\mathbb{PT}$ и фактически расположена в $\mathbb{PN}$, являющемся подпространством $\mathbb{PT}$ (см. рис. 4.1).

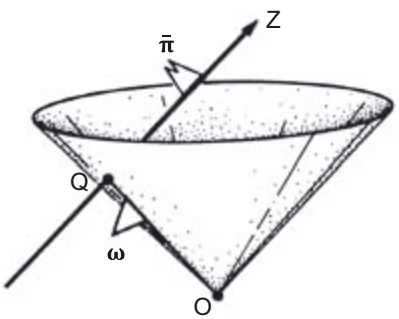


Рис. 4.5. Направление флагштока ω-части нулевого твистора $\mathbf{Z} = (\boldsymbol{\omega}, \pi)$. Если луч света **Z** пересекается со световым конусом, исходящим из начала координат O, в некоторой конечной точке Q, то направление флагштока ω будет соответствовать отрезку OQ; более того, сам ω фиксирован (задается π) радиус-вектором Q, $\omega\bar{\omega}\left(i\bar{\omega}\cdot\pi\right)^{-1}$.

[1] Индексная форма этого отношения: $\omega^{A} = ix^{AB'}\pi_{B'}$.

Когда $s \neq 0$ (то есть твистор $\mathbf{Z}$ ненулевой), отношение инцидентности $\omega = i\mathbf{x} \cdot \pi$ не удовлетворяется ни в одной вещественной точке $\mathbf{x}$, и в данном случае невозможно выделить единственную мировую линию. В таком случае положение частицы становится в известной степени *нелокальным*, как упоминалось ранее [Penrose и Rindler, 1986, разделы 6.2 и 6.3]. Однако отношению инцидентности могут удовлетворять *комплексные* точки $\mathbf{x}$ (точки на *комплексификации* $\mathbb{CM}$ пространства Минковского $\mathbb{M}$), что на самом деле важно, поскольку лежит в основе условия положительности частоты, которому удовлетворяют твисторные волновые функции; о них мы поговорим чуть позже.

Важная и даже немного магическая характеристика такой трактовки физики с позиции твисторного пространства (связанная с описанной ранее геометрией) — это та простота, с которой теория твисторов дает все решения уравнений поля для безмассовых частиц с любой заданной спиральностью [Penrose, 1969b; см. также Penrose, 1968; Hughston, 1979; Penrose and MacCallum, 1972; Eastwood et al., 1981; Eastwood, 1990]. Фактически некоторые варианты такой формулы были найдены гораздо раньше [Whittaker, 1903; Bateman 1904, 1910]. Эта формула складывается естественным образом, стоит лишь задуматься, как описать волновую функцию безмассовой частицы в твисторных категориях. В традиционных физических описаниях волновая функция частицы (см. разделы 2.5 и 2.6) может быть представлена как комплексная функция $\psi(\mathbf{x})$ (пространственного) положения $\mathbf{x}$ или, иным образом, как комплексная функция $\tilde{\psi}(\mathbf{p})$ 3-импульса $\mathbf{p}$. Теория твисторов позволяет представить волновую функцию безмассовой частицы и другими способами, а именно как комплексную функцию $f(\mathbf{Z})$ твистора $\mathbf{Z}$, именуемую просто *твисторной функцией* частицы, либо как комплексную функцию $\tilde{f}(\mathbf{W})$ *дуального* твистора $\mathbf{W}$, которую называют дуально-твисторной функцией частицы. Оказывается, что функции f и $\tilde{f}$ обязательно должны быть *голоморфными*, то есть комплексно-аналитическими (таким образом, чтобы они не «привлекали» комплексно-сопряженных переменных $\bar{\mathbf{Z}}$ или $\bar{\mathbf{W}}$; см. раздел A.10). В разделе 2.13 мы отмечали, что $\mathbf{x}$ и $\mathbf{p}$ — это так называемые канонически сопряженные переменные; соответственно, в каноническом сопряжении друг с другом находятся $\mathbf{Z}$ и $\bar{\mathbf{Z}}$.

Такие твисторные функции (а также дуально-твисторные функции — но для определенности давайте сосредоточимся лишь на твисторных) обладают некоторыми примечательными свойствами. Наиболее заметное из них заключается в том, что для частицы с определенной спиральностью твисторная функция *однородна*, то есть для некоторого числа d, именуемого *порядком однородности*,

$$f(\lambda\mathbf{Z}) = \lambda^d f(\mathbf{Z}),$$

если λ — любое ненулевое комплексное число. Число d зависит от спиральности s и определяется по формуле

$$d = -2s - 2.$$

Согласно условию однородности, f можно считать просто некоторой функцией в *проективном* твисторном пространстве $\mathbb{PT}$ (такие функции иногда называются *скрученными функциями* в $\mathbb{PT}$, причем степень «скрученности» определяется d).

Следовательно, твисторная форма волновой функции для безмассовой частицы с заданной спиральностью s примечательно проста, хотя есть важный нюанс, к которому мы вскоре подойдем. *Простой* аспект этого представления заключается в том, что уравнения поля (в принципе, все подходящие уравнения Шрёдингера), описывающие соответствующую волновую функцию положения в пространстве $\psi(\mathbf{x})$, фактически исчезают почти в полном составе! Нам всего лишь требуется твисторная функция $f(\mathbf{Z})$ нашей твисторной переменной $\mathbf{Z}$, которая будет *голоморфна* (то есть без $\bar{\mathbf{Z}}$; см. раздел А.10) и *однородна*.

Эти уравнения поля важны и в классической физике. Например, когда $s = \pm 1$, что соответствует порядкам однородности $d = -4$ и 0, мы получаем общие решения электромагнитных уравнений Максвелла (см. раздел 2.6). Когда $s = \pm 2$ (порядки однородности $d = -6$ и $+2$), мы получаем общие решения вакуумных эйнштейновских уравнений для *слабого поля* (то есть «линеаризованных»); $\mathbf{G} = \mathbf{0}$, а «вакуумный» означает $\mathbf{T} = \mathbf{0}$; см. раздел 1.1. В каждом случае решения уравнений поля автоматически получаются из твисторной функции согласно простой процедуре, которая известна из теории функций комплексной переменной и называется *контурное интегрирование* (см., например, ПкР, раздел 7.2).

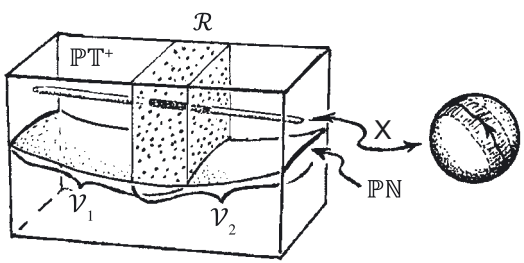


Рис. 4.6. Твисторная геометрия соотносится с контурным интегрированием уравнения поля с положительной частотой (уравнения Шрёдингера) для свободной безмассовой частицы с заданной спиральностью. Твисторная функция может быть определена в области пересечения $\mathcal{R} = \mathcal{V}_1 \cap \mathcal{V}_2$ двух открытых множеств $\mathcal{V}$ и $\mathcal{V}_2$, которые вместе покрывают $\mathbb{PT}^+$

В квантовом контексте проявляется еще одна особенность волновых функций (свободных безмассовых частиц), выводимая непосредственно из твисторного формализма. Дело в том, что волновые функции свободных частиц должны удовлетворять важнейшему условию — *положительности частоты*, которое фактически сводится к тому, что энергия не вносит отрицательного вклада в волновую функцию (см. раздел 4.2). Это происходит автоматически, если гарантировать, что твисторная функция странным, но уместным в данном случае образом будет *определена* в верхней половине проективного твисторного пространства $\mathbb{PT}^+$. Такая геометрия схематически показана на рис. 4.6. Здесь мы полагаем, что волновая функция вычисляется в *комплексной* пространственно-временной точке **x**, представленной в виде линии, обозначенной на рисунке буквой **X**. Эта линия, целиком расположенная в $\mathbb{PT}^+$, фактически представляет собой риманову сферу, как показано в правой части рисунка.

«Любопытный» способ, которым f фактически «определяется» в $\mathbb{PT}^+$, понятен на примере области $\mathcal{R}$ в $\mathbb{PT}^+$, обозначенной прерывистой линией; эта область фактически и является областью определения функции. Таким образом, f может иметь особенности в тех частях $\mathbb{PT}^+$, которые лежат за пределами $\mathcal{R}$ (справа или слева от $\mathcal{R}$ на рисунке). Линия (риманова сфера) **X** попадает в $\mathcal{R}$ в кольцевой области, а контурный интеграл получается в результате обхода замкнутого контура в этой кольцевой области. Так мы получаем значение пространственно-координатной волновой функции $\psi(\mathbf{x})$ для (комплексной) пространственно-временной точки **x**, причем в такой конструкции автоматически удовлетворяются соответствующие уравнения поля *и* соблюдается условие положительности энергии!

В чем же заключается загвоздка, о которой я упоминал выше? Все дело в том, какое именно значение вкладывается в вышеупомянутое любопытное замечание о том, что f «определяется в $\mathbb{PT}^+$», хотя в действительности область определения этой функции — гораздо более компактная область $\mathcal{R}$. Какой математический смысл во всем этом?

На самом деле здесь есть некоторые тонкости, которые я не смогу как следует описать, не вдаваясь в технические подробности. Но важнейшая черта $\mathcal{R}$ заключается в том, что ее можно считать областью пересечения двух открытых областей $\mathcal{V}_1$ и $\mathcal{V}_2$, которые вместе покрывают $\mathbb{PT}^+$:

$$\mathcal{V}_1 \cap \mathcal{V}_2 = \mathcal{R} \text{ и } \mathcal{V}_2 \cup \mathcal{V}_1 = \mathbb{PT}^+;$$

см. рис. 4.6. (Символы $\cap$ и $\cup$ означают *пересечение* и *объединение* соответственно; см. раздел А.5.) В более общем виде можно рассмотреть накрытие $\mathbb{PT}^+$ *большим* количеством открытых множеств, и в таком случае наша твисторная функция

должна определяться в контексте *совокупности* голоморфных функций, определяемых в пределах различных попарных пересечений между этими открытыми множествами. Из этой совокупности мы будем извлекать конкретную величину, именуемую элементом 1-й *когомологии*. Именно этот элемент 1-й когомологии и будет той величиной, которая позволяет построить твисторное представление волновой функции!

Все это кажется довольно сложным — и определенно таковым и окажется, если я попытаюсь все объяснить в деталях. Однако такие осложнения помогают донести важную идею, которая, на мой взгляд, фундаментально связана с таинственной *нелокальностью*, проявляющейся в квантовом мире и описанной в разделе 2.10. Попробую все упростить, сократив терминологию и назвав элемент 1-й когомологии просто *1-функцией*. В таком случае обычная функция будет считаться 0-функцией, и мы также сможем оперировать сущностями более высокого порядка, именуемыми 2-функциями (это элементы 2-й когомологии, определяемые в контексте наборов функций, которые, в свою очередь, определяются на *тройных* пересечениях между открытыми множествами покрытия), и т. д. — работать с 3-функциями, 4-функциями... (Когомология того типа, которой я пользуюсь здесь, называется *когомологией Чеха*; существуют и иные (эквивалентные, но устроенные совершенно иначе) процедуры, например когомология Дольбо) [Gunning and Rossi, 1965; Wells, 1991].)

Итак, как же понять, чему именно соответствует 1-функция? Легче всего, на мой взгляд, объяснить эту проблему в простых категориях; для этого предлагаю читателю рассмотреть невозможный треугольник, показанный на рис. 4.7. Складывается впечатление, что это трехмерная структура, которая фактически не может существовать в обычном евклидовом трехмерном пространстве. Допустим, у нас есть полный ящик деревянных брусочков и уголков и инструкция о том, как их склеивать. Предполагаем, что на инструкции по соединению каждых двух элементов изображена картинка, которая с точки зрения наблюдателя *локально* непротиворечива, но неоднозначна до неопределенности, если задуматься, на каком расстоянии от зрителя расположены ее элементы. Тем не менее в случае, представленном на рис. 4.7, такой целостный объект невозможно собрать в трехмерном пространстве, поскольку зритель не сможет правильно просчитать расстояния до различных элементов объекта.

Именно нелокальная невозможность, наглядно продемонстрированная на этом рисунке, дает хорошее представление о 1-й когомологии, которая, в сущности, и выражается 1-функцией. Действительно, если выполнить инструкции по склеиванию треугольника в точном соответствии с процедурой 1-й когомологии, то получится точная 1-функция, позволяющая измерить степень невозможности результирующей фигуры. Поэтому если данная величина получится ненулевой,

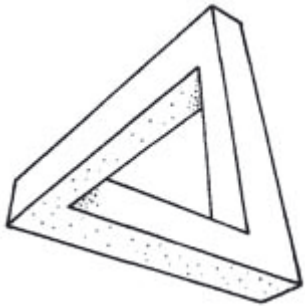


Рис. 4.7. Невозможный треугольник хорошо иллюстрирует 1-ю когомологию. Степень невозможности фигуры — это нелокальная величина, которая может быть точно определена как элемент 1-й когомологии. Если разрезать треугольник *в любом месте*, то невозможность исчезнет — так становится понятно, почему это свойство невозможно локализовать в конкретной точке. Твисторная функция играет очень схожую нелокальную роль и должна интерпретироваться как элемент (голоморфной) 1-й когомологии

то перед нами невозможный объект, как на рис. 4.7. Обратите внимание: если закрыть на картинке любой угол или соединительное ребро, то получится фигура, которую вполне можно представить в евклидовом трехмерном пространстве. Следовательно, невозможность фигуры носит не локальный, а *глобальный* характер, присущий всему изображению в целом. Соответственно, 1-функция, описывающая степень невозможности, действительно нелокальна и относится к структуре в целом, а не к одной из ее частей [Penrose, 1991; Penrose and Penrose, 1958]. Ранее изображать невозможные структуры пытались М. К. Эшер, Оскар Ройтерсвард (см. [Ernst, 1986, p. 125—134; Seckel, 2004].

Аналогично твисторная волновая функция отдельно взятой частицы является нелокальной сущностью, а именно 1-функцией, которую можно получить из локальных совпадающих функций на пересечениях областей, в сущности, точно таким же образом, как и при построении такого невозможного объекта. Теперь такая когомология возникает не из-за негибкости деревяшек в обычном трехмерном евклидовом пространстве, а из-за негибкости голоморфных функций, проявляющейся в процессе аналитического продолжения, на что я указываю в разделе A.10 (а также в разделе 3.8). Эта примечательная «негибкость», свойственная голоморфным функциям, словно наделяет функцию собственным «норовом» — она идет куда хочет, и свернуть с пути ее не заставишь. В описываемой ситуации такая негибкость может помешать определить эту функцию на всем $\mathbb{PT}^+$ сразу. Можно сказать, что голоморфная 1-функция описывает своеобразное препятствие, мешающее такой глобальности, — именно в этом и заключается нелокальная природа твисторной волновой функции.

Таким образом, в формализме твисторной теории некоторая нелокальность прослеживается даже у волновых функций единичных частиц, которые каким-то образом остаются цельными объектами («частицами»), несмотря на то что область распространения волновой функции может быть колоссальна — вплоть до многих световых лет, как в случае, когда мы перехватываем отдельные фотоны от звезд в далеких галактиках (см. раздел 2.6). Следует считать, что твисторная 1-функция действительно напоминает невозможный треугольник, раскинувшийся на таких огромных расстояниях. Наконец, как только мы находим частицу в конкретной точке, с невозможностью покончено; при этом совершенно не имеет значения, где именно удалось наблюдать конкретный фотон. 1-функция сработала, и данный фотон не удастся найти ни в каком ином месте.

В случае с многочастичными волновыми функциями ситуация тем более усложняется с учетом того, что твисторное описание волновой функции для n безмассовых частиц является (голоморфной) n-функцией. Кажется вероятным, что загадки с нарушением неравенств Белла (см. раздел 2.10) в состояниях с запутанностью n частиц удастся прояснить, исследовав такие твисторные описания. Тем не менее, насколько мне известно, серьезных попыток такого рода не предпринималось (см. [Penrose, 1998b, 2005, 2015a]).

Виттен, выдвигая в 2003 году свои инновационные идеи относительно твисторной теории струн, смог обойти такие проблемы когомологии, применив хитроумный «антивиковский поворот» в пространстве Минковского (ср. с разделом 1.9) и «повернув» одно пространственное измерение так, что оно становится временным; в результате получается плоское четырехмерное «пространство-время» с двумя времениподобными измерениями и двумя пространственноподобными. Тогда его (проективное) «твисторное пространство» оказывается *вещественным* проективным трехмерным пространством $\mathbb{RP}^3$, а не комплексным $\mathbb{CP}^3$, которое, по сути, является $\mathbb{PT}$. На первых порах это немного упрощает ситуацию, поскольку когомологии удается избежать, и δ-функции Дирака можно использовать примерно так, как это традиционно принято в стандартной квантовой механике (см. раздел 2.5). Однако, пусть я и вполне представляю себе пользу этой процедуры, меня не оставляет ощущение, что в ней упускается значительная часть потенциала твисторной теории, если заняться исследованием более глубоких физических проблем.

В исходных идеях Виттена было и заманчивое предположение о том, что римановы сферы (линии) в $\mathbb{PT}$, представляющие точки в $\mathbb{CM}$, должны обобщаться до римановых поверхностей более высокого порядка, например до конических сечений, кривых третьего и четвертого порядка и т. д. («струн», как в разделе 1.6), и позволять сформулировать процедуры, дающие возможность непосредственно

вычислить глюонное рассеяние. Подобные идеи выдвигались и ранее [Shaw and Hughston, 1990], но авторы намеревались применять их в существенно ином контексте. Эти идеи значительно подогрели интерес к теории, лежащей в основе глюонных рассеяний и их вычислительного вывода, причем эти исследования особенно заметным образом перекликались с первопроходческими работами, которыми намного раньше и практически в одиночку на протяжении 30 лет занимался Эндрю Ходжес; они связаны с развитием теории твисторных диаграмм [Hodges and Huggett, 1980; Penrose and MacCallum, 1972; Hodges, 1982, 1985a,b, 1990, 1998, 2006b], которые формально строятся по тем же принципам, что и диаграммы Фейнмана в стандартной физике частиц (см. раздел 1.5). Хотя в последние годы интерес к струнным аспектам этих новых разработок как будто немного поугас (или сместился в область так называемых *амбитвисторов*, то есть комбинированных твисторных/дуально-твисторных представлений комплексных нулевых геодезических [LeBrun, 1985, 1990]), разработки в области твисторов развивались сами по себе, а в последнее время появилось множество новых работ, в которых еще значительнее упрощаются глюонные рассеяния, так что удается рассчитывать все более и более сложные процессы. Среди этих примечательных новых разработок особенно интересны полезная концепция *импульсных твисторов* (исходно предложенная Эндрю Ходжесом) и *амплитуэдр* — концепция, введенная Нимой Аркани-Хамедом и разработанная на основе более ранних идей Ходжеса, обретающаяся в надразмерных версиях твисторного пространства («грассманианах»), которые, по-видимому, позволяют заново описать амплитуды рассеяния удивительно исчерпывающим образом (см., например, [Hodges, 2006a, 2013a, b; Bullimore et al., 2013; Mason and Skinner, 2009; Arkani-Hamed et al., 2010, 2014; Cachazo et al., 2014]).

Однако все эти захватывающие разработки по природе своей являются вариантами теории возмущений (пертурбативной теории), где интересующие нас величины вычисляются при помощи некоторых степенных рядов (см. разделы A.10, A.11, 1.5, 1.11 и 3.8). Какими бы мощными ни были эти методы, они не позволяют исследовать многие интересные аспекты. В этой категории наиболее интересны аспекты гравитационной теории, связанные, по сути, с кривизной пространства. Хотя многие задачи можно решить, подступившись к проблемам общей теории относительности со стороны степенных рядов и при этом с большой точностью дать возмущающие поправки к ньютоновской теории плоского пространства в условиях сравнительно слабых гравитационных полей, ситуация совершенно меняется, когда требуется как следует разобраться со свойствами черных дыр. Такая же ситуация должна складываться и с теорией твисторов, когда ее подходы к нелинейным полям, например полям теории Янга — Миллса, и общей теории относительности методами пертурбативной теории рассеяния практически не

позволяют коснуться того потенциала, который, на мой взгляд, должен скрываться в твисторной трактовке физики элементарных частиц.

Существует один мощный вариант применения твисторной теории в нелинейной физике, который, однако, до сих пор фундаментально неполон. Речь идет о попытке дать *не*перетурбативное описание основных нелинейных физических полей, эйнштейновской теории относительности, теории Янга — Миллса и взаимодействий из теории Максвелла. Такая неполнота, представлявшая собой фундаментальное и совершенно удручающее препятствие для разработки теории твисторов на протяжении почти 40 лет, связана с любопытной однобокостью вышеописанных представлений безмассовых частиц в виде твисторных функций, где обнаруживается странный дисбаланс между однородностями для левых и правых спиральностей. Это не так страшно, если ограничиться вышеприведенным описанием свободных линейных безмассовых полей (твисторных волновых функций). Однако до сих пор удавалось *не*пертурбативно трактовать нелинейные взаимодействия лишь для *левоориентированных* частей этих полей.

Действительно, одно из поразительных достоинств вышеприведенного твисторного описания безмассовых полей заключалось в том, что взаимодействия (и самовоздействия) этих полей также удавалось описать удивительно лаконично — это опять же касалось лишь левоориентированной части поля. Феномен «нелинейного гравитона», обнаруженный мною в 1975 году, позволяет построить *искривленное* твисторное пространство, в котором можно представить все левоориентированные решения эйнштейновских уравнений, и фактически описывает, как левоориентированные гравитоны взаимодействуют друг с другом. Примерно спустя год Ричард С. Уорд смог расширить эту процедуру таким образом, чтобы можно было работать с левоориентированными калибровочными (максвелловскими и янг-миллсовскими) полями электромагнитных, слабых или сильных взаимодействий [Penrose, 1976b; Ward, 1977, 1980]. Однако до сих пор не решена фундаментальная проблема, так называемая *googly-проблема* (в крикете «googly» называют ситуацию, когда мяч, который был подан так, что создается впечатление, что он был закручен вправо, на самом деле вращается влево, и наоборот). Проблема заключалась в том, чтобы найти процедуру для *правоориентированных* гравитационных и калибровочных взаимодействий, которые также подходили бы для нелинейного гравитона, так чтобы оба варианта можно было совместить и дать полную твисторную формулировку известных основополагающих физических взаимодействий.

Следует пояснить, что, если бы мы использовали *дуальное* твисторное пространство, то соотношение между ориентированностью спина и степенью гомогенности (теперь уже дуальной) твисторной функции просто изменилось бы на противо-

положное. Использование дуального твисторного пространства для обработки противоположной спиральности не решает googly-проблему, поскольку нам требуется однородная процедура для работы с *обеими* спиральностями одновременно — не в последнюю очередь потому, что придется иметь дело с безмассовыми частицами (например, с плоскополяризованными фотонами; см. раздел 2.5), то есть с ситуациями, где вовлечены квантовые суперпозиции обеих спиральностей одновременно. Разумеется, мы могли бы все время использовать не обычное, а дуальное твисторное пространство, но googly-проблема сохраняется. По-видимому, ключевая сложность до конца неразрешима в рамках исходного аппарата твисторной теории, где используются деформированные варианты твисторного пространства, несмотря на то что в последние годы в этой области были сделаны некоторые обнадеживающие разработки (см., например, [Penrose, 2000a]).

Однако в твисторной повестке дня за последние несколько лет появилась новая концепция, которую я называю *дворцовая твисторная теория*. Название связано с необычным источником ключевой затравочной идеи, породившей эту теорию; все началось с краткого разговора между мной и Майклом Атьей перед обедом в пафосных интерьерах Букингемского дворца [Penrose, 2015a, b]. Эта беседа открыла перед нами новые перспективы применения твисторных идей. Она связана со старинной чертой теории твисторов, игравшей важную роль во многих ранних разработках (хотя ранее по тексту я о ней прямо не упоминал). Это базовая взаимосвязь между твисторной геометрией и идеями квантовой механики, заключающаяся в процедуре квантования твисторов, причем твисторные переменные $\mathbf{Z}$ и $\overline{\mathbf{Z}}$ считаются каноническими сопряжениями друг друга (ранее подобные сопряжения уже упоминались — таковы, например, переменные координаты и импульса ($\mathbf{x}$ и $\mathbf{p}$) элементарной частицы; см. раздел 2.13); кроме того, эти сопряжения комплексные. В стандартной процедуре канонического квантования такие сопряженные переменные заменяются некоммутирующими операторами (см. раздел 2.13). Та же идея применяется и в различных наработках теории твисторов, появившихся за прошедшие годы [Penrose, 1968, 1975b; Penrose and Rindler, 1986], где такая некоммутативность ($\mathbf{Z}\overline{\mathbf{Z}} \neq \overline{\mathbf{Z}}\mathbf{Z}$) является физически естественной, а оба оператора, $\mathbf{Z}$ и $\overline{\mathbf{Z}}$, являются по отношению друг к другу операторами *дифференцирования* (см. раздел A.11). Новизна дворцовой теории твисторов заключается в том, что она инкорпорирует алгебру таких некоммутативных твисторных переменных в нелинейные геометрические конструкции — нелинейный гравитон и упоминавшаяся ранее конструкция калибровочного поля Уорда. Некоммутативная алгебра действительно кажется логичной с геометрической точки зрения — в контексте исследованных выше структур (она включает идеи из области некоммутативной геометрии и геометрического квантования) (см. [Connes and Berberian, 1995 и Woodhouse, 1991]). Представляется, что эта процедура в самом деле обеспечивает формализм, до-

статочно широкий для того, чтобы с его помощью можно было одновременно учесть как левоориентированные, так и правоориентированные спиральности; потенциально она также позволяет описать полностью искривленные пространства-времена таким образом, чтобы в нее легко вписывались вакуумные уравнения Эйнштейна (с Λ или без нее), но еще требуется выяснить, на самом ли деле она в полной мере позволяет всего этого достичь.

Можно отметить, что всплеск интереса к теории твисторов, заметный в последние годы и начавшийся с твисторно-струнных идей Виттена и др., оказался особенно полезен при изучении высокоэнергетических процессов. Особая роль твисторов в данном контексте обусловлена прежде всего тем фактом, что в таких обстоятельствах частицы можно считать *безмассовыми*. Теория твисторов определенно хороша для изучения безмассовых частиц, но этим ее польза ни в коем случае не ограничивается. На самом деле существуют различные идеи по поводу учета масс в данной схеме [Penrose, 1975b; Perjés, 1977, pp. 53–72, 1982, pp. 53–72; Perjés and Sparling, 1979; Hughston, 1979, 1980; Hodges, 1985b; Penrose and Rindler, 1986], но пока они, по-видимому, не играют никакой роли в этих новых разработках. Повышенный интерес явно привлекает следующий вопрос: как следует развивать в будущем теорию твисторов, чтобы учитывать массы (покоя) частиц?

4.2. Откуда берутся квантовые основания?

В разделе 2.13 я пытался продемонстрировать, что, независимо от того, насколько убедительно формализм квантовой механики подтверждается многочисленными экспериментами — ни один из этих экспериментов до сих пор не требует никаких модификаций квантовой механики, — все-таки существуют серьезные основания полагать, что эта теория неокончательная и что линейность, столь важная для наших современных представлений о квантовой теории, в конечном итоге может быть упразднена. В таком случае ее (взаимно несовместимые) составляющие **U** и **R** станут всего лишь отличными приближениями некоторой более целостной, всеобъемлющей модели. Еще в 1963 году сам Дирак красноречиво утверждал:

> Все согласны относительно формализма. Он работает настолько хорошо, что никто не в состоянии позволить себе не соглашаться с ним. Однако картина, которую мы представляем себе, так сказать, «позади» этого формализма, все еще является предметом разногласий. Мне бы хотелось высказать мысль, что не следует особенно огорчаться по поводу этих разногласий. Я уверен, что та стадия физики, которая достигнута на сегодняшний день, не является оконча-

тельной. Она представляет собой именно одну из стадий в эволюции нашей картины природы, и нам следует ожидать, что этот процесс эволюции будет продолжаться и в будущем точно так же, как будет продолжаться процесс биологической эволюции. Современная стадия физической теории — это только переходная ступень к более совершенным стадиям, которых мы достигнем в будущем. Можно выразить уверенность в том, что более совершенные стадии появятся просто потому, что современная физика сталкивается с трудностями.

Если согласиться с такой точкой зрения, то нужна некая подсказка: какую же форму могут приобрести законы усовершенствованной квантовой теории? Нужно как минимум знать условия экспериментов, обеспечив которые можно постепенно зафиксировать явные отклонения от прогнозов стандартной квантовой теории.

В разделе 2.13 я уже указывал, что это должны быть такие обстоятельства, в которых гравитация начинает оказывать серьезное воздействие на квантовые суперпозиции. Идея такова: как только достигается этот уровень, начинает проявляться некоторая нелинейная нестабильность, ограничивающая длительность таких суперпозиций, и по истечении некоторого периода времени суперпозиция разрешается в одну или другую альтернативу. Более того, берусь утверждать, что *все* случаи редукции квантового состояния (**R**) происходят именно таким **OR**-образом (объективная редукция).

Обычно подобные соображения парируют так: якобы сила гравитации настолько слаба, что никакие мыслимые в настоящее время лабораторные эксперименты не позволят выявить каких-либо квантовых свойств гравитации; тем более маловероятно, что вездесущий феномен редукции квантового состояния может произойти под действием абсурдно малых квантовых составляющих гравитационных сил или энергий — они гораздо меньше величин, которые могли бы фиксироваться в описываемых ситуациях. Судя по всему, возникает следующая дополнительная проблема: если мы ожидаем, что квантовая гравитация должна играть роль в обычном настольном квантовом эксперименте, то каким образом наш процесс редукции квантового состояния может высвобождать просто гигантскую (квантово-гравитационную) планковскую энергию E_{P} — порядок этой энергии должен примерно в 10^{15} раз превышать ту, которой могут достигать отдельные частицы в БАК (см. разделы 1.1 и 1.10) и сопоставим с энергией взрыва крупного артиллерийского снаряда? Более того, если мы полагаем, что эффекты квантовой гравитации на уровне фундаментальной структуры пространства-времени должны существенным образом отражаться на обычных физических явлениях, то нужно учитывать, что масштабы воздействия квантовой гравитации на обычные явления будут проявляться на расстояниях порядка планковской длины l_{P} и в периоды порядка планковского времени t_{P}, а эти величины столь

смехотворно малы (см. раздел 3.6), что вряд ли могут играть какую-либо заметную роль в обычной макроскопической физике.

Однако здесь я выступаю с совершенно иных позиций. Я в самом деле не пытаюсь доказать, что крошечные гравитационные силы, действующие в квантовых экспериментах, могут вызывать объективную редукцию, равно как не утверждаю, что в процессе квантовой редукции должна высвобождаться планковская энергия. Я просто указываю на то, что мы должны серьезно пересмотреть наши представления о квантовой физике, в полной мере учитывая при этом эйнштейновскую трактовку гравитации как явления, связанного с искривлением пространства-времени. Также нужно учитывать, что планковскую длину l_{P} и планковское время t_{P}

$$l_{\mathrm{P}} = \sqrt{\frac{\gamma\hbar}{c^3}}\,,\ t_{\mathrm{P}} = \sqrt{\frac{\gamma\hbar}{c^5}}$$

можно получить (как квадратные корни), перемножив две величины — гравитационную постоянную γ и (редуцированную) постоянную Планка $\hbar$ (обе эти величины крайне малы с точки зрения обыденного опыта), а затем разделив результат на очень большую величину — скорость света. Неудивительно, что эти расчеты приводят нас к масштабам непостижимо крошечных порядков, примерно в 10^{20} раз меньше тех, на которых происходят наиболее стремительные взаимодействия в фундаментальной физике частиц.

Однако гипотеза об объективной редукции состояния, которую я обосновывал в разделе 2.13 и которую теперь называю **OR** (в принципе, она похожа на гипотезу, выдвинутую несколькими годами ранее Лайошом Дьёши [1984, 1987, 1989], но без такой привязки к принципам общей теории относительности, которыми пользуюсь я [Penrose, 1993, 1996, 2000b, p. 266—282]), позволяет более реалистично описать масштаб времени, требующегося на редукцию состояния. Действительно, гипотетическая средняя длительность **OR** — $\tau \approx \hbar/E_{\mathbf{G}}$, время, необходимое для распада суперпозиции на два стационарных объекта, находящихся в разных местах (как показано на рис. 2.13), — требует рассматривать такие периоды, при вычислении которых фактически задействуется частное двух малых величин ($\hbar/\gamma$), а не произведение (причем скорость света в формулу даже не входит), и эта гравитационная (собственная) энергия E_{G} (в теории Ньютона) пропорциональна γ. Поскольку сейчас нет никаких причин полагать, что $\hbar/E_{\mathrm{G}}$ может быть особенно велико или особенно мало, нам в каждом конкретном случае нужно тщательно проверять, приводит ли нас данная формула к таким временным масштабам, на которых могли бы происходить реальные физические процессы, подчиняющиеся объективной редукции квантового состояния. Также можно отметить, что планковская энергия

$$E_{\text{Р}} = \sqrt{\frac{\hbar c^5}{\gamma}}$$

хотя и зависит все от того же частного $\hbar/\gamma$, имеет в числителе высокую степень скорости света, из-за чего порядок этой энергии сильно увеличивается.

Так или иначе, в теории Ньютона расчеты собственной гравитационной энергии всегда дают выражения, пропорциональные γ, поэтому мы убеждаемся, что τ масштабируется пропорционально частному $\hbar/\gamma$. Поскольку γ мала, величина E_{G} определенно будет исчезающе малой в тех экспериментальных ситуациях, которые я собираюсь рассматривать (в частности, из-за распределения важного в данной ситуации *сдвига массы*; общая масса в таком разностном распределении будет равна нулю). В таком случае можно ожидать, что период распада квантовой суперпозиции должен быть очень длительным — пропорциональным γ^{-1}, — и это вполне согласуется с тем фактом, что в стандартной квантовой механике время существования квантовой суперпозиции должно быть бесконечным (в соответствии с пределом $\gamma \to 0$). Но в таком случае необходимо учитывать, что величина $\hbar$ также очень мала с обыденной точки зрения, так что не стоит списывать со счетов и вариант, при котором отношение $\hbar/E_{\text{G}}$ может иметь вполне измеримое значение. Выражаясь иначе, в естественных (планковских) единицах (см. раздел 3.6) секунда — крайне длительный период, равный примерно 2×10^{43}. Поэтому, чтобы дать измеримый эффект, фиксируемый, скажем, в течение нескольких секунд, рассматриваемая гравитационная собственная энергия в естественных единицах действительно должна быть совсем крошечной.

Здесь также важно отметить, что в нашем выражении $\hbar/E_{\text{G}}$ нет скорости света c. Поэтому ситуация упрощается: можно рассматривать случаи, в которых сдвиг массы происходит очень медленно. С этим связаны различные практические преимущества, но есть и теоретические преимущества, поскольку нам более не приходится вдаваться в частности эйнштейновской общей теории относительности и можно ограничиться в целом ньютоновскими трактовками. Более того, можно временно абстрагироваться от всех сложностей, связанных с нелокальными аспектами *физически реальной* редукции квантового состояния, которые нарушают «причинно-следственные связи» (как может показаться в столь непостижимых ЭПР-ситуациях, какие были рассмотрены в разделе 2.10), так как в теории Ньютона скорость света, будучи бесконечной, абсолютно не ограничивает скорость передачи сигнала и гравитационные взаимодействия можно считать мгновенными.

Я рассматриваю **OR**-гипотезу в минималистической форме (с минимальным числом дополнительных допущений), когда перед нами примерно равноампли-

тудная суперпозиция пары состояний, каждое из которых само по себе будет стационарным. Эта ситуация рассматривалась в разделе 2.13, где я в общих чертах обосновал, почему должен существовать временной масштаб τ, ограничивающий вероятную продолжительность существования такой суперпозиции, и по истечении этого времени суперпозиция спонтанно разрешается в одну из альтернатив, а время распада рассчитывается по формуле

$$\tau \approx \frac{\hbar}{E_G},$$

где E_G — гравитационная собственная энергия *разности* между распределением масс в первом и во втором состояниях, которые образуют суперпозицию. Если сдвиг — это просто жестко заданный переход из одного места в другое, то можно дать более простое определение E_G, приведенное в разделе 2.13, а именно описать *энергию взаимодействия*. Это будет энергия, затрачиваемая на такой сдвиг, при условии, что на каждое из состояний в суперпозиции воздействует лишь *гравитационное* поле другого.

В обобщенном виде гравитационная собственная энергия гравитационно-связанной системы — это энергия, которая потребовалась бы, чтобы разнести гравитационно-связанные компоненты системы на бесконечность, игнорируя все остальные силы; гравитация рассматривается в рамках теории Ньютона. Так, гравитационная собственная энергия однородного шара массы m и радиуса r описывается формулой $3m^2\gamma / 5r$. Для определения E_G в рассматриваемой здесь ситуации мы вычисляли бы эту величину исходя из теоретического распределения масс, получаемого путем *вычитания* распределения масс одного такого состояния из распределения масс другого, поэтому результирующее распределение масс в одних областях получилось бы положительным, а в других — отрицательным (рис. 4.8), что не является нормальной ситуацией при расчете гравитационных собственных энергий!

В качестве примера рассмотрим приведенный выше случай с однородным шаром (массой m и радиусом r), помещенным в (примерно равноамплитудную) *суперпозицию* двух горизонтально сдвинутых местоположений, центры которых разделены отрезком q. (Ньютоновский) расчет E_G (собственной энергии *разницы* двух распределений масс) дает следующий результат:

$$E_G = \begin{cases} \dfrac{m^2\gamma}{r}\left(2\lambda^2 - \dfrac{3\lambda^3}{2} + \dfrac{\lambda^5}{5}\right), & 0 \le \lambda \le 1, \\ \dfrac{m^2\gamma}{r}\left(\dfrac{6}{5} - \dfrac{1}{2\lambda}\right), & 1 \le \lambda, \end{cases} \quad \text{где } \lambda = \frac{q}{2r}.$$

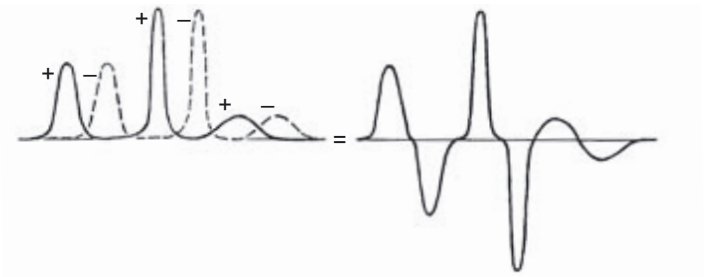


Рис. 4.8. E_G — это гравитационная собственная энергия распределения масс, получаемая вычитанием ожидаемого распределения масс одного квантового состояния в составе суперпозиции из распределения масс другого. Для каждого отдельно взятого состояния можно обнаружить, что в некоторых областях наблюдается сильная концентрация масс (например, вокруг атомных ядер), и из-за таких неравномерностей получается «лоскутное одеяло» из участков с положительной и отрицательной массой; поэтому величина E_G получается относительно большой

Очевидно, что с увеличением расстояния q возрастает и значение E_G, достигающее

$$\frac{7}{10}\times\frac{m^2\gamma}{r},$$

когда два экземпляра шара примыкают друг к другу ($\lambda = 1$); при дальнейшем увеличении q рост E_G замедляется, достигая предельного общего значения

$$\frac{6}{5}\times\frac{m^2\gamma}{r}$$

при бесконечном удалении ($\lambda = \infty$). Следовательно, основное воздействие E_G проявляется на этапе возрастающего разделения, от совпадения к контакту, и лишь немного меньший прирост может быть достигнут при дальнейшем разделении.

Разумеется, никакое реальное вещество по структуре своей не будет совершенно однородным, масса вещества будет сконцентрирована преимущественно в атомных ядрах. Это позволяет предположить, что значительного эффекта можно достичь уже при минимальном сдвиге, при квантовой суперпозиции материального тела в двух местах; в таком случае атомные ядра потребуется сдвинуть лишь на расстояние, равное диаметру такого ядра. Вполне возможно, что так и есть, но существует важное осложнение: ведь мы, в сущности, рассматриваем *квантовые* состояния, поэтому ядра должны быть «размазаны» в соответствии с принципом неопределенности Гейзенберга (см. раздел 2.13). В противном случае мы могли бы озаботиться тем, что нам требуется рассматривать даже не ядра, а протоны и нейтроны, из которых они состоят, либо даже отдельные кварки, образующие

протоны и нейтроны. Поскольку кварки, подобно электронам, также рассматриваются как точечные частицы, то есть их радиус r в вышеприведенной формуле будет нулевым, мы, по-видимому, получим $E_G \approx \infty$ при $\tau \approx 0$, а это подталкивает нас к выводу о том, что почти все суперпозиции должны мгновенно коллапсировать (см. также [Ghirardi et al., 1990]). Следовательно, согласно такой гипотезе, никакой квантовой механики быть не должно!

Да, мы *обязаны* принимать в расчет «квантовое размазывание», если всерьез исследуем $\tau \approx \hbar / E_G$. Как мы помним из раздела 2.13, принцип неопределенности Гейзенберга гласит: чем точнее известно положение частицы, тем неопределеннее ее импульс. Соответственно, невозможно ожидать, что локализованная частица останется стационарной, а в излагаемых здесь рассуждениях ее стационарность — обязательное требование. Разумеется, когда имеешь дело с протяженными телами, с которыми мы будем работать, вычисляя E_G, приходится учитывать положения огромных совокупностей частиц, каждая из которых внесет свой вклад в стационарное состояние такого тела. Нам придется вывести для этого тела стационарную волновую функцию ψ, а затем вычислить так называемое *ожидаемое значение* плотности масс в каждой точке (стандартная квантово-механическая процедура), и это даст нам ожидаемое распределение массы для всего тела. Эта процедура будет проделываться с телом в обоих местоположениях, участвующих в квантовой суперпозиции, и одно из этих (ожидаемых) распределений масс потребуется вычесть из другого, вычислив таким образом требуемую гравитационную собственную энергию E_G (здесь вновь всплывает техническая проблема, обозначенная еще в разделе 1.10, — строго говоря, каждое квантовое состояние должно простираться на всю Вселенную, но эту проблему можно обойти при помощи невинной уловки, а именно трактовать центр масс с *классической* точки зрения. Мы вернемся к этой проблеме позже в данном разделе).

Далее логично спросить, можно ли подвести под такую **OR**-гипотезу рациональное обоснование более серьезное, чем общие соображения, приведенные в разделе 2.13. Идея такова, что базовая нестыковка между принципами общей теории относительности Эйнштейна и законами квантовой механики может быть снята, лишь если удастся принципиально изменить сами эти базовые законы. В данном случае я скорее склонен доверять базовым принципам общей теории относительности и сомневаться в фундаментальных принципах квантовой механики. В большинстве трактовок квантовой гравитации акценты расставлены иначе. Среди физиков более распространено мнение, что при таком столкновении принципов придется отказаться именно от общей теории относительности, поскольку она подтверждается на экспериментах не столь убедительно, как квантовая механика. Я собираюсь отстаивать практически противоположную точку зрения и считаю эйнштейновский *принцип эквивалентности* (см. разделы 1.12 и 3.7) более фундаментальным,

чем квантовый принцип *линейной суперпозиции*, — в основном потому, что именно этот аспект квантового формализма уводит нас в область парадоксов, когда мы пытаемся применять квантовую механику на уровне макроскопических объектов (см., например, парадокс кота Шрёдингера, разделы 1.4, 2.5 и 2.11).

Читатель вполне может вспомнить принцип эквивалентности (Галилея—) Эйнштейна, согласно которому локальное воздействие гравитационного поля неотличимо от действия ускорения — иными словами, локальный наблюдатель, свободно падающий под действием гравитации, не будет ее ощущать. Этот тезис можно сформулировать иначе: сила гравитации, действующая на тело, пропорциональна инертной массе этого тела (сопротивлению его ускорению) — таким свойством не обладает больше ни одна из сил природы. Сегодня всем известно, что космонавты в орбитальном полете на космической станции находятся в состоянии невесомости и не ощущают никакого притяжения. Об этом принципе были хорошо осведомлены Галилей и Ньютон, а Эйнштейн заложил его в основу своей общей теории относительности.

Теперь давайте вообразим квантовый эксперимент, в котором учитывается гравитационное воздействие Земли. Можно предвосхитить две различные процедуры — назову их *перспективами*, — позволяющие вписать гравитационное поле Земли в этот эксперимент. Первая перспектива, более бесхитростная, называется ньютоновской; в таком случае мы полагаем, что гравитационное поле просто порождает направленную вниз силу $m\mathbf{a}$, воздействующую на любую частицу массы m (вектор гравитационного ускорения $\mathbf{a}$ в данном случае считается постоянным как в пространстве, так и во времени). Ньютоновские координаты обозначим $(\mathbf{x}, t)$, где трехмерный вектор $\mathbf{x}$ означает положение в пространстве, а t — время. На стандартном языке квантовой механики мы в таком случае должны были бы трактовать гравитационное поле в рамках стандартной квантовой процедуры и говорили бы о нем: «приплюсовываем член гравитационного потенциала к гамильтониану» — в соответствии с той самой процедурой, которая бы применялась и к любому другому физическому взаимодействию. Альтернативная эйнштейновская перспектива потребовала бы представить, что описание делается относительно пространственно-временных координат $(\mathbf{X}, T)$, которыми оперирует наблюдатель, находящийся в состоянии свободного падения, — таким образом, с точки зрения этого наблюдателя, гравитационное поле Земли обнуляется. Затем данное описание вновь переформулируется для экспериментатора, находящегося в покоящейся лаборатории (рис. 4.9). Между двумя системами координат сформируется отношение

$$\mathbf{x} = \mathbf{X} + \frac{1}{2}t^2\mathbf{a}, \quad t = T.$$

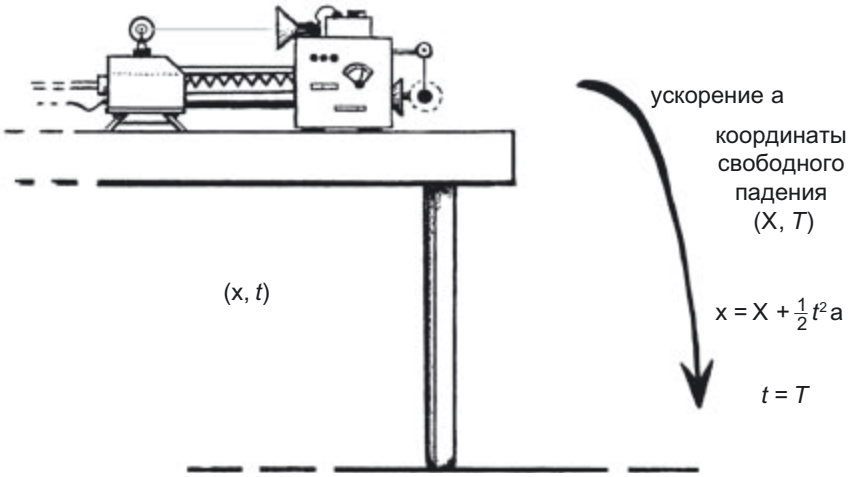


Рис. 4.9. Ставится (причудливый) квантовый эксперимент, в котором участвует сила тяготения Земли. В (традиционной) ньютоновской трактовке гравитационного поля Земли используются координаты (**x**, *t*), зафиксированные в лабораторных условиях; при этом сила притяжения Земли трактуется точно так же, как и любая другая сила. В эйнштейновской перспективе рассматриваются координаты (**X**, *T*), в которых гравитационное поле Земли обнуляется

Оказывается [Penrose, 2009a, 2014a; Greenberger and Overhauser, 1979; Beyer and Nitsch, 1986; Rosu, 1999; Rauch and Werner, 2015], что соотношение волновой функции ψ_E (см. разделы 2.5–2.7), с которой мы имеем дело при эйнштейновской перспективе, и волновой функции ψ_N из ньютоновской перспективы описывается следующей формулой:

$$\psi_E = \exp\left(i\left(\frac{1}{6}t^3\mathbf{a}\cdot\mathbf{a} - t\bar{\mathbf{x}}\cdot\mathbf{a} \right)\frac{M}{h} \right)\psi_N$$

(при правильно выбранных координатах), где M — общая масса рассматриваемой квантовой системы, а $\bar{\mathbf{x}}$ — ньютоновский радиус-вектор центра масс этой системы. Расхождение представляет собой обычный фазовый множитель, поэтому не должно давать никаких заметных различий между двумя перспективами (см. раздел 2.5) — или должно? В рассматриваемой здесь ситуации два описания действительно должны быть эквивалентны, и существует хорошо известный эксперимент, впервые поставленный в 1975 году [Colella et al., 1975; Colella and Overhauser, 1980; Werner, 1994; Rauch and Werner, 2015]. Этот эксперимент демонстрирует согласованность между двумя перспективами, тем самым подтверждая, что квантовая механика *не противоречит* эйнштейновскому принципу эквивалентности в данном контексте.

Однако следует отметить любопытный факт: этот фазовый множитель содержит член

$$\frac{1}{6}t^3\mathbf{a}\cdot\mathbf{a}$$

в показателе экспоненты (умноженный на коэффициент $\mathrm{i}M/\hbar$), что подводит нас к важному следствию: если рассматривать лишь те решения уравнений Шрёдингера, которые предполагают положительную энергию («физические» решения с *положительной частотой*; см. раздел 4.1), отделив их от тех, что предполагают отрицательную энергию («нефизические» решения), то обнаружится расхождение между эйнштейновскими и ньютоновскими волновыми функциями. Соображения квантовой теории поля (актуальные и в обычной квантовой механике; см. [Penrose, 2014a]) позволяют предположить, что эйнштейновская и ньютоновская перспективы дают нам два разных вакуума (см. разделы 1.6 и 3.9), так что выстраиваемые на основе двух этих перспектив гильбертовы пространства в определенном смысле несовместимы друг с другом, и мы не можем согласованно складывать векторы состояний из одного такого пространства с векторами состояний из другого.

Фактически речь идет всего лишь об *эффекте Унру* в пределе $c \to \infty$, вкратце описанном в разделе 3.7 и обычно рассматриваемом в контексте черных дыр, где движущийся с ускорением в вакууме наблюдатель должен ощущать *температуру* $\hbar a/2\pi kc$ (то есть всего $a/2\pi$ в естественных единицах). Вакуум, фиксируемый ускоряющимся наблюдателем, — это так называемый *тепловой вакуум*, дающий ненулевую температуру окружающей среды, которая в данном случае принимает значение $\hbar a/2\pi kc$. Если рассмотреть эту ситуацию в контексте ньютоновского предела $c \to \infty$, то температура опускается до нуля, но два вакуума (моделируемые с ньютоновской и с эйнштейновской точки зрения) остаются *разными*, поскольку есть указанный выше нелинейный фазовый множитель, сохраняющийся и при применении предела $c \to \infty$ к вакууму Унру[1].

Это не вызывало бы никаких сложностей, если бы, как в данном случае, мы рассматривали всего одно фоновое гравитационное поле, например гравитационное поле Земли, и все состояния, оказывающиеся в суперпозиции, обладали бы одним и тем же вакуумным состоянием вне зависимости от того, идет ли речь о ньютоновской или эйнштейновской перспективе. Однако предположим, что рассматривается ситуация, где есть суперпозиция двух гравитационных полей. Такой случай исследуется в лабораторном эксперименте с суперпозицией массивного объекта

[1] Благодарю Бернарда Кэя, расчеты которого помогли мне подтвердить эту гипотетическую версию.

в двух разных местах (см. рис. 2.28 в разделе 2.13). Слабенькое гравитационное поле этого объекта будет слегка различаться в обоих местах, и в квантовом состоянии, описывающем суперпозицию двух местоположений объекта, должна учитываться и суперпозиция двух этих гравитационных полей. В таком случае уже *не все равно*, какую именно перспективу мы берем в рассмотрение.

В данной ситуации в составе общего гравитационного поля также учитывается вклад гравитационного поля Земли, но при расчете *разницы*, которая, как мы убедимся, необходима при вычислении E_G, оказывается, что гравитационное поле Земли обнуляется, и остается учитывать лишь гравитационное поле объекта, претерпевшего *квантовый сдвиг*, которое вносит свой вклад в E_G. Однако в таком обнулении есть один тонкий момент, на котором следует остановиться отдельно. Когда рассматриваемый массивный объект сдвигается в некотором направлении, Земля должна сдвигаться в противоположном направлении, так, чтобы центр гравитации в системе Земля — объект оставался неизменным. Разумеется, Земля сдвигается на исчезающе малую величину, ведь Земля неизмеримо массивнее рассматриваемого объекта. Но сама массивность Земли, возможно, подводит нас к вопросу: может ли прямо в эту минуту сдвиг Земли давать значительный вклад в E_G? К счастью, проверив детали, мы убедимся в том, что обнуление действительно происходит и вклад сдвига Земли в E_G можно полностью игнорировать.

Однако почему мы вообще должны рассматривать величину E_G? Если принять ньютоновскую перспективу и трактовать гравитационные поля исходя из нее, то не составит труда интерпретировать и квантовую суперпозицию двух местоположений объекта, и гравитационное поле будет учитываться по такому же принципу, что и любое другое, в соответствии с обычными квантово-механическими процедурами; такая перспектива допускает и линейную суперпозицию гравитационных состояний, причем возникает всего один вакуум. Правда, мне кажется, что с учетом солидных экспериментальных подтверждений общей теории относительности в мегамасштабах мы должны сделать ставку на эйнштейновскую перспективу, которая в конечном итоге должна значительно точнее соответствовать законам природы, нежели ньютоновская. Тогда мы вынуждены склониться к точке зрения, что при суперпозиции двух рассматриваемых гравитационных полей оба поля должны интерпретироваться согласно эйнштейновской перспективе. В таком случае мы пытаемся сложить суперпозицию двух состояний, относящихся к двум разным вакуумам, то есть к двум несовместимым гильбертовым пространствам, — а такие суперпозиции считаются *недопустимыми* (см. раздел 1.16 и начало раздела 3.9).

Необходимо рассмотреть эту ситуацию немного подробнее, и в данном случае речь пойдет о субмикроскопических явлениях в основном на масштабе атомных

ядер физического тела, вступающего в суперпозицию, но также и за пределами этого объекта. Хотя в нескольких предыдущих абзацах мы рассуждали о гравитационном поле ускорения **a**, *постоянного* в пространстве, можно предположить, что эти выкладки останутся актуальными (хотя бы приблизительно) и локальными в обширных преимущественно пустых областях, где будет складываться суперпозиция двух разных гравитационных полей. В данном случае я пытаюсь аргументировать, что каждое из этих полей в отдельности должно трактоваться согласно эйнштейновской перспективе, так что структурно физика этого примера включает «недопустимую» суперпозицию квантовых состояний, относящихся к двум разным гильбертовым пространствам. Состояние свободного падения в одном гравитационном поле будет соотноситься с состоянием свободного падения в другом на уровне фазового множителя, с которым мы сталкивались ранее, и будет включать член, нелинейный во времени t в показателе экспоненты $e^{iMQt^3/\hbar}$ для некоторого Q, как $\frac{1}{6}\mathbf{a}\cdot\mathbf{a}$ ранее. Но поскольку теперь мы рассматриваем переход от одного состояния свободного падения (с вектором ускорения $\mathbf{a}_1$) в другое состояние свободного падения (с вектором ускорения $\mathbf{a}_2$), наше Q принимает вид $\frac{1}{6}(\mathbf{a}_1-\mathbf{a}_2)\cdot(\mathbf{a}_1-\mathbf{a}_2)$, а не просто $\frac{1}{6}a\cdot a$, как ранее, поскольку в данном случае важна разность $\mathbf{a}_1-\mathbf{a}_2$ между полями объекта, расположенными в разных местах, а *сами* величины ускорений $\mathbf{a}_1$ и $\mathbf{a}_2$ лишь *условно* важны относительно Земли как системы отсчета.

На самом деле и $\mathbf{a}_1$, и $\mathbf{a}_2$ теперь являются функциями местоположения, но я предполагаю, что как минимум с хорошим приближением в любой небольшой локальной области именно этот член (Q) — корень проблемы. Суперпозиция состояний из столь различных гильбертовых пространств (то есть с разными вакуумами) технически недопустима, и теперь между первым и вторым состояниями в локальных областях возникает фазовый множитель

$$\exp\left(\frac{iM(\mathbf{a}_1-\mathbf{a}_2)^2t^3}{6\hbar}\right).$$

Таким образом, состояния относятся к несовместимым гильбертовым пространствам, пусть даже разность $\mathbf{a}_1-\mathbf{a}_2$ между двумя значениями ускорения свободного падения почти наверняка будет совсем крошечной в любом мыслимом эксперименте.

Строго говоря, феномен альтернативных вакуумов взят из КТП, а не из обсуждаемой здесь нерелятивистской квантовой механики, но данная проблема непосредственно важна и для последней. Стандартная квантовая механика

требует, чтобы энергия оставалась положительной (то есть частоты должны оставаться положительными), но в обычной квантовой механике это обычно не представляет проблемы (по причине технического характера: нормальная квантовая механика управляется положительно определенным гамильтонианом, который будет сохранять такую положительность). Однако в описываемом случае ситуация не такая, и мы, по-видимому, вынуждены нарушать данное условие, если только не держать вакуумы отдельно друг от друга (то есть не допускать сложения векторов состояний из разных гильбертовых пространств или суперпозиции таких векторов) (см. [Penrose, 2014a]).

Следовательно, мы, по-видимому, выходим за пределы нормального аппарата квантовой механики, и однозначного пути далее не просматривается. На данном этапе я предлагаю придерживаться той же линии, которую мы избрали в разделе 2.13, а именно уходить от парадокса с суперпозицией вакуумов из разных гильбертовых пространств и просто пытаться оценить *погрешность*, возникающую из-за игнорирования этой проблемы. Как и ранее (см. раздел 2.13), проблема возникает с членом $(\mathbf{a}_1 - \mathbf{a}_2)^2$, и предлагается просуммировать эту величину по всему трехмерному пространству, то есть взять ее пространственный интеграл, который и использовать в качестве величины погрешности, связанной с игнорированием недопустимых суперпозиций. Неопределенность, возникающая при игнорировании этой проблемы, вновь приводит нас к величине E_G как к мере энергетической неопределенности, присущей системе, как и ранее в разделе 2.13 [Penrose, 1996].

Для того чтобы можно было прикинуть длительность периода, в течение которого могла бы существовать наша суперпозиция, прежде чем стали бы возникать математические противоречия, обусловленные недопустимостью такой суперпозиции, можно воспользоваться принципом неопределенности времени-энергии $\Delta E \Delta t \geq \frac{1}{2}\hbar$, предложенным Гейзенбергом, который позволяет оценить возможную длительность суперпозиции, где $\Delta E \approx E_G$ (и опять точно так же, как в разделе 2.13). Просто ситуации именно такого рода возникают в нестабильном атомном ядре, которое распадается по истечении некоторого отрезка времени τ. В разделе 2.13 мы, в сущности, приравнивали τ к $\Delta\tau$ в соотношении Гейзенберга, поскольку именно благодаря этой неопределенности атом может вообще распасться в течение конечного промежутка времени. Следовательно, мы обязательно находим фундаментальную неопределенность энергии ΔE (или, по эйнштейновскому уравнению $E = Mc^2$, неопределенность массы $c^2\Delta M$), которая приблизительно соотносится с временем распада τ по формуле Гейзенберга, то есть $\tau \approx \hbar/2\Delta E$. Следовательно (игнорируя числовые множители), мы вновь получаем

$$\tau \approx \frac{\hbar}{E_G}$$

в качестве оценки возможной продолжительности суперпозиции, как и в разделе 2.13.

Несмотря на изложенные ранее комментарии, согласно которым эта гипотеза описывает *объективное* событие **R** (то есть **OR**), которое могло бы происходить в «обычных» временных масштабах с немикроскопическими объектами, здесь и в самом деле легко усмотреть прямую связь с планковским временем и планковской длиной. На рис. 4.10 я попытался набросать историю такого **OR**-события в пространстве-времени, где комок вещества оказывается в квантовой суперпозиции в двух разных местах. На рисунке изображено пространство-время, претерпевающее постепенную бифуркацию, прежде чем происходит **OR**-событие.

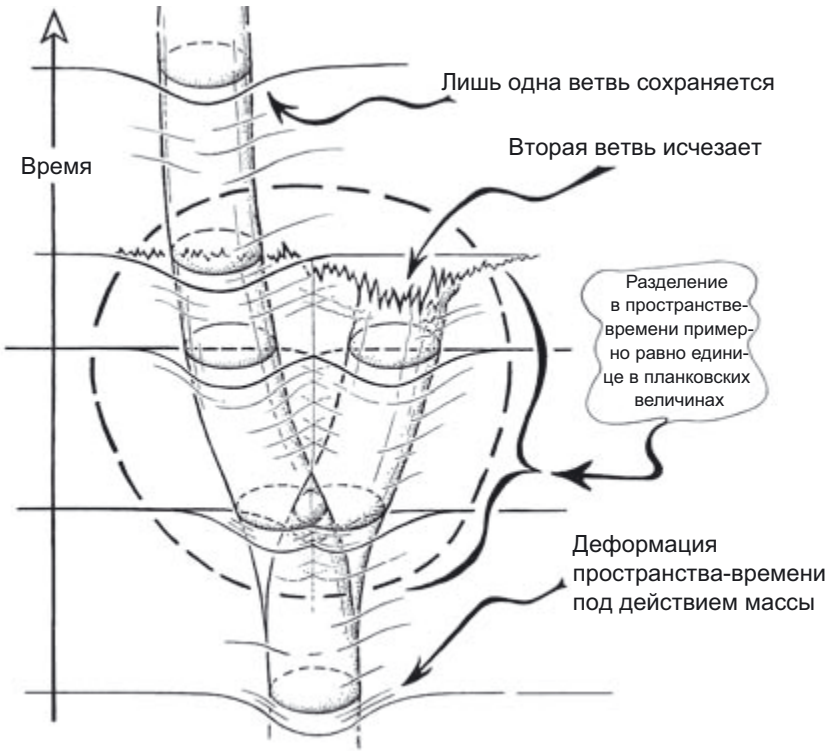


Рис. 4.10. Модель пространства-времени, иллюстрирующая, как квантовая суперпозиция двух разных сдвигов массивного объекта приводит к существенному расхождению находящихся в суперпозиции пространств-времен, каждое из которых деформируется по-своему, в зависимости от местоположения комка вещества. Согласно гравитационной **OR**-гипотезе, один из вариантов пространства-времени «исчезает» примерно к тому моменту, как разрыв между двумя составляющими достигает порядка единицы в планковских величинах

В момент **OR**-события как такового одно из направлений бифуркации исчезает, и остается единственное пространство-время, где и находится комок вещества. Я обозначил ограниченную область пространства-времени, где возникает бифуркация, прежде чем ее обрывает **OR**-процесс. Связь с планковскими единицами обусловлена тем фактом, что в данной гипотезе четырехмерный объем, в рамках которого сохраняется бифуркация, примерно равен *единице* в планковских величинах! Следовательно, чем меньше *пространственный* разрыв при ветвлении пространства-времени, тем дольше продержится бифуркация, и наоборот, чем больше этот разрыв, тем краткосрочнее период бифуркации. (Однако величину пространственно-временного разрыва следует понимать как соответствующую симплектическую меру в пространстве пространств-времен, что непросто осознать; правда, таким образом можно вывести довольно грубое приближение оценки $\tau \approx \hbar/E_G$) [Penrose, 1993, p. 179—189] (см. также [Hameroff and Penrose, 2014]).

В данном случае логично спросить: а есть ли какие-нибудь эмпирические данные, которые либо подтверждают, либо опровергают эту гипотезу? Легко представить себе ситуации, где $\hbar/E_G$ — либо очень долгий, либо очень краткий период. Например, в случае с котом Шрёдингера, рассматриваемом в разделах 1.4 и 2.13, сдвиг массы между местоположениями кота у *A*-дверцы и *B*-дверцы был бы более чем достаточен, чтобы период τ оказался предельно кратким (гораздо короче планковского времени 10^{-43} с), так что спонтанный переход из любой кошачьей суперпозиции, в сущности, получился бы мгновенным. Напротив, в других экспериментах, где мы имеем дело с квантовыми суперпозициями нейтронов, расположенных в разных местах, значение τ было бы колоссальным (астрономическим). Аналогичная ситуация складывалась бы даже с фуллеренами C_{60} и C_{70} (фуллерен — молекула, состоящая из 60 или 70 атомов углерода); считается, что это самые крупные объекты, для которых уже удалось зафиксировать квантовую суперпозицию двух разных местоположений [Arndt et al., 1999], а срок, в течение которого эти молекулы находились в суперпозиции отдельно друг от друга, составил бы чрезвычайно малую долю секунды.

Фактически в обеих этих ситуациях необходимо учитывать, что рассматриваемое квантовое состояние вполне может быть не изолировано от окружающего пространства, поэтому в нем может участвовать значительное количество лишней материи, *окружение* системы, которое, вероятно, должно вступать в квантовую запутанность с рассматриваемым состоянием. Соответственно, при сдвиге масс в данной суперпозиции также придется учитывать сдвиг всех масс в этой возмущенной среде, и именно такой сопутствующий сдвиг (в котором участвует огромное количество частиц, движущихся во всевозможных направлениях) зачастую будет давать основной вклад в E_G. Проблема декогеренции с окружающей

средой играет важную роль в большинстве традиционных трактовок унитарной эволюции (**U**) квантовой системы, которая, как считается, приводит к редукции квантового состояния (**R**) по правилу Борна (см. раздел 2.6). Идея такова, что окружающая среда неконтролируемым образом влияет на рассматриваемую квантовую систему, а процедура, описанная в разделе 2.13, — это среднее от всех степеней свободы окружающей среды. При этом свойства состояния квантовой суперпозиции фактически таковы, как будто оно представляет собой смесь вероятностей всех причастных альтернатив. Хотя в разделе 2.13 я и утверждал, что вовлечение беспорядочной окружающей среды в квантовое состояние отнюдь не решает математический квантово-механический парадокс, такое вовлечение должно играть важную роль в рассматриваемой здесь **OR**-модификации стандартной квантовой теории. Как только окружающая среда начинает всерьез запутываться с квантовым состоянием, сразу накапливается существенный сдвиг масс, вызванный запутанностью системы с окружающей средой, и возникает большая величина E_G, достаточная для стремительной спонтанной редукции состояния в одну или другую альтернативу, участвующую в квантовой суперпозиции (эта идея во многом основана на более ранней **OR**-гипотезе, предложенной Гирарди с коллегами [Ghirardi et al., 1986]).

На момент написания книги еще не было поставлено ни одного достаточно чувствительного эксперимента, который позволил бы подтвердить или опровергнуть эту гипотезу. Вклад окружающей среды в E_G очень мал, что не оставляет надежд зафиксировать какие-либо его эффекты. В настоящее время реализуется несколько проектов [Kleckner et al., 2008, 2015; Pikovski et al., 2012; Kaltenbaek et al., 2012], которые в конечном итоге должны как-то прояснить этот вопрос. Единственный подобный эксперимент, в котором довелось участвовать мне [Marshall et al., 2003; Kleckner et al., 2011], разрабатывался под руководством Дирка Баувместера, сотрудника университетов Лейдена и Санта-Барбары. В этом эксперименте крошечное зеркало, представляющее собой куб около 10 микрометров в поперечнике (10^{-5} м — человеческий волос примерно в десять раз толще), помещалось в квантовую суперпозицию двух местоположений, причем расстояние между этими местоположениями сравнимо по размеру с диаметром атомного ядра. Экспериментаторы намеревались удержать эту суперпозицию в течение нескольких секунд или минут, а затем вернуть ее в исходное состояние и проверить, на самом ли деле неизбежно теряется любая фазовая когерентность.

Для достижения такой суперпозиции квантовое состояние единичного фотона сначала расщепляется при помощи светоделителя (раздел 2.3). Затем одна часть волновой функции фотона воздействует на крошечное зеркало таким образом, что, под действием импульса фотона, это зеркало немного сдвигается (предполо-

жительно, на величину, сравнимую с диаметром атомного ядра), причем зеркало в этот момент аккуратно подвешено на консольке. Поскольку волновая функция фотона расщеплена надвое, состояние зеркала оказывается в суперпозиции, одновременно в сдвинутом и несдвинутом состоянии, такой маленький котик Шрёдингера. Однако если речь идет о фотоне видимого света, то единичного воздействия для этого будет отнюдь не достаточно, поэтому один и тот же фотон должен бить по зеркалу снова и снова, примерно миллион раз, и для этого его будет отражать другое (полусферическое) зеркало. Такого многократного воздействия может хватить, чтобы зеркало сдвинулось на отрезок, сравнимый по размеру с диаметром атомного ядра, а возможно, и больше, за период порядка нескольких секунд.

С теоретической точки зрения не вполне понятно, насколько тонким в таком случае должно быть распределение масс. Поскольку каждый из двух элементов суперпозиции сам по себе считается стационарным, распределение масс будет условно «размазано», причем степень такого размазывания должна зависеть от используемого вещества. В стационарных решениях уравнения Шрёдингера распределение масс обязательно должно получаться несколько размазанным — это следует из принципа неопределенности Гейзенберга (так что E_G *не должно* рассчитываться таким образом, словно частицы точечные, и это хорошо, поскольку, как отмечалось выше, вычисления с точечными частицами давали бы бесконечное E_G). Равномерно размазанное распределение масс, вероятно, нам также не подойдет (с экспериментальной точки зрения этот случай максимально *не*благоприятен, поскольку дает минимально возможное E_G для заданных размера/формы и разделения исследуемой массы). Для точной оценки E_G потребовалось бы хотя бы приблизительно решить стационарное уравнение Шрёдингера, так, чтобы можно было оценить ожидаемое распределение масс. Чтобы такой эксперимент удался, нужно идеально изолировать систему от всех вибраций, а кроме того, охладить ее до экстремально низких температур и держать в почти абсолютном вакууме. Самое главное, что качество зеркал должно быть превосходным.

Существует техническая проблема, обозначенная ранее в этом разделе (а также в разделе 1.10), связанная со стационарными решениями уравнений Шрёдингера: такое решение обязательно должно распространяться на всю Вселенную. Эту проблему можно решить либо путем своеобразной импровизации — считать, что центр масс жестко расположен в фиксированном месте, либо (что, пожалуй, предпочтительнее) воспользоваться уравнением Шрёдингера — Ньютона (ШН). Речь идет о нелинейном обобщении стандартного шрёдингеровского уравнения, в котором учитывается гравитационное воздействие ожидаемого распределения масс. Для этого сама волновая функция включается в уравнение как ньютоновское гравитационное поле, которое приплюсовывается к гамильтониану [Ruffini and Bonazzola, 1969; Diòsi, 1984; Moroz et al., 1998; Tod and Moroz, 1999; Robertshaw

and Tod, 2006]. До сих пор основная ценность уравнения ШН в связи с **OR** заключается в том, что оно позволяет рассуждать об альтернативных стационарных состояниях, в которые может редуцироваться система в результате **OR**.

Консолька, на которой удерживается крошечное зеркало, позволяет ему отклоняться от исходного положения в течение некоторого заранее отмеренного временного интервала — скажем, нескольких секунд или минут. Чтобы проверить, на самом ли деле состояние зеркальца спонтанно редуцировалось под ударами фотонов или же квантовая когерентность сохранилась, фотон будет освобождаться из этой отражательной ловушки (состоящей из зеркала-отражателя и полусферического зеркала) так, чтобы он мог вернуться по прежней траектории в светоделитель. Тем временем вторая часть волновой функции фотона будет использоваться для фиксации времени, попадая в другую отражательную ловушку, состоящую из двух стационарных зеркал. Если, как утверждает стандартная квантовая теория, между двумя раздельными частями волновой функции фотона действительно должна сохраняться фазовая когерентность, то это можно будет подтвердить при помощи детектора фотонов, установленного в нужной точке по другую сторону от светоделителя (рис. 4.11). Если когерентность в системе сохраняется, то возвращающийся фотон должен всегда (или, при измененной установке, никогда) попадать именно в этот детектор, и тогда детектор будет срабатывать.

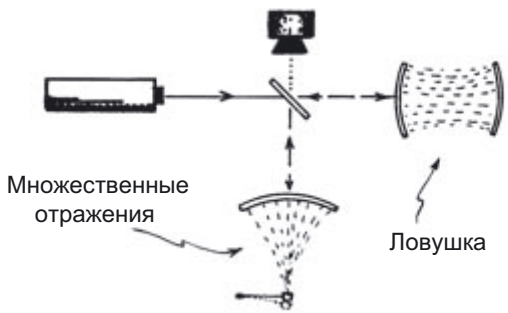


Рис. 4.11. Здесь проиллюстрирован эксперимент Баувместера, призванный проверить, происходит ли в природе гравитационная **OR**. Отдельный фотон испускается из лазерного источника и попадает в светоделитель таким образом, что траектория фотона делится на горизонтальную и вертикальную части. Горизонтальный путь ведет в ловушку, где фотон может подолгу удерживаться, отражаясь туда-сюда от зеркал, образующих эту ловушку, а вертикальный путь ведет в другую ловушку, где находится всего одно крошечное зеркало, подвешенное на консольке, так что под давлением множества отражающихся фотонов оно должно немного сдвигаться. Согласно **OR**-гипотезе, спустя некоторое время два местоположения подвешенного зеркала должны спонтанно редуцироваться в одно, и суперпозиция исчезнет. Это можно будет зафиксировать при помощи детектора, изображенного вверху

В настоящее время этот эксперимент все еще не позволяет достоверно протестировать гипотезу. В известной степени успешность этого эксперимента должна подтверждать прогнозы стандартной квантовой механики, но сдвиги при эксперименте должны быть намного больше, чем достигаются сегодня (речь идет о сдвигах масс между состояниями, находящимися в суперпозиции). Однако ожидается, что, когда в будущем этот эксперимент удастся усовершенствовать, начнется серьезная экспериментальная проверка границ квантовой теории. Вполне возможно, что через несколько лет удастся экспериментально определить, подтверждаются ли в наблюдениях те гипотезы, которые я выдвигаю [Weaver et al., 2016; Eerkens et al., 2015; Pepper et al., 2012] (см. также [Kaltenbaek et al., 2015; Li et al., 2011]).

Завершая этот раздел, следует остановиться на некоторых проблемах, связанных с редукцией квантового состояния, которые могут оказаться важными, если данная модель экспериментально подтвердится. Из вышеизложенного должно быть понятно, что эта гипотеза подлинно объективна в следующем смысле: такая **OR**-схема предполагает, что **R** происходит где-то в объективной реальности и не зависит от того, что квантовую систему наблюдает некое сознающее существо. В тех частях Вселенной, где и в помине нет никаких сознающих наблюдателей, **R**-события будут происходить точно в таких же обстоятельствах, с такой же частотой и с такими же вероятностями, как и здесь, где за этими процессами наблюдают многочисленные разумные существа. С другой стороны, я не раз [Penrose, 1989, 1994, 1997] продвигал идею, согласно которой феномен *сознания как такового* может зависеть от **OR**-событий (происходящих преимущественно в нейронных микротрубочках), и каждое такое событие, в сущности, дает момент «протосознания», первоэлемент, из которого неким образом формируется сознание как таковое [Hameroff and Penrose, 2014].

В рамках этих исследований я делаю небольшую генерализацию рассмотренной выше **OR**-гипотезы. Эта генерализация может касаться ситуаций, когда энергии E_1 и E_2 двух состояний, находящихся в суперпозиции, слегка отличаются друг от друга. В стандартной квантовой механике такая суперпозиция должна осциллировать между двумя этими состояниями с частотой $|E_1 - E_2|/h$, сопровождаясь гораздо более активными квантовыми осцилляциями с частотой около $(E_1 + E_2)/2h$. Обобщенная **OR**-гипотеза заключается в том, что в такой ситуации спустя период времени, в среднем равный около $\tau \approx \hbar/E_G$, состояние должно спонтанно редуцироваться в *классические осцилляции* между двумя этими альтернативами с частотой $|E_1 - E_2|/h$, причем фактическая фаза колебаний будет результатом «случайного» выбора, зависящим от **OR**. Однако эту гипотезу нельзя считать совершенно универсальной, поскольку должен существовать классический энергетический барьер, предотвращающий классические осцилляции.

Естественно, все вышеизложенное далеко не дотягивает до согласованной математической теории обобщенной квантовой механики, предельным случаем которой были бы **U** и **R** (а также классическая общая теория относительности). Какие можно сделать предположения об истинной сути такой теории? Думаю, почти никаких, поскольку такая теория произвела бы настоящую революцию в аппарате квантовой механики, а не свелась бы к простым уточнениям существующего формализма. Конкретнее, я склонен полагать, что в этом должны сыграть какую-то роль элементы теории твисторов, поскольку она позволяет надеяться, что загадочные нелокальные аспекты квантовой запутанности и квантовых измерений удастся увязать с нелокальностями голоморфной когомологии, с которой мы неизбежно сталкиваемся, углубляясь в технические тонкости теории твисторов (см. раздел 4.1). Я с оптимизмом отношусь к некоторым новейшим наработкам дворцовой твисторной теории, вкратце упомянутой в конце раздела 4.1, — возможно, они позволят нам продвинуться в этих исследованиях [Penrose, 2015a, b].

4.3. Конформная невообразимая космология?

Кроме обоснований, изложенных в разделах 2.13 и 4.2, существуют и другие причины полагать, что квантовая теория в ее стандартном виде неприменима к гравитационному полю как таковому в системах, где гравитация начинает играть заметную роль на квантово-механическом уровне. Одна из подобных причин связана с так называемым информационным парадоксом, который сопряжен с хокинговским испарением черных дыр. Есть все основания полагать, что эта проблема связана с возможной сущностью квантовой гравитации, и чуть ниже я ее рассмотрю. Но есть и другая проблема, которая маячит за всеми выкладками из главы 3. Речь об очень странной природе Большого взрыва, которую мы особенно подробно обсуждали в разделах 3.4 и 3.6, а именно: Большой взрыв обладал очень ограниченными гравитационными степенями свободы, причем очевидно, это касалось *только* гравитационных степеней свободы.

Традиционный подход к описанию физики Большого взрыва таков: это был единственный наблюдаемый (пусть и опосредованно) феномен, в котором отчетливо проявлялись эффекты *квантовой гравитации* (какова бы ни была эта теория). Поэтому нередко утверждают, что нужно лучше разобраться в Большом взрыве, чтобы всерьез подступиться к удручающе сложной квантовой гравитации. На самом деле я иногда и сам прибегал к этому аргументу, высказываясь в поддержку исследований квантовой гравитации (см. введение к книге «Квантовая гравитация» [Isham et al., 1975]).

Но можно ли всерьез ожидать, что какая-либо традиционная квантовая теория (поля), примененная к гравитационному полю, позволит объяснить исключительно странную структуру, которой очевидно обладал Большой взрыв, независимо от того, последовал ли инфляционный этап сразу за этим мгновенным событием? Полагаю, что по причинам, которые я постарался изложить в главе 3, — нет, *не позволит*. Требуется объяснить экстраординарное подавление гравитационных степеней свободы при Большом взрыве. Если все эти $10^{10^{124}}$ возможностей потенциально присутствовали при Большом взрыве, как того требует формализм квантовой механики, то следует ожидать, что каждая из этих возможностей внесла вклад в общее исходное состояние. Это противоречит обычным процедурам КТП, поэтому приходится просто *объявить*, что эти возможности должны были отсутствовать. Более того, сложно понять, как модели из раздела 3.11, включающие события *до* Большого взрыва, помогли бы избежать этих сложностей, поскольку логично ожидать, что эти гравитационные степени свободы должны были бы самым выразительным образом отпечататься в геометрии пространства *после* отскока, поскольку они явно должны были существовать *до* отскока.

Существует и схожее свойство, которым должна была бы обладать любая теория квантовой гравитации обычного типа, а именно динамическая *симметрия относительно времени*. В данном случае нас интересует нечто вроде уравнения Шрёдингера (**U**-процесс), симметричного относительно времени при замене $i \to -i$, поскольку оно будет применяться к симметричным относительно времени уравнениям общей теории относительности Эйнштейна. Если предполагается, что такая квантовая теория будет применяться к исключительно высокоэнтропийным сингулярностям, ожидаемым в черных дырах, весьма возможно, общего БКЛМ-типа, то такие же пространственно-временные сингулярности (обращенные во времени вспять) также должны применяться и к Большому взрыву, и это допускается все той же «обычной» квантовой теорией. Тем не менее при Большом взрыве ничего такого не происходило. Более того — надеюсь, мне удалось это доказать в разделе 3.10, — антропный аргумент практически бесполезен при объяснении таких колоссальных ограничений, присущих Большому взрыву.

Да, Большой взрыв обладал крайне специфическими ограничениями, которые абсолютно не обнаруживаются в сингулярностях черных дыр. Наблюдения убедительно свидетельствуют о том, что именно там, где эффекты квантовой гравитации «должны были» наиболее отчетливо отразиться на физических явлениях близ таких сингулярностей, наблюдается абсолютнейшая *асимметрия относительно времени*. Так не должно быть, если привлекать для объяснения квантовую теорию обычного типа, даже с учетом антропного объяснения, которое так выручает. Как я говорил ранее, должно быть другое объяснение.

Сам я в данном случае предпочитаю на время отвлечься от квантовой теории и поразмыслить над *геометрией*, которая должна была существовать сразу после Большого взрыва, а затем сравнить эту геометрию с совершенно диким типом геометрии (вполне возможно, БКЛМ-типа; см. в конце раздела 3.2), которая должна обнаруживаться близ сингулярностей черных дыр. Первая проблема — просто охарактеризовать условия, при которых гравитационные степени свободы во время Большого взрыва могли быть *подавлены*.

Долгие годы (фактически с 1976 года) я излагал эту идею в категориях так называемой *гипотезы кривизны Вейля* [Penrose, 1976a, 1987a, 1989, глава 7; ПкР раздел 28.8]. *Конформный тензор Вейля* **C** измеряет тип пространственно-временной кривизны, важной для *конформной* геометрии пространства-времени. Как было показано в разделах 3.1, 3.5, 3.7 и 3.9, это геометрия, определяемая при помощи системы световых конусов (или нулевых конусов) в пространстве-времени. Чтобы записать формулу для определения **C**, мне пришлось бы здесь сильно углубиться в тензорное исчисление, а эта тема технически серьезно выходит за рамки данной книги. Поэтому очень хорошо, что в этом рассказе я могу обойтись без такой формулы. Однако некоторые свойства **C**, связанные с конформным изменением масштаба ($\hat{\mathbf{g}} = \Omega^2\mathbf{g}$), вскоре нас всерьез заинтересуют.

Следует отметить особую геометрическую роль тензора **C**. Дело в том, что уравнение $\mathbf{C} = \mathbf{0}$, которое справедливо в некоторой не слишком обширной односвязной открытой области пространства-времени $\mathcal{R}$, утверждает, что $\mathcal{R}$ (с метрикой $\mathbf{g}$) является *конформно плоским*. Это означает, что существует такое вещественное скалярное поле Ω (именуемое *конформным множителем*), что конформно связанная с ним пространственно-временная метрика $\hat{\mathbf{g}} = \Omega^2\mathbf{g}$ является плоской метрикой Минковского в $\mathcal{R}$ (см. разделы A.6 и A.7, где доступно объясняется, что такое «односвязный» и «открытый», однако здесь эти термины не играют для нас особой роли).

Полный риманов тензор кривизны **R** содержит 20 независимых компонентов на точку, и его фактически можно разбить на эйнштейновский тензор **G** (см. разделы 1.1 и 3.1) и тензор Вейля **C**, в каждом из которых по 10 компонентов на точку. Вспомните эйнштейновские уравнения : $\mathbf{G} = 8\pi\gamma\mathbf{T} + \Lambda\mathbf{g}$. Здесь **T** — материальный тензор энергии-импульса, который описывает, как все степени свободы материи (в том числе степени свободы электромагнитного поля) напрямую влияют на кривизну пространства-времени через **G**, являющийся частью полной пространственно-временной кривизны **R**; также имеем вклад $\Lambda\mathbf{g}$, обусловленный космологической постоянной. Оставшиеся 10 компонентов кривизны в **R** описывают *гравитационное* поле, и их удобно представлять с помощью вейлевского тензора **C**.

Согласно гипотезе кривизны Вейля, любая пространственно-временная сингулярность *прошлого* типа — это такая сингулярность, из которой могут выходить времениподобные линии, уходящие в будущее, но в которую не могут приходить такие линии из прошлого (см. рис. 4.12 *а* и *б*). Сингулярность прошлого типа должна иметь *тензор Вейля, обращающийся в нуль* в пределе приближения к этой сингулярности из будущего вдоль любой их таких времениподобных линий. Согласно этой гипотезе, Большой взрыв (и любая другая сингулярность «взрывного» типа, которая могла бы существовать, например, в момент «хлопка» при окончательном исчезновении черной дыры под действием хокинговского излучения; см. рис. 4.12 *в* и обсуждение далее в этом разделе[1]) не должен содержать независимых гравитационных степеней свободы. Эта гипотеза ничего не сообщает о сингулярностях будущего типа или о голых сингулярностях (см. рис. 4.12 *г*), где будут присутствовать как входящие, так и исходящие времениподобные линии (считается, что классических сингулярностей такого типа быть не может согласно принципу сильной космической цензуры; см. разделы 3.4 и 3.10).

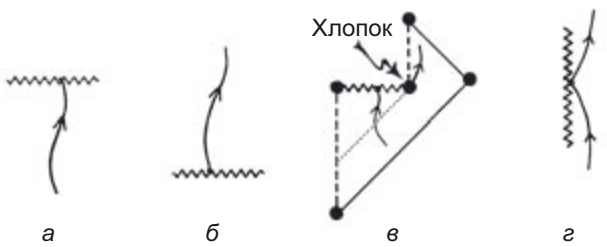


Рис. 4.12. Различные типы пространственно-временной сингулярности: *а*) будущий тип — в такую сингулярность попадают лишь те мировые линии, которые приходят в нее из прошлого; *б*) прошлый тип — в такой сингулярности присутствуют лишь те мировые линии, которые возникают в ней и устремляются в будущее; *в*) черная дыра, испаряющаяся под действием хокинговского излучения; внутри нее должна быть сингулярность будущего типа, но последний «хлопок», по всей видимости, относится к прошлому типу; *г*) гипотетическая голая сингулярность, куда попадают и мировые линии прошлого, и мировые линии будущего. В соответствии с гипотезой космической цензуры, сингулярности типа, приведенного на рис. 4.12 *г*, не должны возникать в типичных классических ситуациях. Согласно гипотезе кривизны Вейля, сингулярности типа, приведенного на рис. 4.12 *б*, например Большой взрыв, должны иметь очень сильные ограничения, поскольку в них подавляется кривизна Вейля

[1] В отличие от времениподобной сингулярности, представленной на рис. 4.12 *г*, «хлопок» с рис. 4.12 *в*, на мой взгляд, нельзя считать нарушением гипотезы космической цензуры, поскольку он больше похож на две отдельные сингулярности, а именно: в будущем расположена БКЛМ-часть, представленная волнистой линией неправильной формы, и на ней четко заметен «хлопок». Причинные структуры обеих действительно сильно отличаются, и я не вижу смысла их идентифицировать.

Здесь следует прояснить одну деталь относительно гипотезы кривизны Вейля. Она позиционируется просто как геометрическое утверждение, которое достаточно четко постулирует, что при Большом взрыве и других сингулярностях прошлого типа (если такие существуют) гравитационные степени свободы должны быть почти полностью подавлены. Эта гипотеза не содержит никаких предложений по поводу определения *энтропии* в гравитационном поле (как о некоторой скалярной величине, алгебраически конструируемой из **C**, — в таком виде ее предлагали некоторые специалисты, что на самом деле было не слишком удачно). Колоссальный эффект данной гипотезы в плане объяснения *низкоэнтропийности* Большого взрыва (см. раздел 3.6) напрямую следует из нее, поскольку она подразумевает, что не должно было существовать никаких первозданных белых дыр (или черных дыр). Фактическая «низкая энтропия» и вычисленное значение невероятности ($10^{-10^{124}}$) выводятся непосредственно из формулы Бекенштейна — Хокинга (см. раздел 3.6).

Однако есть некоторые технические сложности, связанные с точной математической интерпретацией гипотезы кривизны Вейля. Одна сложность связана с тем, что **C** как таковая, будучи тензорной величиной, не определяется *на самой* пространственно-временной сингулярности, где, строго говоря, такие тензорные понятия нормально не определяются. Соответственно, утверждение $\mathbf{C} = \mathbf{0}$ для такой сингулярности нужно формулировать в смысле предела по мере приближения к сингулярности. Сложности связаны с тем, что есть несколько неэквивалентных способов сформулировать это условие и непонятно, какой из них наиболее подходящий. С учетом таких неопределенностей очень хорошо, что мой оксфордский коллега Пол Тод предложил и тщательно изучил альтернативную формулировку математических условий Большого взрыва, которая в явном виде вообще не привязана к категориям **C**.

В соответствии с гипотезой Тода [2003] (как и в случае с ФЛРУ-сингулярностями Большого взрыва, см. в конце раздела 3.5), наш Большой взрыв можно *конформно* представить как гладкую пространственноподобную 3-поверхность $\mathcal{B}$, в пределах которой пространство-время, в принципе, можно конформно расширить в прошлое. Это означает, что, имея подходящий конформный множитель Ω, можно перекалибровать нашу *физическую* метрику $\breve{\mathbf{g}}$, актуальную после Большого взрыва, в новую метрику $\mathbf{g}$:

$$\mathbf{g} = \Omega^2 \breve{\mathbf{g}},$$

в соответствии с которой пространство-время получит гладкую границу прошлого $\mathcal{B}$ (где $\Omega = \infty$), в пределах которой эта *новая* метрика $\mathbf{g}$ остается строго определенной и гладкой. Таким образом, появляется возможность продолжить $\mathbf{g}$

в гипотетический «добольшевзрывный» регион пространства-времени; рис. 4.13 (надеюсь, что слегка странная нотация, где $\check{\mathbf{g}}$ соответствует фактической *физической* метрике, не запутает читателя; просто она позволяет ссылаться на величины, определяемые в $\mathcal{B}$, без использования всяких «штрихов» и «крышечек», что пригодится нам впоследствии). Однако следует отметить, что гипотеза Тода не утверждает, что $\mathbf{C} = \mathbf{0}$ на $\mathcal{B}$, а требует, чтобы $\mathbf{C}$ оставался на $\mathcal{B}$ *конечным* (поскольку конформное пространство-время здесь гладкое), что, тем не менее, очень сильно ограничивает гравитационные степени свободы на $\mathcal{B}$ и определенно исключает любые БКЛМ-подобные явления.

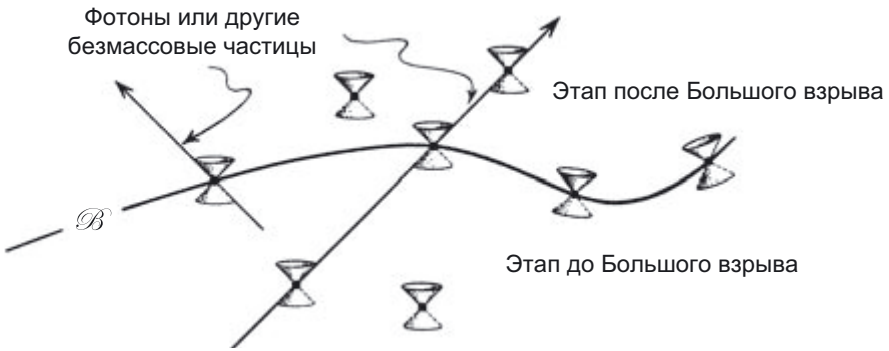


Рис. 4.13. Гипотеза Тода об ограничениях, характерных для Большого взрыва (по типу гипотезы кривизны Вейля) предполагает, что пространство-время можно расширить в прошлое конформно гладким образом, так, что исходная сингулярность становится гладкой гиперповерхностью $\mathcal{B}$, в пределах которой конформное пространство-время гладко продолжается в гипотетическую область, предшествующую Большому взрыву. Если счесть такую предшествующую область физически реальной, то безмассовые частицы, например фотоны, должны беспрепятственно проникать через $\mathcal{B}$ из прошлого в будущее

В исходной версии гипотезы Тода эта дополнительная область, существовавшая до Большого взрыва, не считалась физически «реальной»; она вводилась просто как аккуратный математический артефакт, позволяющий четко сформулировать гипотезу кривизны Вейля, без необходимости вводить неуклюжие и словно взятые с потолка математические граничные условия. Примерно в таком духе обычно исследовались *будущие* асимптотики общерелятивистских пространств-времен (эту идею я предложил в 1960-е годы), предположительно позволявшие анализировать поведение исходящего гравитационного излучения [Penrose, 1964b, 1965b, 1978; Penrose and Rindler, 1986, глава 9]. В этих исследованиях асимптотическое *будущее* могло рассматриваться с геометрической точки зрения; для этого будущее пространственно-временного многообразия снабжалось гладкой конформной границей (см. раздел 3.5). В данном случае давайте называть *физической* физическую метрику отдаленного будущего $\hat{\mathbf{g}}$ (опять же прошу

прощения за путаную нотацию, в том числе за изменение нотации по сравнению с разделом 3.5 и за то, что физическую метрику теперь обозначаю не $\check{\mathbf{g}}$, а $\hat{\mathbf{g}}$ — это изменение я поясню чуть ниже), и мы перекалибруем ее в новую, конформно связанную с ней метрику $\mathbf{g}$, по формуле:

$$\mathbf{g} = \omega^2 \hat{\mathbf{g}}\,.$$

Теперь метрика $\mathbf{g}$ гладко продолжается за пределы гладкой 3-поверхности $\mathcal{F}$, где $\omega = 0$. Вновь вернемся к рис. 4.13, но теперь рассмотрим расширение физического пространства-времени в нижней части рисунка через будущую бесконечность $\mathcal{F}$ (а не через $\mathcal{B}$) в гипотетическую область «бесконечно отдаленного будущего».

Я уже многократно прибегал к таким фокусам в этой книге, например в случаях с конформными диаграммами в разделе 3.5, когда речь шла о представлении космологических ФЛРУ-моделей, в которых, согласно условиям строгих конформных диаграмм (см., например, рис. 3.22 в разделе 3.5), Большой взрыв $\mathcal{B}$ в каждой модели изображается в виде зазубренной линии на границе с прошлым, а бесконечно отдаленное будущее $\mathcal{F}$ — в виде плавной линии на границе с будущим. По условиям этих диаграмм, при повороте относительно оси симметрии мы в каждом случае получаем трехмерные конформные границы пространств-времен. Особенность выкладок из этого раздела состоит в том, что здесь мы рассматриваем более универсальные пространственно-временные модели, так что в них *не* требуется наличия никакой вращательной симметрии, а в ФЛРУ-моделях не предполагается наличия никаких высших симметрий.

Откуда мы знаем, что такие уловки применимы и в более общих случаях? Здесь обнаруживается огромная логическая разница между случаями гладкой $\mathcal{B}$ и гладкой $\mathcal{F}$. Оказывается, что при очень широких физических допущениях (например, с предположением о положительности космологической постоянной Λ, что, по-видимому, подтверждается наблюдениями) с математической точки зрения можно ожидать существования гладкого конформного бесконечно отдаленного будущего $\mathcal{F}$ (что подразумевается в теоремах Гельмута Фридриха [1986]). С другой стороны, существование гладкой *исходной* 3-поверхности Большого взрыва $\mathcal{B}$ вносит чрезвычайно жесткие граничные условия в космологическую модель, чего и следует ожидать в свете гипотезы Тода, выдвинутой именно с учетом таких ограничений и с надеждой математически закодировать невероятность, даже если ее масштаб составляет порядка $10^{-10^{124}}$.

С математической точки зрения существование гладких конформных границ ($\mathcal{B}$ в прошлом и $\mathcal{F}$ в будущем) удобно представлять в контексте теоретической возможности *продолжения* пространства-времени на другую сторону границы такой 3-поверхности, но подобное продолжение следует считать не более чем

математической уловкой, вводимой просто для облегчения формулировки условий, которые было бы неудобно выражать каким-то иным образом, и такая уловка позволяет оперировать локальными геометрическими категориями, а не громоздкими асимптотическими пределами. Теоретики придерживаются подобной перспективы, пользуясь такими идеями о конформных границах как для $\mathcal{B}$, так и для $\mathcal{F}$. Однако приходится признать, что и сама физика довольно хорошо стыкуется с этими математическими процедурами, и существует довольно убедительная, пусть и дерзкая (фантастическая?) возможность, что *физика* мира может быть осмысленно продолжена за эти трехмерные конформные границы как для $\mathcal{B}$, так и для $\mathcal{F}$. Поэтому мы размышляем, мог ли в самом деле существовать мир до Большого взрыва, а также может ли скрываться другой мир за нашим бесконечно отдаленным будущим!

Ключевой момент таков, что большая часть физики — в сущности, вся физика, не имеющая дела с массивными частицами, — по-видимому, не изменяется (является инвариантной) при рассматриваемой здесь конформной перекалибровке. Это исключительно хорошо подтверждается для максвелловских уравнений электромагнетизма, причем не только для уравнений свободного электромагнитного поля, но и для законов взаимодействия токов и зарядов, являющихся источниками электромагнитного поля. То же касается (классических) уравнений Янга — Миллса, описывающих сильные и слабые ядерные взаимодействия; они являются расширениями максвелловских уравнений, где группа калибровочных симметрий поворотов фазы расширяется до более обширных групп, необходимых для описания слабых или сильных взаимодействий (см. разделы 1.8 и 1.15).

Однако когда речь заходит о квантовых версиях этих теорий, требуется отдельная оговорка (особенно это касается уравнений Янга — Миллса), поскольку могут возникать конформные аномалии, при которых квантовая теория не обладает полноценной симметрией, характерной для классической теории [Polyakov, 1981a, b; Deser, 1996]. Можно вспомнить, что эта проблема была особенно важна при разработке теории струн (см. раздел 1.6, а также раздел 1.11). Хотя я и считаю, что эта проблема конформных аномалий, скорее всего, окажется немаловажной в контексте описываемых здесь идей, берусь утверждать, что эти аномалии нисколько не отменяют правильности общей модели.

Данная конформная инвариантность присуща исключительно уравнениям поля для безмассовых частиц, которые являются носителями этих взаимодействий. Переносчики электромагнетизма — это фотоны, а переносчики сильного взаимодействия — глюоны. Правда, в случае со слабым взаимодействием есть одна сложность, поскольку его переносчиками считаются очень массивные W- и Z-бозоны. Можно предположить, что если отмотать время назад, то по мере

приближения к Большому взрыву температура будет постоянно повышаться, пока массы покоя всех рассматриваемых частиц (как и проблема конформных аномалий) не станут абсолютно пренебрежимыми по сравнению с невообразимо высокой кинетической энергией движущихся частиц. Соответствующая физика Большого взрыва — по сути физика безмассовых частиц — будет конформно инвариантной, и, следовательно, если проследить ее вплоть до ограничивающей 3-поверхности $\mathcal{B}$, то материя, по сути, вообще не заметит $\mathcal{B}$. Эта материя должна обладать «прошлым» даже на $\mathcal{B}$, точно так же как повсюду в физике, и это «прошлое» будет описываться тем, что происходит в пределах предлагаемого *теоретического* расширения пространства-времени, как этого и требует гипотеза Тода.

Но какие космологические процессы теоретически могли происходить в той гипотетической продолженной области Вселенной, о которой говорит Тод и где мы теперь пытаемся всерьез представить себе *до*большевзрывную физическую реальность? Наиболее вероятно, что тогда Вселенная претерпевала коллапс, как в условиях расширенной фридмановской модели ($K > 0$), рассмотренной в разделе 3.1 (см. рис. 3.6 или 3.8 в разделе 3.1), либо в условиях многих альтернативных моделей «отскока», например в экпиротической модели, описанной в разделе 3.11. Правда, всем им свойственна проблема, связанная со вторым законом термодинамики и неоднократно упоминаемая в главе 3, а именно: либо до отскока второй закон термодинамики действовал, как и сейчас, и тогда становится очень сложно увязать крайне хаотичное Большое схлопывание (показанное на рис. 3.48) с размеренным Большим взрывом, либо второй закон действовал *обратным* образом (то есть диаметрально противоположно по обе стороны от отскока), но тогда невозможно объяснить существование столь маловероятного момента (момента отскока), невероятность которого выражается числом $10^{-10^{124}}$ (см. раздел 3.6).

Я выдвигаю совершенно иную идею, а именно: предлагаю исследовать противоположный конец шкалы времени/расстояния и обратиться к другой конформно-математической уловке, которую мы недавно рассмотрели. Речь идет о конформном «расплющивании» отдаленного будущего, как показано на многих примерах в разделе 3.5 (см., например, рис. 3.25 и рис. 3.26 *а*). Таким образом мы сможем расширить модель за пределы гладкой 3-поверхности $\mathcal{I}$ в направлении будущей бесконечности. Наличие положительной космологической постоянной Λ подразумевает, что, во-первых, $\mathcal{I}$ — это *пространственноподобная* 3-поверхность [Penrose, 1964b; Penrose and Rindler, 1986, глава 9], а во-вторых, как отмечалось выше, что продолжение через гладкую поверхность $\mathcal{I}$ представляет собой общее явление, что прямо продемонстрировал Фридрих [1998], при определенных широких допущениях. Как подчеркивалось ранее, весьма *мало*вероятно, что второй аспект действительно был характерен для типичного большого взрыва, ведь гипотеза Тода, требующая гладкого продолжения пространства-времени

за пределы $\mathcal{B}$, налагает огромные (весьма желательные) *ограничения* на физику Большого взрыва.

Такое ровное конформное продолжение за пределы $\mathcal{F}$ могло бы происходить в отдаленном будущем как минимум при условии, что в тот период материальное содержимое Вселенной уже состояло бы преимущественно из *безмассовых* частиц — ведь это подразумевается в вышеприведенном утверждении. Насколько вероятно, что в столь отдаленном будущем во Вселенной останутся лишь безмассовые частицы? В данном случае нужно рассмотреть две основные проблемы: одна связана с природой частиц, которые сохранятся в отдаленном будущем, а другая — с черными дырами.

Начнем с черных дыр. Поначалу они должны неудержимо расти, поглощая все больше и больше материи, а потом они станут поглощать космическое реликтовое излучение, когда больше никакой «еды» не останется! Однако после того как температура РИ наконец окажется ниже температуры хокинговского излучения черной дыры, дыра начнет очень медленно испаряться, пока не сгинет в финальном взрыве (относительно слабом по астрофизическим меркам). В целом этот процесс будет очень длителен у сверхмассивных черных дыр в ядрах галактик и более скоротечным у более мелких черных дыр с массами в несколько солнечных. Согласно этой модели примерно через 10^{100} лет (в зависимости от того, насколько массивными будут эти черные дыры) все закончится. Подобный сценарий впервые предложил Стивен Хокинг в 1974 году, и этот прогноз кажется мне наиболее правдоподобным.

Что же можно сказать о частицах, которые сохранятся в столь отдаленном будущем? В численном отношении абсолютное большинство из них составят фотоны. Уже подтверждено, что на каждый барион приходится 10^9 фотонов и большинство из этих фотонов относится к реликтовому излучению. Это число должно оставаться практически постоянным, несмотря на то что звезды постепенно погаснут, а многие барионы упокоятся в черных дырах. Дополнительный вклад внесет хокинговское излучение сверхмассивных черных дыр — это будут крайне низкочастотные фотоны.

Однако остается какое-то количество массивных частиц, которые следует упомянуть. Некоторые из них, сегодня считающиеся стабильными, вполне могут распасться — часто утверждают, что когда-нибудь распадутся все протоны. Однако протоны обладают (положительным) электрическим зарядом, и до тех пор, пока закон сохранения электрического заряда остается законом природы, какие-то реликтовые заряженные частицы должны уцелеть. Наименее массивной подобной частицей может быть позитрон — античастица, парная электрону. Из рассуждений о горизонтах и т. д. (см. рис. 4.14 и [Penrose, 2010, раздел 3.2, рис. 3.4]) понятно,

что на неопределенно долгий срок должны сохраниться и электроны, и позитроны (а возможно, и другие массивные частицы). Им некуда деться, поскольку не существует безмассовых заряженных частиц (об этом известно по изучению процесса аннигиляции пар [Bjorken and Drell, 1964]). Можно допустить, что закон сохранения заряда не абсолютен, но эта крайне маловероятная возможность нам ничем не поможет, поскольку, согласно теоретическим выкладкам, тогда массу обретут сами фотоны [Bjorken and Drell, 1964]. Что касается оставшихся частиц, не имеющих заряда, должны сохраниться самые легкие нейтрино, хотя я понимаю, что современные эксперименты по-прежнему не исключают существования особых, безмассовых нейтрино (см. [Fogli et al., 2012]).

Рис. 4.14. На этой схематичной конформной диаграмме показано, как в пространственноподобной бесконечности $\mathcal{F}$, которая должна существовать при положительной Λ, отдельные заряженные частицы, например электроны и позитроны, должны настолько сильно отдалиться друг от друга, что уже не смогут аннигилировать

Вышеизложенные соображения, по-видимому, означают, что, пусть условия для существования гладкой (пространственноподобной) конформной границы будущего $\mathcal{F}$ и удовлетворяются почти полностью, в предельно далеком будущем все-таки должны попадаться отдельные массивные частицы, нарушающие эту стройную картину. Здесь я собираюсь описать модель под названием *конформная циклическая космология* (КЦК), и она получается максимально непротиворечивой, если к наступлению $\mathcal{F}$ сохранятся лишь *безмассовые* частицы. Соответственно, я предполагаю, что в крайне отдаленном будущем исчезнет масса покоя как таковая — она обнулится к наступлению асимптотического бесконечного временного предела. Темпы ее исчезновения должны быть немыслимо медленными, чтобы не противоречить имеющимся наблюдениям. Можно сравнить такое явление с обратным хиггсовским механизмом, который фактически включается лишь тогда, когда температура окружающей среды достигает экстремально низких значений. На самом деле переменные значения масс покоя частиц — характерная особенность некоторых физических теорий [Chan and

Tsou, 2007, 2012; Bordes et al., 2015], поэтому вполне допустимо предположить, что в конце концов все массы частиц сведутся к нулю, хотя, конечно, «в конце концов» может означать «очень нескоро». Такие теории позволяют предположить, что разные частицы будут утрачивать массу покоя разными темпами, поэтому данное явление нельзя увязать с общим уменьшением гравитационной постоянной. Общая теория относительности требует, чтобы вдоль любой времениподобной мировой линии было определено непрерывное собственное время. Пока массы покоя считаются постоянными, такую меру времени лучше всего выражать в категориях, описанных в разделе 1.8, где комбинация эйнштейновского уравнения $E = mc^2$ с планковским $E = h\nu$ позволяет заключить, что любая стабильная частица массой m принципиально работает как идеальные часы с частотой mc^2/h. Но такой принцип перестанет работать в далеком будущем, если частицы станут терять массу разными темпами.

КЦК требует положительной космологической постоянной Λ (для того чтобы $\mathcal{F}$ была пространственноподобной). Следовательно, Λ в некотором смысле отслеживает масштаб, так что эйнштейновские уравнения (включающие Λ) остаются применимыми в конечных областях пространства-времени; правда, сложно понять, как при помощи Λ можно было бы создать локальные часы. Важная базовая составляющая КЦК заключается в том, что часы утрачивают свой смысл по мере приближения к $\mathcal{F}$, так что идеи конформной геометрии приобретают решающее значение, и как на $\mathcal{B}$, так и на $\mathcal{F}$ следует руководствоваться иными физическими принципами.

Перейдем к самой гипотезе КЦК и рассмотрим, почему она может быть желательна. Эта идея заключается в том [Penrose, 2006, 2008, 2009a, b, 2010, 2014b; Gurzadyan and Penrose, 2013], что наши современные представления о вечно расширяющейся Вселенной, возникшей из Большого взрыва (но *без* инфляционного этапа) и устремленной в экспоненциально расширяющееся бесконечное будущее, — это всего лишь один *эон* в бесконечной череде таких эонов, $\mathcal{F}$ каждого из которых гладко конформно совпадает с $\mathcal{B}$ следующего (см. рис. 4.15), и получающееся в результате конформное четырехмерное многообразие является гладким на всех таких стыках. В некотором отношении эта схема напоминает циклическую/экпиротическую гипотезу Стейнхардта — Турока (см. раздел 3.11), но без каких-либо сталкивающихся бран или других деталей из теории струн/М-теории. Кроме того, она в чем-то похожа на гипотезу Венециано (см. раздел 3.11), поскольку не предполагает инфляционного этапа после каждого большого взрыва[1], но в некотором смысле именно экспоненциальное

[1] Как и в разделах 3.4 и 3.5, я пишу «Большой взрыв» с прописной буквы, имея в виду конкретное событие, с которого начался наш эон, и со строчной буквы, имея в виду другие эоны, то есть употребляя термин в общем смысле.

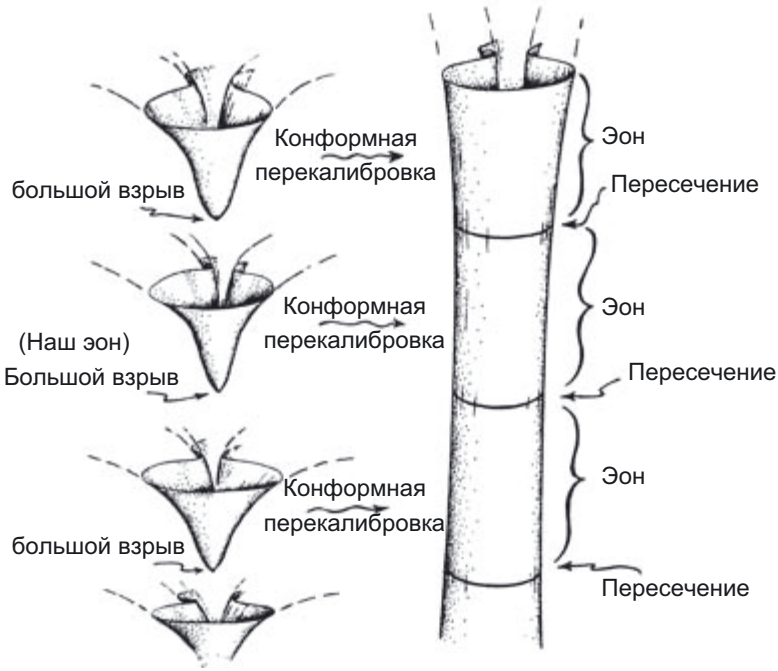


Рис. 4.15. Модель конформной циклической космологии (КЦК). Согласно данной гипотезе, традиционная картина (показанная на рис. 3.3) всей истории Вселенной (но без инфляционного этапа), только каждый «эон» — всего лишь один в бесконечной последовательности подобных эонов. Переход от одного эона к другому — это конформно гладкое продолжение от будущей бесконечности предыдущего эона к большому взрыву следующего (а потенциальный инфляционный этап каждого эона в данной модели меняется на окончательное экспоненциальное расширение в конце предыдущего эона)

расширение на последнем этапе эона, предшествовавшего нашему, позволяет выдвинуть серьезные доводы в пользу инфляции. Эти вопросы упоминались в разделе 3.9: 1) приблизительная масштабная инвариантность температурных флуктуаций РИ; 2) наличие корреляций в РИ на масштабах, превышающих радиус горизонта; 3) требование, чтобы в ранней Вселенной локальная плотность материи ρ была исключительно близка к критическому значению ρ_c — кажутся весьма вероятными следствиями КЦК.

Наконец, нужно остановиться на некоторых моментах, связанных с жизнеспособностью КЦК. Один из вопросов: как подобная модель может согласовываться со вторым законом термодинамики? Ведь он, по-видимому, в принципе отрицает цикличность. Однако ключевое значение имеет тот факт (уже отмеченный в разделе 3.6), что даже сейчас большая часть вселенской энтропии сосредоточена

в сверхмассивных черных дырах, расположенных в центрах галактик, и доля этой энтропии в будущем радикально увеличится. Но что в итоге произойдет с этими черными дырами? Есть все основания ожидать, что они полностью испарятся под действием хокинговского излучения.

Здесь также следует оговориться относительно хокинговского испарения: хотя детали его зависят от тонких деталей квантовой теории поля в искривленном пространстве, это явление полностью ожидаемо уже на базе общих свойств второго закона термодинамики. Здесь нужно учитывать, что колоссальная энтропия, присваиваемая черной дыре (в сущности, пропорциональная квадрату ее массы по формуле Бекенштейна — Хокинга; см. раздел 3.6), позволяет уверенно спрогнозировать хокинговскую температуру черной дыры (по сути, обратно пропорциональную ее массе; см. раздел 3.7). Отсюда можно сделать вывод о том, что черная дыра должна когда-нибудь потерять всю массу и испариться [Bekenstein, 1972, 1973; Bardeen et al., 1973; Hawking, 1975, 1976a, b]. Я ни в коей мере не оспариваю эти явления, которые, в сущности, *обусловлены* вторым законом термодинамики. Но есть важный вывод, к которому Хокинг пришел еще в самом начале своих размышлений: сама динамика черных дыр подразумевает потерю информации — или, как выразился бы я, в черной дыре теряются динамические степени свободы, и поэтому в нашей дискуссии появляется совершенно новый элемент.

На мой взгляд, потеря степеней свободы — *явное* следствие пространственно-временной геометрии коллапса черной дыры, что видно из конформных диаграмм, демонстрирующих такой коллапс, — хотя многие физики придерживаются противоположной точки зрения. На строгой конформной диаграмме с рис. 3.29 *а*

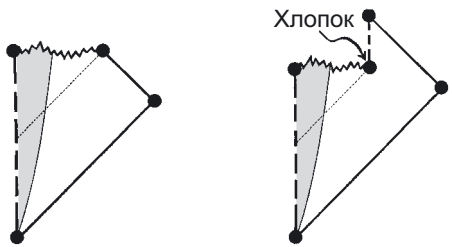


Рис. 4.16. Волнообразность зубчатой линии, обозначающей сингулярность черной дыры на этих конформных диаграммах, демонстрирует, что черная дыра обладает типичной (возможно, БКЛМ) природой, но остается пространственноподобной в соответствии с гипотезой сильной космической цензуры: *а*) типичный классический коллапс с образованием черной дыры; *б*) коллапс в черную дыру с последующим ее исчезновением под действием хокинговского испарения. Области, закрашенные темным, обозначают распределение материи (ср. с рис. 3.19 и 3.29)

из раздела 3.5 показана исходная картина сферически симметричного коллапса, предложенная Оппенгеймером и Снайдером. Как мы можем убедиться, все материальные тела после пересечения горизонта событий с неизбежностью устремляются к своему разрушению в сингулярности и никоим образом не могут передать информацию о своем внутреннем состоянии во внешний мир — если мы продолжаем придерживаться идеи классической причинности. Более того, если придерживаться принципа строгой космической цензуры (см. разделы 3.4, 3.10 и Penrose [1998a, ПкР, раздел 28.8]), общая картина типичного коллапса будет почти такой же, и я постарался показать это на конформной диаграмме с рис. 4.16 *а*, где можно себе представить, что зубчатая линия сверху представляет собой своеобразную БКЛМ-сингулярность. К тому же все материальные тела, попадающие за горизонт событий, неизбежно разрушатся на сингулярности. На рис. 4.16 *б* я немного изменил картинку, так, чтобы черная дыра испарялась под действием хокинговского излучения, — но мы видим, что ситуация с попаданием материи в дыру остается прежней. Рассматривая предположение, что на практике ситуация была бы совершенно иной, если учесть в ней локальные квантовые эффекты, необходимо принимать во внимание, на каких временны́х масштабах происходят эти явления. Тело, падающее в сверхмассивную черную дыру, после пересечения горизонта может лететь до сингулярности несколько недель и даже лет, и сложно представить себе, чтобы классическое описание такого движения к неизбежному концу оказалось бы непригодным для такого случая. Если предполагается, что информация об этом теле по мере его приближения к сингулярности может каким-то образом передаваться на уровне квантовой запутанности, а горизонт при этом находится на расстоянии световых недель и даже световых лет от тела, то возникает очень серьезный конфликт с ограничениями передачи сигнала в условиях квантовой запутанности (см. разделы 2.10 и 2.12).

Здесь я хочу обратиться к проблеме *файерволов*, предлагаемых некоторыми теоретиками в качестве альтернативы горизонтам событий черных дыр [Almheiri et al., 2013; см. также Susskind et al., 1993; Stephens et al., 1994]. В рамках этой гипотезы предлагаются аргументы, основанные на общих принципах квантовой теории поля (близкие к аргументам, связанным с хокинговской температурой) и призванные доказать, что наблюдатель, пытающийся проникнуть за горизонт событий, окажется перед огненной стеной невероятно высокой температуры и будет полностью уничтожен. На мой взгляд, это лишний раз свидетельствует о том, что базовые принципы современной квантовой механики (особенно унитарность **U**) не могут полностью соблюдаться в контексте гравитации. С точки зрения общей теории относительности локальная физика горизонта событий черной дыры не должна отличаться от локальной физики в любой другой точке. Действительно, сам горизонт даже не поддается локальному определению, поскольку его точная локализация зависит от того, сколько вещества упадет в черную дыру в будущем.

В конце концов, при всех бесчисленных подтверждениях современной квантовой теории в микромасштабах, в макромасштабе непревзойденным успехом пользуется все-таки общая теория относительности с космологической постоянной Λ.

Тем не менее большинство физиков, всерьез изучающих эту проблему, по-видимому, изрядно обеспокоены такой перспективой потери информации в черной дыре. Данное явление даже назвали *информационным парадоксом черной дыры*. Он считается парадоксом, поскольку грубо нарушает фундаментальный квантово-механический принцип *унитарности* **U**, что всерьез подорвало бы всю квантовую *веру*! Те из вас, кто смог дочитать до этой страницы, уже в курсе: я не считаю, что принцип **U** должен быть верен на всех уровнях реальности; более того, считаю, что нарушение этого принципа (которое, как правило, все равно будет происходить при измерении) неизбежно, когда в дело вступает гравитация. В случае черных дыр гравитация действительно первостепенна, и я без малейших проблем готов признать нарушение принципа унитарности **U** в квантовой динамике черных дыр. Как бы то ни было, я давно считаю проблему потери информации в черной дыре серьезным доводом в пользу того, что нарушение унитарности **U**, неизбежно происходящее при объективном **R**-процессе, должно быть обусловлено гравитацией и как раз может вызывать информационный парадокс черной дыры [Penrose, 1981; ПкР, раздел 30.9]. Соответственно, в данном случае я твердо придерживаюсь точки зрения (непопулярной среди большинства физиков — после 2004 года от нее отказался даже Хокинг [Hawking, 2005]), — согласно которой в сингулярности черной дыры информация действительно теряется. Следовательно, в результате такого процесса *объем фазового пространства* должен радикально уменьшаться, что и происходит между образованием черной дыры и ее окончательным испарением по хокинговскому принципу.

Как это помогает решить проблему со вторым законом термодинамики в КЦК? Для правильной аргументации нужно тщательно рассмотреть определение энтропии. Из раздела 3.3 мы помним, что энтропия по Больцману определяется в терминах логарифма объема фазового пространства V:

$$S = k \log V,$$

где V охватывает все состояния, представляющие рассматриваемое макросостояние, и учитывает все релевантные макроскопические параметры. Теперь, когда в рассматриваемой ситуации оказывается черная дыра, возникает вопрос, учитывать ли все те степени свободы, которые описывают предметы, упавшие в черную дыру. Предметы будут двигаться к сингулярности и в какой-то момент разрушатся — перестанут учитываться в любых процессах, происходящих за пределами черной дыры.

Когда дыра полностью испарится, можно предположить, что вместе с ней исчезнут и все поглощенные ею степени свободы. В качестве альтернативы можно попросту не учитывать такие степени свободы на всех последующих этапах существования черной дыры после падения предметов за горизонт событий. Еще один вариант интерпретации — предположить, что потеря информации будет постепенной, растянется на весь длительный жизненный цикл черной дыры. Правда, это практически ничего не меняет, так как нас интересует лишь потеря информации в целом, за всю историю черной дыры.

Как мы помним из раздела 3.3, логарифм из формулы Больцмана позволяет записать общую энтропию системы $S_{\text{общ}}$, где поглощенные степени свободы *учитываются* как сумма

$$S_{\text{общ}} = S_{\text{внеш}} + k \log V_{\text{погл}},$$

где $S_{\text{внеш}}$ — это энтропия, вычисляемая с использованием фазового пространства, в котором *не* учитываются поглощенные степени свободы, а $V_{\text{погл}}$ — объем фазового пространства, соответствующий всем поглощенным степеням свободы. Энтропия $S_{\text{погл}} = k \log V_{\text{погл}}$ по соглашению считается на момент, когда черная дыра уже испарилась, поэтому с физической точки зрения целесообразно изменить определение энтропии с $S_{\text{общ}}$ на $S_{\text{внеш}}$, когда дыры не будет.

Таким образом, убеждаемся, что, согласно КЦК, второй закон не нарушается, и свойства черных дыр, а также их испарение в значительной степени действительно можно вывести из второго закона. Однако поскольку в черной дыре утрачиваются степени свободы, второй закон в определенном смысле выходит за собственные рамки. К тому времени, как в определенном эоне полностью испарятся все черные дыры (это произойдет примерно через 10^{100} лет после большого взрыва), то определение энтропии, которое изначально было вполне приемлемо, устареет. В такую эпоху будет действовать уже новое определение, при котором энтропия гораздо ниже; и это определение станет применимым за некоторое время до перехода в следующий эон.

Для того чтобы понять, почему в следующем эоне это явление выразится в виде подавления гравитационных степеней свободы, нужно внимательнее изучить уравнения, описывающие переход от эона к эону. Используя нотацию, которой мы уже пользовались в этом разделе, получим картинку, примерно как на рис. 4.17. Здесь $\hat{\mathbf{g}}$ — это эйнштейновская физическая метрика в отдаленном будущем *предшествующего* эона, незадолго до перехода, а $\check{\mathbf{g}}$ — эйнштейновская физическая метрика сразу после большого взрыва, с которого начинается *следующий* эон. Как мы помним, гладкость геометрии в непосредственной близости от $\mathcal{I}$ выражается в терминах метрики $\mathbf{g}$, локально определяемой в узкой области,

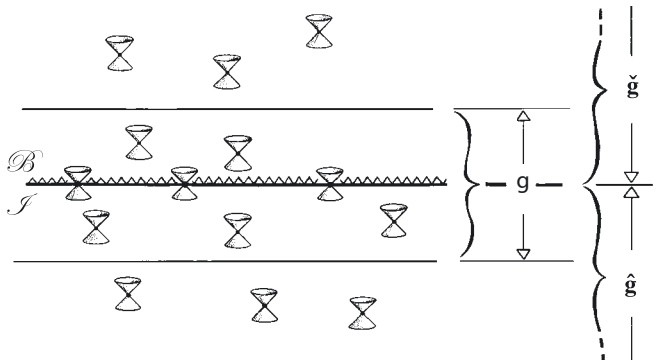


Рис. 4.17. 3-поверхность перехода соединяет предыдущий эон с последующим и является как будущей бесконечностью $\mathcal{F}$ предыдущего, так и поверхностью большого взрыва $\mathcal{F}$ последующего ($\mathcal{F} = \mathcal{B}$). Метрика $\mathbf{g}$ должна быть полностью гладкой во всем открытом «опоясывающем» регионе, содержащем поверхность перехода, и конформна эйнштейновской физической метрике $\hat{\mathbf{g}}$ до перехода ($\mathbf{g} = \omega^2\hat{\mathbf{g}}$), а также эйнштейновской физической метрике $\check{\mathbf{g}}$ после перехода ($\mathbf{g} = \Omega^2\check{\mathbf{g}}$). ω-поле полагается гладким во всем опоясывающем регионе, и считается, что в этой области справедлива обратная гипотеза: $\Omega = -\omega^{-1}$

содержащей $\mathcal{F}$, относительно которой геометрия $\mathcal{F}$ оказывается геометрией обычной пространственноподобной 3-поверхности $\mathbf{g}$, конформно относящейся к существовавшей до $\mathcal{F}$ физической метрике $\check{\mathbf{g}}$ по формуле $\mathbf{g} = \omega^2\hat{\mathbf{g}}$.

Аналогично гладкость $\mathcal{B}$ выражается в терминах некоторой метрики, которую мы вновь обозначим $\mathbf{g}$. Она локально определяется в узкой области, содержащей $\mathcal{B}$, относительно которой $\mathcal{B}$ геометрически уподобляется обычной пространственноподобной 3-поверхности, а $\mathbf{g}$ конформно относится к существующей после $\mathcal{B}$ физической метрике $\check{\mathbf{g}}$ по формуле $\mathbf{g} = \omega^2\check{\mathbf{g}}$. КЦК предполагает, что эти две метрики «$\mathbf{g}$» суть одно и то же (так называемая *опоясывающая* метрика), и такая метрика накрывает область перехода, содержащую $\mathcal{F}$ предшествующего эона, которая теперь отождествляется с $\mathcal{B}$ последующего. В итоге получаем

$$\omega^2\hat{\mathbf{g}} = \mathbf{g} = \Omega^2\check{\mathbf{g}},$$

где, кроме того, я признаю и обратную гипотезу, согласно которой Ω обратна ω, но со знаком «минус»:

$$\Omega = -\omega^{-1},$$

где ω плавно меняет знак с минуса на плюс в процессе перехода из предшествующего эона в последующий, и $\omega = 0$ на 3-поверхности пересечения ($\mathcal{F} = \mathcal{B}$) (см. рис. 4.17). Таким образом, и Ω, и ω могут оставаться положительными в тех областях, где играют роль конформных множителей.

Однако этого еще недостаточно для обеспечения единственности перехода из предыдущего эона в последующий, и остается определить, как лучше всего обеспечить такую единственность (по-видимому, при этом неизбежно нарушение симметрии, связанное с хиггсовским механизмом и повторным возникновением массы после перехода). Все эти технические вопросы выходят за рамки нашей книги, но следует отметить, что подобная процедура серьезно противоречит общему убеждению, что для понимания природы Большого взрыва во всех подробностях требуется некая теория квантовой гравитации. В данном случае мы обходимся лишь классическими дифференциальными уравнениями, и прогностический потенциал такой трактовки гораздо выше, особенно с учетом того, что никакой общепризнанной теории квантовой гравитации пока нет! Я считаю, что *нет необходимости* вступать на поле квантовой гравитации по той причине, что все колоссальные искривления пространства-времени (то есть крайне малые радиусы такой кривизны — порядка планковских величин), встречающиеся на $\mathcal{B}$, имеют форму эйнштейновской кривизны $\mathbf{G}$ (эквивалентной кривизне Риччи; см. раздел 1.1), а в таком случае они не измеряют гравитацию. Гравитационные степени свободы относятся не к $\mathbf{G}$, а к $\mathbf{C}$, и $\mathbf{C}$ поблизости от пересечения, согласно КЦК, остается полностью конечной, поэтому квантовая гравитация в данном контексте может быть совершенно неважна.

Несмотря на то что формулировка уравнений КЦК во всех деталях пока не определена, насчет распространения гравитационных степеней свободы на линии пересечения можно уверенно утверждать одно. Эта проблема довольно любопытная и тонкая, но суть ее формулируется легко и недвусмысленно: поскольку тензор Вейля $\mathbf{C}$ описывает конформную кривизну, он действительно должен быть конформно инвариантной сущностью; но есть и другая величина, которую я назову $\mathbf{K}$ и которую можно считать равной $\mathbf{C}$ в метрике $\hat{\mathbf{g}}$ предшествующего эона, так что

$$\hat{\mathbf{K}} = \hat{\mathbf{C}}\,.$$

Однако у двух этих тензоров разные конформно инвариантные интерпретации. Тогда как интерпретация $\mathbf{C}$ (в любой метрике) представляет собой конформную кривизну Вейля, $\mathbf{K}$ интерпретируется как гравитонное поле, удовлетворяющее конформно инвариантному волновому уравнению (на самом деле речь о том же уравнении из твисторной теории, что упоминалось в разделе 4.1, при спине 2, то есть $|s| = 2\hbar$, так что $d = -6$ или $+2$). В данном случае интересно, что конформная инвариантность данного волнового уравнения требует, чтобы у $\mathbf{K}$ и $\mathbf{C}$ различались значения конформного веса. Поэтому если вышеприведенное уравнение выполняется, мы находим, что в метрике $\mathbf{g}$

$$\mathbf{K} = \Omega\mathbf{C}.$$

В силу конформной инвариантности волнового уравнения **K** величина **K** достигает конечного значения на $\mathcal{F}$, и отсюда мы сразу делаем вывод о том, что **C** здесь должен обращаться в нуль (так как Ω становится бесконечной). Поскольку конформная геометрия по обе стороны от перехода должна совпадать ($\mathcal{F} = \mathcal{B}$), получаем, что **C** действительно обращается в нуль и при большом взрыве последующего эона. Следовательно, КЦК явно удовлетворяет гипотезе кривизны Вейля в ее исходной форме $\mathbf{C} = 0$, а не только в конечной форме **C**, которую мы получаем непосредственно из гипотезы Тода, касающейся единственного эона.

У нас имеются классические дифференциальные уравнения, описывающие всю информацию, достигающую $\mathcal{F}$ из предшествующего эона и попадающую в $\mathcal{B}$ последующего эона. Информация, заключенная в гравитационных волнах, достигает $\mathcal{F}$ в форме **K** и проникает в следующий эон в облике Ω. Выясняется (в соответствии с обратной гипотезой), что конформный множитель Ω должен «воплотиться» в виде нового скалярного поля в последующем эоне, и это скалярное поле определяет природу материи, возникающей при большом взрыве в начале следующего эона. Я полагаю, что Ω-поле на самом деле является первичной формой темной материи в следующем эоне — из раздела 3.4 мы помним, что в настоящее время на эту таинственную субстанцию приходится примерно 85 % материи во Вселенной. Ω-поле в последующем эоне должно интерпретироваться как своеобразная энергонесущая материя; она *должна* там быть в составе энергетического тензора этого эона в соответствии с уравнениями КЦК (которые в последующем эоне станут просто эйнштейновскими Λ-уравнениями). Она должна давать дополнительный вклад, кроме того, что мы получаем от всех безмассовых полей (например, от электромагнитного), распространяющихся из старого эона в новый — кроме гравитационного поля. Именно Ω-поле подхватывает информацию в **K** из предыдущего эона, так что информация **K** не утрачивается, а появляется в последующем эоне в виде возмущений в Ω, а не в виде гравитационных степеней свободы [Gurzadyan and Penrose, 2013].

Идея заключается в том, что в тот момент, как в последующем эоне вступает в свои права механизм Хиггса, Ω-поле приобретает массу и становится темной материей, отсылка к которой позволяет непротиворечиво объяснить астрофизические наблюдения (см. раздел 3.4). Должна существовать какая-то тесная связь между Ω-полем и механизмом Хиггса. Более того, в ходе последующего эона такая темная материя должна будет полностью распасться и превратиться в другие частицы, поэтому она не будет накапливаться от эона к эону.

Наконец, встает вопрос о проверке КЦК на результатах наблюдений. Фактически вся модель довольно жесткая, поэтому должно быть много областей, где выкладки КЦК можно было бы убедительно экспериментально проверить. Работая над

книгой, я сосредоточился всего на двух свойствах этой модели. Во-первых, я рассматриваю столкновения сверхмассивных черных дыр в эоне, предшествовавшем нашему. В истории каждого эона такие столкновения, в принципе, должны происходить довольно часто. Так, в нашем эоне Млечный Путь несется навстречу Туманности Андромеды. Они столкнутся примерно через 10^9 лет, и достаточно велик шанс, что наша галактическая черная дыра, масса которой составляет $\sim 4 \times 10^6$ солнечных, по спирали устремится к центральной черной дыре Туманности Андромеды, масса которой $\sim 10^8$ солнечных. Такие столкновения должны приводить к почти мгновенным колоссальным всплескам гравитационных волн, и такая энергия, согласно КЦК, должна проявляться в начальных импульсных возмущениях в первичном распределении темной материи в последующем эоне. Такие события из предшествующего эона в нашем проявлялись бы как (зачастую концентрические) *кольцевые* возмущения в картине распределения РИ, которые были бы вполне различимы. Есть серьезные признаки того, что такие явления в РИ действительно просматриваются, — об этом свидетельствуют данные космических обсерваторий WMAP и Планк (см. раздел 3.1), проанализированные двумя независимыми группами исследователей [Gurzadyan and Penrose, 2013; Meissner et al., 2013]. По-видимому, это и есть четкие доказательства в пользу существования более раннего — и при этом удивительно неоднородного — эона в истории Вселенной, согласующиеся с гипотезой КЦК. Если такая интерпретация верна, то напрашивается вывод, что в эоне, предшествовавшем нашему, супермассивные черные дыры были распределены достаточно неравномерно. Хотя модель КЦК подобного и не предполагала, эти данные легко в нее вписываются. Гораздо сложнее объяснить такую неоднородность при помощи привычных инфляционных представлений, где температурные флуктуации РИ объясняются случайными квантовыми факторами.

Есть и еще одно эмпирическое следствие КЦК, на которое я обратил внимание с подсказки Пола Тода в начале 2014 года. Дело в том, что КЦК позволяет объяснить источник *первичных магнитных полей*. В самом начале Большого взрыва (безотносительно КЦК) явно должны были существовать магнитные поля — дело в том, что магнитные поля наблюдаются в обширных *войдах*, пустынных областях межгалактического пространства (см. [Ananthaswamy, 2006]). Существование галактических и межгалактических магнитных полей обычно связывают с динамическими галактическими процессами, происходящими в *плазме* (не связанных в атомы протонах и электронах, сосуществующих в обширных регионах пространства), и эти процессы растягивают и усиливают магнитные поля, издавна существующие в пределах и за пределами галактик. Однако такие процессы не могут протекать там, где нет галактик, то есть в войдах, поэтому обнаруженные в них магнитные поля остаются загадкой. Напрашивается вывод, что такие магнитные поля должны быть *первозданными*, то есть что они существовали уже на момент Большого взрыва.

По версии Тода, такие магнитные поля могли в целости и сохранности просочиться через Большой взрыв из тех областей, где в предыдущем эоне существовали галактические скопления. В конце концов магнитные поля подчиняются уравнениям Максвелла, которые, как упоминалось ранее, *конформно инвариантны*. Следовательно, такие поля могли попасть из отдаленного будущего предыдущего эона в Большой взрыв нашего.

Такие первичные магнитные поля, возможно, являются источниками так называемых *B-мод* в поляризации фотонов в РИ; по-видимому, именно такие моды удалось наблюдать команде BICEP2, о чем сообщалось 17 марта 2014 года [Ade et al., 2014] как о «явной улике» в пользу инфляции! На момент написания книги значимость этих наблюдений вызывает некоторый скепсис — считается, что была недооценена роль космической пыли. Тем не менее КЦК предлагает альтернативное объяснение таких мод, и интересно, какой вариант будет точнее соответствовать фактам. (Однако на момент написания книги оставались сомнения в самой интерпретации данных BICEP2, поскольку воздействие межгалактической пыли как следует не учитывалось [Mortonson and Seljak, 2014]). В заключение отмечу, что наличие таких галактических скоплений в предыдущем эоне, согласно КЦК, вполне можно проверить, изучая столкновения сверхмассивных черных дыр. Поэтому два упомянутых здесь эмпирических аспекта КЦК определенным образом связаны друг с другом. Все это подводит нас к интересным задачам, также требующим экспериментальной проверки, и интересно посмотреть, насколько КЦК оправдает связанные с ней ожидания.

4.4. Личное послесловие

Несколько лет тому назад я давал интервью одному голландскому журналисту, и он спросил, считаю ли я себя «диссидентом». Думаю, отвечая на этот вопрос, я вложил в слово «maverick» несколько иное значение, нежели тот журналист (причем *Краткий оксфордский словарь*, мне кажется, поддерживает скорее мою трактовку). Я считаю, что диссидент — это тот, кто не просто идет наперекор устоявшимся взглядам, но и тот, кто делает это в известной мере умышленно, сторонясь толпы. Я ответил интервьюеру, что себя таковым не считаю, и в основном — что касается базовых физических теорий, лежащих в основе современных представлений об устройстве реальности, — я довольно консервативен и придерживаюсь гораздо более традиционных взглядов, чем многие из моих знакомых, которые очертя голову стремятся расширять границы научных представлений о мире.

Хороший пример, помогающий понять мою точку зрения, — общая теория относительности Эйнштейна (с космологической постоянной Λ). Она абсолютно

мила мне как красивая классическая теория тяготения и пространства-времени, и ей можно полностью доверять до тех пор, пока мы не приблизимся вплотную к сингулярности, где кривизна пространства становится немыслимой и теория Эйнштейна в ее привычном виде перестает работать. Определенно следствия общей теории относительности устраивают меня гораздо больше, чем даже самого Эйнштейна на закате жизни. Если, согласно теории Эйнштейна, должны существовать странные объекты, состоящие практически из одной пустоты и способные поглощать целые звезды, то пусть так — но сам Эйнштейн отверг такой феномен, именуемый сегодня *черной дырой*, и сам же пытался доказать, что такие необратимые гравитационные коллапсы определенно происходить не могут. Очевидно, Эйнштейн считал, что его собственная общая теория требует фундаментальной доработки даже на классическом уровне; ведь в последние годы (работая в Принстоне) он потратил массу времени, пытаясь модифицировать свою величественную теорию относительности разными (зачастую математически непривлекательными) способами, чтобы увязать ее с электромагнетизмом, в основном игнорируя остальные физические поля.

Разумеется, как я писал в разделе 4.3, мне очень нравится достраивать общую теорию относительности Эйнштейна в необычных направлениях, поскольку, если строго ее придерживаться, Большой взрыв должен быть началом всего, а какие-либо явления до этого знаменательного события в великую теорию Эйнштейна не укладываются. Однако хотелось бы указать, что мои дополнения к этой теории весьма скромные и позволяют всего лишь немного расширить ее концепции, а значит, и область применения. КЦК отлично согласуется с общей теорией относительности, содержащей Λ, в том виде, как ее сформулировал Эйнштейн в 1917 году, и совершенно не противоречит такой трактовке этой теории, которую можно найти в старых книгах по космологии (хотя она учитывает такие источники материи во Вселенной, которые в этих книгах не упоминаются). Более того, КЦК принимает эйнштейновскую космологическую постоянную Λ именно в том виде, в каком ее ввел сам Эйнштейн. КЦК не вводит никаких таинственных сущностей вроде «темной энергии», «ложного вакуума» или «квинтэссенции», описываемых уравнениями, которые могут радикально отклоняться от классической теории Эйнштейна.

Даже когда речь заходит о квантовой механике — как вы помните, в разделе 2.13 я скептически отзываюсь о слепой квантовой вере, которой, по-видимому, придерживаются столь многие физики, — я полностью признаю практически все ее самые странные допущения, например нелокальность, свойственную для эффектов ЭПР (Эйнштейна — Подольского — Розена). Моя лояльность начинает пошатываться лишь тогда, когда эйнштейновская кривизна пространства-времени вступает в противоречие с квантовыми принципами. Соответственно, меня устраивают все

эти эксперименты, продолжающие подтверждать странность квантовой теории, поскольку пока что все они далеко не дотягивают до уровня, где могли бы обозначиться противоречия между ними и общей теорией относительности.

Когда речь заходит о модных аспектах высших пространственных измерений (и в несколько меньшей степени о суперсимметрии), я вновь становлюсь очень консервативен, но сейчас должен сделать одно признание. Я выдвигал здесь возражения против дополнительных пространственных измерений практически исключительно с точки зрения сложностей, связанных с заведомо избыточной функциональной свободой, заключенной в этих измерениях. Я действительно считаю, что это веские возражения, и ни разу не видел, чтобы сторонники дополнительных измерений всерьез их обсуждали. Однако еще есть *реальные*, мои глубоко личные возражения против дополнительных измерений!

Что же это за возражения? Иногда журналисты или просто друзья и знакомые спрашивали меня, почему я возражаю против введения в теорию дополнительных измерений — и на это я могу ответить, что некоторые суждения высказываю публично, а другие оставляю при себе. Публичные возражения связаны в основном с проблемами, возникающими из-за избыточной функциональной свободы, а что насчет моих личных? Здесь я должен сделать небольшой экскурс в историю — рассказать, как развивал эти идеи.

Я впервые принялся разрабатывать концепции, которые сочетали бы пространственно-временное описание и квантовые принципы, еще в начале 1950-х годов, будучи магистрантом-математиком, а затем — когда работал научным сотрудником в Кембриджском колледже Святого Иоанна. В ту пору меня серьезно мотивировали долгие дискуссии с моим другом и наставником Деннисом Сиамой, а также с некоторыми другими, например с Феликсом Пирани. Я вдохновлялся лекциями некоторых выдающихся преподавателей, особенно Германа Бонди и Поля Дирака. Кроме того, мои старшие курсы прошли в Университетском колледже Лондона, где я был просто очарован силой и волшебством комплексного анализа и геометрии и с тех пор убежден, что эта магия также наверняка лежит глубоко в основе нашего мира. Я видел, что в контексте двухкомпонентного спинорного формализма (тема, которую объяснил на своих лекциях Дирак) существует не только тесная связь между трехмерной пространственной геометрией и квантово-механическими амплитудами, но и довольно своеобразная связь между группой Лоренца и сферой Римана (см. раздел 4.1). Взаимоотношения обоих этих типов требовали конкретной пространственно-временной размерности — именно такой, которую мы наблюдаем вокруг, но мне не удавалось выявить ключевую взаимосвязь между ними, которую впоследствии (в 1963 году) продемонстрировала теория твисторов.

Для меня это была кульминация многолетних поисков, и хотя были и другие мотивы, благодаря которым эти идеи развивались в совершенно конкретном направлении [Penrose, 1987c], принципиальная «лоренцева» комбинация трехмерности пространства с одномерностью времени была глубоко вплетена во всю эту затею. Более того, многие позднейшие разработки (например, твисторное представление волновых функций в безмассовых полях, о чем шла речь в разделе 4.1), казалось, подтверждают ценность этих мотивировок. Поэтому когда я узнал, что теория струн, изначально отчетливо меня привлекавшая, отчасти потому, что именно она одной из первых стала работать с римановыми поверхностями, стала развиваться с опорой на все эти дополнительные пространственные измерения, я просто пришел в ужас, меня далеко не привлекали романтические красоты Вселенной с высшими размерностями. Я счел невозможным, чтобы Природа могла отказаться от всех этих прекрасных связей с лоренцевым 4-пространством, — и по-прежнему так думаю.

Разумеется, читатель может счесть мою упрямую приверженность лоренцеву 4-пространству очередным примером моего личного консерватизма в основах науки. Но в самом деле, когда физики правильно развивают свои идеи, я не вижу никаких причин их менять. Только когда физики не совсем правы или, пожалуй, совсем не правы, я начинаю волноваться. Конечно же, в любой теории могут потребоваться фундаментальные изменения, даже когда она так красиво работает. Характерный пример — ньютоновская механика. Я считаю, что примерно такой же случай возникает и с квантовой теорией. Но это ничуть не разубеждает меня в том, что обе эти величественные теории тормозят развитие фундаментальной науки. Прошло почти два века, прежде чем стало ясно, что ньютоновская модель Вселенной нуждается в доработке и в нее нужно включить сплошные поля Максвелла, а потом еще полвека — чтобы в дело постепенно пошли изменения, связанные с теорией относительности и с квантовой теорией. Интересно посмотреть, сможет ли квантовая теория так долго продержаться без изменений.

Наконец, скажу еще несколько слов о том, как модность научных идей связана с их живучестью. Я искренне восхищаюсь тем и поражаюсь тому, как современные технологии, и в особенности интернет, обеспечивают мгновенный доступ к постоянно расширяющемуся корпусу научных знаний. Однако я боюсь, что сама эта доступность может загнать нас в тиски моды. Сейчас так много доступной информации, что исключительно сложно понять, какие материалы из всего этого множества содержат новые идеи, действительно заслуживающие внимания. Как судить, что важно, а что все время на виду только лишь благодаря собственной популярности? Как продраться через множество, которого просто *много*, но которое бедно идеями, старыми ли, новыми, идеями, которые

обладали бы подлинной важностью, непротиворечивостью и истинностью? Это сложный вопрос, и у меня нет на него внятного ответа.

Однако роль моды в науке определенно не нова, и в разделе 1.1 я немного описал, какова она была в прошлом. Научиться вырабатывать собственные непротиворечивые суждения, не ведясь без нужды на моду, — это баланс, достичь которого нелегко. Мне в этом отношении повезло — я рос, вдохновляясь примером своего исключительно талантливого отца Лионеля. Он был биологом и специализировался в области генетики человека, но, вообще говоря, обладал обширными интересами и умениями: в математике, искусстве и музыке, а также имел литературный талант. Правда, боюсь, что его личностные качества оставляли желать лучшего, поэтому он так и не смог наладить семейные отношения, несмотря на то, как нам с ним было интересно и какое образование мы получили, приобщаясь к его разнообразным интересам и оригинальным идеям. Вообще, атмосфера в нашей семье была подчеркнуто интеллектуальной, и я многому научился от моего любимейшего старшего брата Оливера — особенно в области физики.

Лионель явно отличался изрядным свободомыслием, и если он считал, что некоторое общепринятое убеждение ошибочно, то не стеснялся на это указать. Хорошо помню случай, как один из его коллег решил поместить на обложке своей книги одно знаменитое генеалогическое древо. Считалось, что это известное семейство демонстрирует классический пример наследования Y-хромосомы, поскольку страдает заболеванием, напрямую передаваемым больным отцом каждому из сыновей. Такая тенденция прослеживалась на протяжении нескольких поколений, причем болезнь ни разу не поражала женщин. Речь шла о тяжелом кожном заболевании (*иглистый ихтиоз*), а больного, страдавшего этим недугом, иногда называли «*человек-дикобраз*». Лионель сказал коллеге, что просто не верит в такое генеалогическое древо, поскольку не признает, что это заболевание действительно передается через Y-хромосому. Кроме того, таких больных показывали в цирках, и Лионелю казалось весьма вероятным, что владельцам цирков было бы выгодно продвигать историю о прямом наследовании такого уродства от отца к сыну. Коллега всерьез усомнился в том, что скепсис Лионеля оправдан, поэтому отец взялся доказать свою правоту и в компании моей матери Маргарет отправился изучать старые церковные метрики, чтобы проверить, как *на самом деле* выглядит генеалогическое древо человека-дикобраза. Через несколько недель он гордо представил совсем иное, гораздо более правдоподобное родословное древо, продемонстрировав, что болезнь вообще никак не связана с наследованием Y-хромосомы, но вполне может быть объяснена как напрямую наследуемое доминантное заболевание.

Мне всегда казалось, что Лионель обладал мощным инстинктивным чутьем насчет того, какие идеи должны быть истинны (хотя *мог* и ошибаться). Его чутье срабатывало не только в естественных науках, а, например, и в проблеме шекспировского авторства. Лионель прочел книгу Томаса Луни [1920] о том, что автором шекспировских пьес на самом деле был Эдуард де Вер, 17-й граф Оксфорд, и книга показалась ему такой убедительной, что он даже пытался проверить авторство при помощи статистического анализа, взяв за основу тексты, действительно принадлежавшие Веру, и сравнив их с пьесами (результаты получились несколько неубедительными). Большинство коллег Лионеля считали, что он слишком упорствует в своем мнении. Мне же эти доводы против общепризнанного авторства представлялись очень сильными (я считал крайне маловероятным, что у автора этих великих произведений не было книг и нигде не сохранилось следов его почерка, если не считать нескольких неразборчивых автографов, но я не имел ни малейшего понятия, кто же мог быть истинным автором). Интересно отметить, что сильные аргументы в пользу авторства Вера привел в своей недавно опубликованной книге Марк Андерсон [2005]. Независимо от того, как сложно бывает изменить *научные* взгляды, которые уже считаются общепризнанными, я думаю, что сделать то же самое в области *литературы*, — особенно в случае такой догматической веры, с которой связаны огромные коммерческие интересы, — *неизмеримо* сложнее!

ПРИЛОЖЕНИЕ А

Математическое приложение

А.1. Итерационные степени

В этом разделе я хочу поговорить о возведении чисел в *степень*. Возведение в степень — это умножение числа на это же число некоторое количество раз. Следовательно, запись

$$a^b,$$

где a и b — положительные целые числа, означает a умножить на себя всего b раз (так что $a^1 = a$, $a^2 = a \times a$ и $a^3 = a \times a \times a$ и т. д.); следовательно, $2^3 = 8$, $2^4 = 16$, $2^5 = 32$, $3^2 = 9$, $3^3 = 27$, $4^2 = 16$, $5^2 = 25$, $10^5 = 10\,000$ и т. д. Кроме того, можно без труда распространить этот принцип и на такие случаи, где ни a, ни b не обязаны быть положительными, если $a \neq 0$ (например, $a^{-2} = 1 / a^2$), а также на случаи, где ни a, ни b не обязаны быть целыми (то есть они могут быть вещественными или даже *комплексными* числами — об этом мы поговорим далее, в разделах А.9 и А.10). (Тут, правда, могут возникать проблемы многозначности (см., например, ПкР, раздел 5.4).) Мелкое замечание, касающееся терминологии, которой я строго придерживаюсь в этой книге: вместо терминов вроде «миллиард», «триллион» или «квадриллион», которые несколько неинформативны (отчасти это связано с неоднозначностью их использования, по-прежнему актуальной), а также достаточно неудобны при обсуждении очень больших чисел, встречающихся в этой книге (особенно в главе 3), я систематически использую экспоненциальную нотацию, например 10^{12}, для любых чисел больше миллиона (10^6).

Все это достаточно очевидно, поэтому читателю может быть интересна такая операция на уровень выше, например в случае с величиной

$$a^{b^c}.$$

Поясню, что означает эта нотация. Она *не* означает $(a^b)^c$, иными словами, a^b в степени c, по той простой причине, что мы вполне могли бы записать такую

величину без повторяющихся степеней, как a^{bc} (то есть a в степени $b \times c$). Выражение a^{b^c}, в свою очередь, должно означать (обычно гораздо более крупную) величину

$$a^{(b^c)},$$

иными словами, a в степени bc. Следовательно, $2^{2^3} = 2^8 = 256$, а вовсе не $(2^2)^3 = 64$.

Теперь я хочу сделать очень простое замечание насчет таких величин, а именно: в случае с достаточно большими числами a, b, c оказывается, что a^{b^c} почти не зависит от значения a, а важнее всего значение c. (Еще более захватывающие сведения, касающиеся этих вопросов, изложены в работах [Littlewood, 1953 and Bollobás, 1986, p. 102—103].) Это становится очевидным, если переписать a^{b^c} в виде логарифма. Поскольку я математик и немного пурист, я предпочитаю использовать *натуральные* логарифмы, то есть под обозначением log я на самом деле имею в виду натуральный логарифм $\log_e$ (хотя он часто обозначается ln). Если вы предпочитаете обычные «логарифмы по основанию 10», пропустите следующий абзац и начинайте читать после него. Но братья-пуристы понимают, что натуральный логарифм — это просто функция, *обратная* стандартной *экспоненциальной* функции. Таким образом, вещественное число

$$y = \log x$$

(для положительного вещественного числа x) определяется при помощи эквивалентного (обратного) уравнения

$$e^y = x,$$

где e^y — стандартная экспоненциальная функция, иногда записываемая в виде «exp y» и определяемая при помощи суммы бесконечного ряда:

$$\exp y = e^y = 1 + \frac{y}{1!} + \frac{y^2}{2!} + \frac{y^3}{3!} + \frac{y^4}{4!} + \ldots,$$

где

$$n! = 1 \times 2 \times 3 \times 4 \times 5 \times \ldots \times n$$

(см. рис. А.1). Подставив в вышеприведенный ряд $y = 1$, находим:

$$e = e^1 = 2{,}7182818284590452\ldots.$$

Мы еще вернемся к этому ряду в разделе А.7.

Следует отметить (что довольно примечательно): нотация (e^y) согласуется с предыдущими выкладками, а именно: если y — положительное целое число, то e^y действительно равно числу e, умноженному на себя y раз. Более того, продолжает соблюдаться и правило приведения сложения к умножению , которому удовлетворяют степени, то есть

$$e^{y+z} = e^y e^z.$$

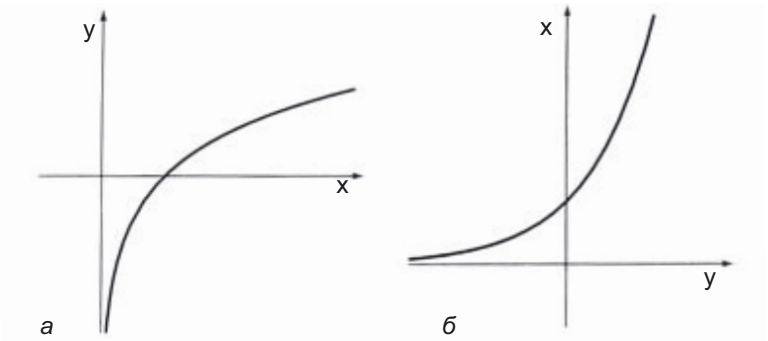


Рис. А.1. Логарифмическая функция $y = \log x$ (*а*) является обратной для экспоненциальной функции $x = e^y$ (*б*) (обозначения осей нестандартные). Обратите внимание: для перехода к обратной функции мы меняем местами оси x и y, то есть делаем отражение по диагонали: $y = x$

Из этого следует, что, поскольку log — это функция, обратная exp, можно записать:

$$\log(ab) = \log a + \log b$$

(что эквивалентно вышеприведенной записи, если подставить $a = e^y$ и $b = e^z$). Также имеем:

$$a^b = e^{b \log a}$$

(поскольку $e^{\log a} = a$, так что $\mathbf{e}^{b \log a} = \left(\mathbf{e}^{\log a}\right)^b = a^b$), и из этого следует

$$a^{b^c} = e^{\mathbf{e}^{c \log b + \log\log a}}$$

(так как $e^{c \log b + \log\log a} = e^{c \log b} e^{\log\log a} = b^c \log a$). Поскольку функция $\log x$ возрастает очень медленно, а функция $\log\log x$ — еще медленнее, то обнаруживаем, что при достаточно больших значениях a, b и c именно c играет наибольшую роль при определении величины $c \log b + \log\log a$, а следовательно, и величины a^{b^c}, а значением a можно практически полностью пренебречь.

Вероятно, неподготовленному читателю будет проще следить за мыслью, если я буду говорить о логарифмах по основанию 10, то есть $\log_{10}$ (в научно-популярном изложении у такого варианта есть свое преимущество — не приходится объяснять, что такое «e»!). Здесь я буду пользоваться нотацией Log в значении $\log_{10}$. Соответственно, вещественное число $u = \mathrm{Log}\, x$ (для любого положительного вещественного числа x) определяется при помощи эквивалентного обратного уравнения

$$10^u = x,$$

и мы имеем

$$a^b = 10^{b\mathrm{Log}\, a},$$

откуда (как и выше) следует, что

$$a^{b^c} = 10^{10^{c\mathrm{Log}b + \mathrm{Log}\mathrm{Log}a}}.$$

Теперь без труда можно показать, как медленно возрастает Log x, отметив, что

$$\mathrm{Log}\ 1 = 0, \quad \mathrm{Log}\ 10 = 1, \quad \mathrm{Log}\ 100 = 2,$$
$$\mathrm{Log}\ 1000 = 3, \quad \mathrm{Log}\ 10\,000 = 4 \quad \text{и т. д.}$$

Соответственно крайне медленное увеличение Log Log x иллюстрируется на следующем примере:

$$\mathrm{Log}\,\mathrm{Log}\ 10 = 0, \quad \mathrm{Log}\,\mathrm{Log}\ 10000000000 = \mathrm{Log}\,\mathrm{Log}\ 10^{10} = 1,$$
$$\mathrm{Log}\,\mathrm{Log}\ (\text{один гугол}) = \mathrm{Log}\,\mathrm{Log}\ 10^{100} = 2, \quad \mathrm{Log}\,\mathrm{Log}\ 10^{1000} = 3 \quad \text{и т. д.},$$

причем, как мы помним, 10^{1000} записывается как единица, за которой следует *тысяча* нулей, тогда как *гугол* — это число, в котором за единицей следует сто нулей.

В главе 3 мы сталкивались с некоторыми очень большими числами, например $10^{10^{124}}$. Это конкретное число (позволяющее грубо оценить, насколько «особенной» была наша Вселенная на момент Большого взрыва) и аргументы, изложенные в главе 3, позволяют рассматривать меньшее число $e^{10^{124}}$. Однако исходя из вышеизложенного, находим, что

$$e^{10^{124}} = 10^{10^{124} + \mathrm{Log}\mathrm{Log}e}.$$

Оказывается, что значение величины Log Log e равно примерно −0,362, поэтому, как можно убедиться, если заменить в левой части e на 10, это равноценно замене 124 примерно на 123,638 в верхней степени, и при округлении этой степени до

ближайшего целого числа все равно получается 124. На самом деле примерное число 124 в этом выражении известно достаточно приблизительно, и более «верное» значение вполне может составлять 125 или 123. Во многих более ранних работах я действительно обозначал эту «степень своеобразия» числом $e^{10^{123}}$, и это число подсказал мне Дон Пейдж примерно в 1980 году; тогда мы еще не вполне представляли себе, насколько во Вселенной преобладает темная материя (см. раздел 3.4). Более крупное значение $e^{10^{124}}$ (или $\mathbf{e}^{10^{125}}$) дается с учетом темной материи. Следовательно, замена e на 10 здесь вообще почти не играет роли! В рассматриваемом случае значение b не очень велико ($b = 10$), поэтому член Log Log e все-таки имеет значение, но непринципиальное, поскольку 124 все равно намного больше, чем Log Log e.

Еще одно свойство столь больших чисел заключается в следующем: если перемножить или разделить пару таких чисел, верхние степени которых хотя бы незначительно различаются, то большее число, вероятно, полностью поглотит меньшее, так что мы попросту сможем игнорировать меньшее число в таком произведении или частном! Чтобы в этом убедиться, для начала рассмотрим

$$10^{10^x} \times 10^{10^y} = 10^{10^x + 10^y} \text{ и } 10^{10^x} \div 10^{10^y} = 10^{10^x - 10^y}.$$

Далее отметим, что если $x > y$, то (нижняя) степень 10 в произведении равна $10^x + 10^y = 1000...001000...00$, а (нижняя) степень в частном — $10^x - 10^y = 1000...000999...99$. Здесь в случае произведения в первом блоке «000...00» $x - y - 1$ нулей, а во втором блоке «000...00» y нулей. Что касается частного, в первом блоке $x - y$ нулей, а во втором $y - 1$ девяток. Естественно, если x гораздо больше y, то единица в середине первого числа или девятки в конце практически не играют роли (правда, здесь нужно с осторожностью формулировать, что такое «не играет роли»), поскольку если вычесть одно число из другого, то разность все равно окажется огромным числом!) Даже если $x - y$ равно всего 2, то нижняя степень изменится не более чем на 1 %, причем она изменится гораздо меньше, если $x - y$ будет гораздо больше 2. Таким образом, можно игнорировать 10^{10^y} в произведении $10^{10^x} \times 10^{10^y}$. Аналогично и при делении меньшее число 10^{10^y} можно не учитывать в выражении $10^{10^x} \div 10^{10^y}$. В разделе 3.5 я показал, в каких случаях это может иметь значение.

А.2. Функциональная свобода полей

В контексте главы 1 более важны числа вида a^{b^c} в «пределе», где a и b становятся бесконечными, и я буду записывать такую величину как ∞^{∞^n}. Возможен вопрос: а что это, в сущности, означает? И почему такие величины важны в физике?

Для ответа на первый вопрос лучше сначала ответить на второй. А для этого необходимо напомнить, что значительная часть физики описывается в терминах так называемых *полей*. Итак, что же такое «поле» с точки зрения физика?

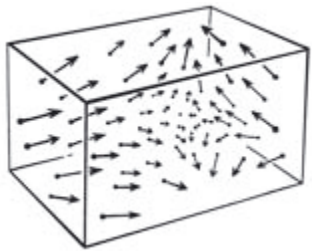


Рис. А.2. Магнитное поле в обычном трехмерном пространстве — хороший пример физического (векторного) поля

Для того чтобы «почувствовать», что такое физическое поле, удобно рассмотреть магнитное поле. В каждой точке пространства у магнитного поля есть *направление* (определяемое двумя углами, указывающими, допустим, наклонение в ориентации восток — запад и верх — низ) и *напряженность* (еще одно число) — всего три числа. Проще говоря, можно представить три компонента *векторной* величины, выражаемых вещественными числами, и три этих компонента полностью характеризуют магнитное поле в заданной точке (см. рис. А.2). (Магнитное поле является *векторным*, данная концепция подробнее описана в разделе А.7.) Итак, сколько магнитных полей потенциально может существовать в пространстве? Очевидно, бесконечное множество. Но бесконечность — очень грубая оценка, и я хотел бы ее значительно уточнить.

Для начала будет полезно представить себе упрощенную модель, где континуум $\mathbb{R}$, состоящий из всех возможных вещественных чисел, заменяется конечным множеством **R**, состоящим всего из N элементов, где N — исключительно большое положительное целое число. Поэтому мы не рассматриваем весь континуум, а аппроксимируем дискретное множество очень плотно расположенных точек (все в одну линию). Три вещественных числа, которые описывают наше магнитное поле в точке P, теперь можно считать всего тремя элементами **R**. Итак, будет всего N вариантов для первого из этих чисел, N вариантов для второго и N вариантов для третьего — всего:

$$N \times N \times N = N^3.$$

В этой упрощенной модели в любой точке P в пространстве будет N^3 возможных магнитных полей. Но мы хотим знать, сколько здесь вообще возможных полей,

причем поле может произвольно варьироваться от точки к точке в пространстве. В этой модели каждое измерение в континууме пространства-времени также должно описываться конечным множеством **R**, так что каждую из трех пространственных координат (обычно это вещественные числа, обозначаемые через x, y, z) теперь можно считать элементом **R**, поэтому в нашей упрощенной модели также будет N^3 возможных точек в пространстве. В любой конкретной точке P есть N^3 возможных магнитных полей. В точках P и Q будет по N^3 возможных магнитных полей в каждой, и, следовательно, в двух точках будет $N^3 \times N^3 = (N^3)^2 = N^6$ возможных значений поля (предполагается, что значения полей в различных точках не зависят друг от друга). В трех разных точках будет $(N^3)^3 = N^9$ возможностей, в четырех $(N^3)^4 = N^{12}$ и т. д. Следовательно, во всех N^3 точках у нас будет всего

$$(N^3)^{N^3} = N^{3N^3}$$

возможностей для всех магнитных полей, присутствующих в пространстве в этой упрощенной модели.

Этот пример получился немного путаным, поскольку число N^3 используется в двух разных ипостасях, где первая тройка — число компонентов, имеющихся у магнитного поля в каждой точке, а вторая тройка — число измерений в пространстве. У полей других типов число компонентов может отличаться. Например, температура или плотность вещества в определенной точке пространства будут иметь по одному компоненту, тогда как у *тензорных* величин, например у напряжений в твердом теле, компонентов на точку будет больше. Можно рассмотреть поле с некоей c-компонентной величиной вместо нашего 3-компонентного магнитного поля, и тогда в упрощенной модели получим всего

$$(N^c)^{N^3} = N^{cN^3}$$

возможностей для такого поля. Также можно рассмотреть пространство с размерностью d, отличающееся от знакомого нам трехмерного пространства. Затем, если в нашей упрощенной модели пространство будет d-мерным, число возможных c-компонентных полей составит

$$(N^c)^{N^d} = N^{cN^d}.$$

Разумеется, нас более интересует реальная физика, а не такая упрощенная модель, а в реальной физике число N считается бесконечным, хотя необходимо учитывать, что на самом деле мы не знаем точной математической структуры истинной физики природы, так что выражение *реальная физика* здесь означает конкретные математические модели, используемые в современных исключитель-

но успешных теориях. В этих успешных теориях N действительно бесконечно, и если заменить в вышеприведенной формуле N на ∞, то получим

$$\infty^{c\infty^d}$$

различных потенциально возможных c-компонентных полей в d-мерном пространстве.

В конкретной физической ситуации, с которой я начал этот экскурс, а именно при рассмотрении числа возможных различных конфигураций магнитных полей, которые могут существовать в полном пространстве, имеем $c = d = 3$, поэтому получим ответ

$$\infty^{3\infty^3}.$$

Правда, необходимо учитывать, что (в упрощенной модели) мы исходили из предположения, что величины полей в различных точках *не зависят* друг от друга. В описываемом контексте, если рассматривать магнитные поля в пространстве, такое допущение в известном смысле неверно. Ведь есть ограничение, которому удовлетворяют магнитные поля, — так называемое *уравнение связи* (записываемое обычно в виде div **B** = 0, где **B** — вектор магнитного поля; это пример *дифференциального* уравнения, они вкратце рассматриваются в разделе А.11). Это означает: два разных магнитных полюса — северный и южный — не могут существовать по отдельности, действуя как независимые источники магнитного поля, и, насколько позволяет судить современная физика, автономных магнитных полюсов действительно не существует (см., однако, раздел 3.9). Данная связь накладывает на магнитное поле ограничение, соотносящее друг с другом величины этого поля в разных точках пространства. Здесь есть и более специфичная подоплека: не все пространственные компоненты магнитного поля являются независимыми, но один из трех (на наш выбор) фиксируется двумя другими, которые и определяют, как этот компонент действует в некотором двумерном субрегионе S трехмерного пространства. Поэтому компонент $3\infty^3$ в степени на самом деле должен иметь вид $2\infty^3 + \infty^2$, но поскольку можно считать, что поправка ∞^2 к степени будет полностью поглощена гораздо более крупным значением $2\infty^3$, можно попросту забыть о нем и выразить *функциональную свободу* магнитных полей в обычном трехмерном пространстве (подчиняющуюся этой связи) в виде

$$\infty^{2\infty^3}.$$

Уточнив данную нотацию с учетом трудов Картана (см. [Bryant et al., 1991; Cartan, 1945], особенно разделы 68 и 69 на с. 75—76 оригинального издания),

действительно можно усмотреть смысл в выражениях вроде $\infty^{2\infty^3+\infty 2}$, где экспонента может считаться полиномом степени «∞», коэффициенты которого — неотрицательные целые числа. В этом примере у нас две свободные функции трех переменных плюс одна свободная функция двух переменных. Однако я в этой книге обойдусь без уточнения такой нотации.

Естественно, придется прояснить некоторые моменты относительно этой полезной нотации (которой, по-видимому, впервые воспользовался, выдающийся и исключительно оригинальный американский физик Джон А. Уилер [1960] (см. также [Penrose, 2003, p. 185—201; ПкР, раздел 16.7]). Во-первых, должен отметить, что к этим бесконечным числам неприменимо понятие *мощности* в обычном (канторовском) смысле, позволяющее сравнивать бесконечные множества. Вполне возможно, что некоторые читатели знакомы с удивительной канторовской теорией бесконечных чисел. Если вы не сталкивались с этой теорией — ничего страшного. Я упоминаю здесь теорию Кантора лишь для того, чтобы подчеркнуть ее отличие от излагаемого здесь материала. Тем не менее если вы знаете теорию Кантора, то следующие ремарки, возможно, помогут вам понять, как возникли такие отличия.

В канторовской системе бесконечных чисел, именуемых *кардинальными*, кардинальное число множества $\mathbb{Z}$ соответствует количеству различных целых чисел в этом множестве и обозначается через $\aleph_0$ («алеф-нуль »), соответственно *количество* различных целых чисел в множестве равно $\aleph_0$. В таком случае количество различных *вещественных чисел* будет равно $2^{\aleph_0}$, что обычно записывается как $C\ (=2^{\aleph_0})$ (вещественные числа можно представлять в виде бесконечных последовательностей двоичных цифр, например 10010111.0100011..., что, грубо говоря, равно количеству альтернатив, соответствующему $\aleph_0$, то есть $2^{\aleph_0}$). Однако такого уточнения в данном контексте недостаточно. Например, если попытаться представить размер d-мерного пространства как N^d, по мере увеличения N до $\aleph_0$, то в канторовской модели мы всегда будем получать $\aleph_0$ независимо от того, как велико будет d. Действительно, в канторовской нотации $(\aleph_0)^d = \aleph_0$ для любого положительного числа d. При $d = 2$ просто констатируем факт, что систему из *пар* целых чисел (r, s) можно сосчитать всего одним целым числом t, как показано на рис. А.3, и это соответствует $(\aleph_0)^2 = \aleph_0$. Данное правило распространяется на любую последовательность из d целых чисел — достаточно всего лишь повторять такой процесс и продемонстрировать, что $(\aleph_0)^d = \aleph_0$. Однако, так или иначе, это не может соответствовать ∞ из вышеприведенных выражений, поскольку мы считаем наше конечное множество $\mathbf{R}$, состоящее из N элементов, моделью *континуума*, содержащего в теории Кантора $2^{\aleph_0} = C$ элементов. (Вполне можно полагать, что C представляет собой предел 2^N, по мере того как N стремится к бесконечности: $N \to \infty$, ведь можно представить себе вещественное число в диапазоне от 0 до 1, выраженное в виде двоичного

разложения, например 0.1101000101110010...). Если оборвать разложение на N-м разряде, то у нас будет 2^N вариантов. Допуская $N \to \infty$, получаем весь континуум возможных вещественных чисел в диапазоне от 0 до 1 с небольшой избыточностью. Но этот факт как таковой для нас бесполезен, поскольку мы просто получим в теории Кантора $C^d = C$ для любого положительного числа d (подробнее о теории Кантора см. в Gardner [2006] и Lévy [1979]).

Теория Кантора о (кардинальных) бесконечностях на самом деле работает лишь с *множествами*, которые не похожи по структуре на некое непрерывное пространство. В данном контексте нам приходится учитывать свойства непрерывности (или гладкости) пространств, с которыми мы работаем. Например, в канторовском смысле количество точек на одномерной линии $\mathbb{R}$ и на двумерной плоскости $\mathbb{R}^2$ (в парах координат x, y, выраженных в вещественных числах) одинаково, как отмечалось в предыдущем абзаце. Однако если речь идет о точках линии $\mathbb{R}$ или плоскости $\mathbb{R}^2$ «из реального мира», представленных в виде *непрерывной* линии или *непрерывной* плоскости соответственно, то последняя должна считаться гораздо более «крупным» объектом в пределе, когда конечное N-элементное множество $\mathbf{R}$ становится непрерывным $\mathbb{R}$ (однако «непрерывным» в узком смысле, когда «близкие» элементы в нашей счетной последовательности всегда образуют «близкие» пары (r, s). Эта закономерность может не работать в *обратном* направлении, то есть она не означает, что близкие *пары* всегда состоят из членов, расположенных близко друг от друга в счетной последовательности).

В нотации Уилера размер нашей непрерывной линии $\mathbb{R}$ обозначается как ∞^1 ($=\infty$), а размер непрерывной плоскости $\mathbb{R}^2$ — как ∞^2 ($> \infty$). Аналогично размер трехмерного пространства $\mathbb{R}^3$ (состоящего из троек вещественных чисел x, y, z) равен

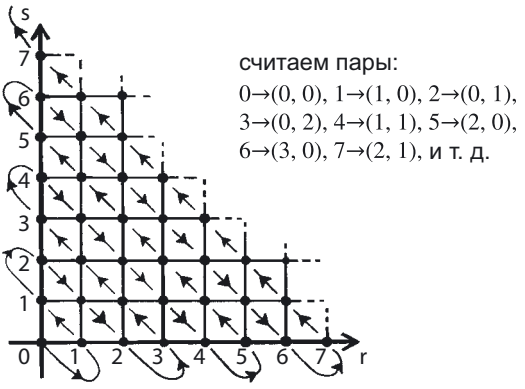


Рис. А.3. Канторовская процедура постановки пар натуральных чисел (r, s) в соответствие отдельным натуральным числам

∞^3 ($>\infty^2$). Пространство гладко изменяющихся магнитных полей в евклидовом трехмерном пространстве ($\mathbb{R}^3$) является бесконечномерным, но все-таки имеет размер, который можно выразить в уилеровской нотации как $\infty^{2\infty^3}$ — см. ранее (при условии, что учитывается связь div **B** = 0; в противном случае, без учета div **B** = 0, получим $\infty^{3\infty^3}$).

Ключевой момент этих выкладок, которым я активно пользовался в главе 1, заключается в том, что хотя (в таком «непрерывном» смысле) мы имеем

$$\infty^{a\infty^d} > \infty^{b\infty^d}, \text{ если } a > b,$$

мы также имеем

$$\infty^{a\infty^c} >> \infty^{b\infty^d}, \text{ если } c > d,$$

и последнее соотношение соблюдается при любом соотношении между положительными числами a и b, причем символ >> означает колоссальное превосходство левой части над правой. Здесь, как и в случае с конечными целыми числами, рассмотренном в разделе А.1, при расчете масштабов гораздо более важна величина *верхней* степени. Данный факт интерпретируется следующим образом: хотя в заданном d-мерном пространстве при увеличении числа компонентов мы получаем более свободно (но непрерывно) изменяющиеся поля, но в пространствах с *разной размерностью* наиболее важна именно эта пространственномерная разница, и любая разница в числе компонентов на точку, которая может существовать между полями, при этом полностью поглощается. В разделе А.8 мы сможем полнее оценить фундаментальный смысл, скрывающийся за этим очевидным фактом.

Выражение «степени свободы» часто используется в физических контекстах, и я в этой книге также часто прибегаю к этой терминологии. Однако следует подчеркнуть, что это понятие не эквивалентно «функциональной свободе». В принципе, если у нас есть физическое поле с n степеней свободы, то, вероятно, имеется в виду функциональная свобода порядка

$$\infty^{n\infty^3},$$

поскольку «количество» степеней свободы связано с количеством параметров *на точку* в трехмерном пространстве. Следовательно, при указанной выше функциональной свободе $\infty^{2\infty^3}$, которой обладают магнитные поля, у нас две степени свободы, что, естественно, больше, чем свобода $\infty^{1\infty^3}$ в однокомпонентном скалярном поле, но у скалярного поля в пятимерном пространстве-времени будет *значительно* большая *функциональная* свобода $\infty^{1\infty^4}$, нежели $\infty^{2\infty^3}$ у нашего магнитного поля в обычном трехмерном пространстве (или четырехмерном пространстве-времени).

А.3. Векторные пространства

Для того чтобы лучше разобраться во всех этих вопросах, важно полнее понимать, как пространства с высшими измерениями трактуются с математической точки зрения. В разделе А.5 мы в общем виде рассмотрим понятие *многообразия* — пространства, которое может иметь любое конечное число измерений, но при этом может быть в определенном смысле *искривлено*. Однако, прежде чем приступить к обсуждению такой геометрии искривленного пространства, по некоторым причинам будет полезно рассмотреть базовую алгебраическую структуру *плоских* пространств высших размерностей. Сам Евклид изучал геометрию двух или трех измерений, но не чувствовал необходимости рассматривать такую геометрию, которая включает большее число измерений, равно как не видел фактов, которые подвигли бы его хотя бы обдумать такие возможности. Тем не менее с появлением систем координат, в первую очередь декартовых (хотя были и другие авторы: например Орем, живший в XIV веке, и даже Аполлоний Пергский, живший в III веке до н. э., по-видимому, высказывали такие идеи), стало очевидно, что алгебраический формализм, применимый к двум или трем измерениям, можно обобщить и для описания высших размерностей, даже если польза от таких многомерных пространств была далеко не очевидна. Тогда как трехмерное евклидово пространство можно изучать при помощи координатных процедур, и точка в пространстве представляется *тройкой* вещественных чисел (x, y, z), можно обобщить эту систему до n-координатного кортежа $(x_1, x_2, x_3 \ldots x_n)$, описывающего точку в некотором n-мерном пространстве. Разумеется, конкретное представление точек в категориях n-координатных кортежей вещественных чисел в таком случае отличается значительной произвольностью; так, например, обозначение точек сильно зависит от выбора координатных *осей*, а также от положения *начала координат*, из которого эти оси выходят, — как мы уже знаем по работе с декартовыми координатами при описании точек в евклидовой плоскости (рис. А.4). Но если мы решаем присвоить особый статус точке O, то геометрию, расположенную *относительно* O, вполне можно представить в виде специальной алгебраической структуры под названием *векторное пространство*.

Векторное пространство состоит из совокупности алгебраических элементов $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$, $\mathbf{x}$, ..., именуемых *векторами*, которые задают отдельные точки в пространстве, а также из чисел, именуемых *скалярами: a*, *b*, *c*, *d*..., при помощи которых можно измерять расстояния (или отрицательные расстояния). Скаляры обычно представляют собой обычные вещественные числа, то есть элементы $\mathbb{R}$, но, как мы могли убедиться, особенно в главе 2, для правильного понимания квантовой механики нужно рассматривать и такие ситуации, где скаляры являются комплексными числами (элементами $\mathbb{C}$, см. раздел А.9). Скаляры, будь они

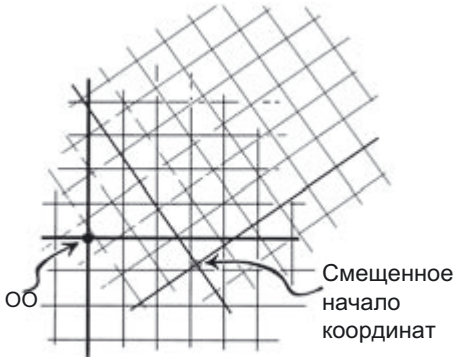


Рис. А.4. Выбор координат в пространстве может быть весьма произвольным даже в случае с обычными прямоугольными декартовыми координатами в евклидовом двумерном пространстве, как, например, в двух системах, показанных на этом рисунке

вещественными либо комплексными, удовлетворяют правилам обычной алгебры, то есть два скаляра можно складывать (+), умножать (×), вычитать (–) и делить (÷) (хотя символ умножения обычно опускают, а символ деления заменяется косой чертой /), причем делить на ноль нельзя. Скаляры подчиняются обычным алгебраическим правилам:

$$a+b=b+a,\ (a+b)+c=a+(b+c),\ a+0=a,$$
$$(a+b)-c=a+(b-c),\ a-a=0,\ a\times b=b\times a,$$
$$(a\times b)\times c=a\times(b\times c),\ a\times 1=a,\ (a\times b)\div c=a\times(b\div c),$$
$$a\div a=1,\ a\times(b+c)=(a\times b)+(a\times c),$$
$$(a+b)\div c=(a\div c)+(b\div c),$$

где a, b, c — произвольно взятые скаляры (хотя деление на ноль не допускается), а 0 и 1 — это *особые* скаляры. Мы пишем $-a$ для $0-a$ и a^{-1} для $1\div a$, а также, как правило, пишем ab в случае $a\times b$ и т. д. (это абстрактные правила, регламентирующие особую систему, которую математики именуют *коммутативным полем*. $\mathbb{R}$ и $\mathbb{C}$ являются примерами таких полей. Не следует путать это понятие с *физическим полем*, описанным в разделе А.2).

Векторы подчиняются двум операциям: *сложению* $(\mathbf{u}+\mathbf{v})$ и *умножению на скаляр* $(a\mathbf{u})$ — по следующим правилам:

$$\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u},\ \ \mathbf{u}+(\mathbf{v}+\mathbf{w})=(\mathbf{u}+\mathbf{v})+\mathbf{w},$$
$$a(\mathbf{u}+\mathbf{v})=a\mathbf{u}+a\mathbf{v},\ \ (a+b)\mathbf{u}=a\mathbf{u}+b\mathbf{u},\ \ a(b\mathbf{u})=(ab)\mathbf{u},$$
$$1\mathbf{u}=\mathbf{u},\ \ 0\mathbf{u}=\mathbf{0},$$

где **0** — *особый* нулевой вектор, и можно записать $-\mathbf{v}$ для $(-1)\mathbf{v}$ и $\mathbf{u} - \mathbf{v}$ для $\mathbf{u} + (-\mathbf{v})$. В случае обычной евклидовой геометрии для двух или трех измерений легко понять геометрическую интерпретацию этих простейших операций над векторами. Нужно зафиксировать точку начала координат O, которую мы обозначим через нулевой вектор **0**; также будем считать, что любой другой вектор **v** обозначает некоторую точку V в пространстве, причем можно сказать, что вектор v соответствует параллельному переносу, то есть *трансляции* всего пространства, и при таком переносе O попадает в V, что можно схематически представить в виде направленного отрезка прямой $\overrightarrow{OV}$, который именно так изображается на рисунках (рис. А.5).

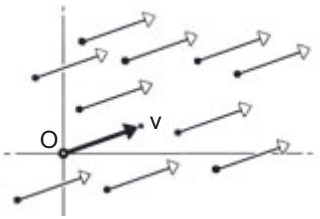


Рис. А.5. (n-мерное) вещественное векторное пространство можно представить в виде совокупности параллельных переносов в евклидовом (n-мерном) пространстве. Сам вектор можно трактовать как направленный отрезок $\overrightarrow{OV}$, где O — выбранное нами начало координат, а V — точка в пространстве; но мы также можем считать v векторным полем, описывающим трансляционные перемещения с переходом O в точку V

В данном случае скаляры являются вещественными числами, и при умножении вектора на положительный вещественный скаляр a направление вектора остается неизменным, но длина вектора увеличивается (или уменьшается) на коэффициент a.

При умножении вектора на отрицательный вещественный скаляр происходит то же самое, но направление результирующего вектора меняется на противоположное. Сумма $\mathbf{w} (= \mathbf{u} + \mathbf{v})$ двух векторов **u** и **v** — это результат сложения двух сдвигов, то есть сдвига u и сдвига v, который схематически соответствует точке W, завершающей параллелограмм OUWV (рис. А.6), причем в вырожденном случае O, U и V расположены на одной прямой, а точка W расположена так, что направленный отрезок OW является суммой отрезков OU и OV. Если бы мы хотели описать ситуацию, в которой три точки U, V и W *коллинеарны* (то есть лежат на одной прямой), то могли бы выразить ее при помощи соответствующих векторов u, v и w следующим образом:

$$a\mathbf{u} + b\mathbf{v} + c\mathbf{w} = 0,$$

где $a + b + c = 0$ для некоторых ненулевых скаляров a, b, c, или, что то же самое, $\mathbf{w} = r\,\mathbf{u} + (1-r)\mathbf{v}$ для некоторого ненулевого скаляра r (причем $r = -a/c$).

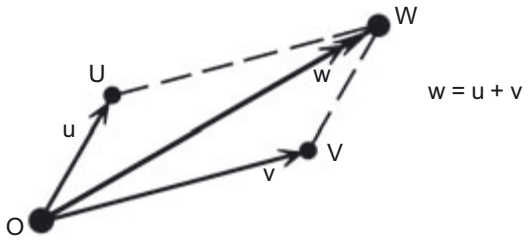


Рис. А.6. Правило параллелограмма, действующее при сложении векторов: если **u** + **v** = **w**, то OUWV — это параллелограмм (возможно, вырожденный)

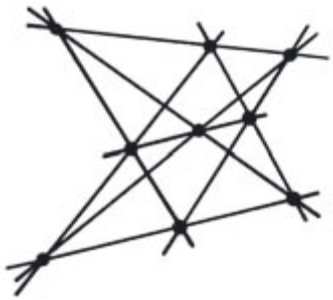


Рис. А.7. Классическую теорему Паппа можно доказать при помощи векторов

Такое алгебраическое описание евклидового пространства очень абстрактно, зато позволяет свести геометрические теоремы Евклида к простым вычислениям, хотя если выполнять эти вычисления «в лоб» (неаккуратно), то они могут серьезно усложниться даже для сравнительно простых геометрических теорем. В качестве примера возьмем теорему Паппа, сформулированную в IV веке (рис. А.7), которая звучит так: «Если две группы коллинеарных точек A, B, C и D, E, F на плоскости соединить пересекающимися прямыми, получив таким образом еще три точки X, Y, Z, так что в X пересекаются отрезки AE и BD, в Y — отрезки AF и CD, а в Z — отрезки BF и CE, то точки X, Y, Z также лежат на одной прямой». Это можно вычислить и «в лоб», однако процесс будет довольно сложен, если не применять специальных вспомогательных процедур.

Именно эта теорема интересна потому, что целиком завязана на понятие коллинеарности. Евклидова геометрия также оперирует понятием *расстояние*, а это понятие удается включить в векторную алгебру при помощи так называемого

внутреннего произведения (оно же скалярное произведение) пар векторов **u**, **v**. В результате получается скаляр, который я записываю (по образцу из квантово-механической литературы) как $\langle \mathbf{u} \,|\, \mathbf{v} \rangle$, хотя встречаются и другие варианты записи, например $(\mathbf{u}, \mathbf{v})$ и $\mathbf{u} \cdot \mathbf{v}$. Далее мы обсудим геометрическую интерпретацию $\langle \mathbf{u} \,|\, \mathbf{v} \rangle$, но сначала рассмотрим алгебраические свойства:

$$\langle \mathbf{u} \,|\, \mathbf{v} + \mathbf{w} \rangle = \langle \mathbf{u} \,|\, \mathbf{v} \rangle + \langle \mathbf{u} \,|\, \mathbf{w} \rangle, \quad \langle \mathbf{u} + \mathbf{v} \,|\, \mathbf{w} \rangle = \langle \mathbf{u} \,|\, \mathbf{v} \rangle + \langle \mathbf{w} \,|\, \mathbf{v} \rangle,$$
$$\langle \mathbf{u} \,|\, a\mathbf{v} \rangle = a \langle \mathbf{u} \,|\, \mathbf{v} \rangle,$$

и во многих типах векторных пространств (например, когда скаляры являются вещественными числами)

$$\langle \mathbf{u} \,|\, \mathbf{v} \rangle = \langle \mathbf{v} \,|\, \mathbf{u} \rangle \text{ и } \langle a\mathbf{u} \,|\, \mathbf{v} \rangle = a \langle \mathbf{u} \,|\, \mathbf{v} \rangle,$$

причем обычно также существует требование

$$\langle \mathbf{u} \,|\, \mathbf{u} \rangle \geq 0,$$

где

$$\langle \mathbf{u} \,|\, \mathbf{u} \rangle = 0, \text{ только если } \mathbf{u} = 0.$$

В случае *комплексных* скаляров (см. раздел А.9) два последних соотношения нередко выглядят иначе, так что мы получаем так называемое *эрмитово* внутреннее произведение, где $\langle \mathbf{u} \,|\, \mathbf{v} \rangle = \overline{\langle \mathbf{v} \,|\, \mathbf{u} \rangle}$, в соответствии с требованиями квантовой механики (в той трактовке, в которой они были описаны в разделе 2.8; см. раздел А.9, где подробнее рассказано о горизонтальной черте). Отсюда следует, что $\langle a\mathbf{u} \,|\, \mathbf{v} \rangle = \overline{a} \langle \mathbf{u} \,|\, \mathbf{v} \rangle$.

Теперь геометрическое понятие *расстояния* можно выразить в категориях внутреннего произведения. Расстояние от начала координат O до точки U, определяемое вектором **u**, — это такой скаляр u, для которого

$$u^2 = \langle \mathbf{u} \,|\, \mathbf{u} \rangle,$$

а поскольку в большинстве типов векторных пространств $\langle \mathbf{u} \,|\, \mathbf{u} \rangle$ является положительным вещественным числом (кроме случаев, когда $\mathbf{u} = \mathbf{0}$), можно определить u как квадратный корень:

$$u = \sqrt{\langle \mathbf{u} \,|\, \mathbf{u} \rangle}.$$

В разделах 2.5 и 2.8 нотация

$$\| \mathbf{u} \| = \langle \mathbf{u} \,|\, \mathbf{u} \rangle$$

обозначает то, что я называю *нормой* **u**, тогда как *длиной* **u** я называю величину $u = \sqrt{\langle \mathbf{u} \mid \mathbf{u} \rangle}$ (хотя некоторые авторы предпочитают называть *нормой* **u** величину $\sqrt{\langle \mathbf{u} \mid \mathbf{u} \rangle}$), а буква, выделенная курсивом, означает длину вектора (тогда как та же самая буква, записанная жирным шрифтом, — это сам вектор). Например, v — длина вектора **v** и т. д. Таким образом, в обычной евклидовой геометрии $\langle \mathbf{u} \mid \mathbf{v} \rangle$ интерпретируется как

$$\langle \mathbf{u} \mid \mathbf{v} \rangle = uv \cos\theta,$$

где θ — это угол[1] между отрезками OU и OV (здесь следует отметить, что θ = 0 и cos 0 = 1, когда U = V). Расстояние между двумя точками U и V будет равно длине u − v, то есть квадратному корню из

$$\| \mathbf{u} - \mathbf{v} \| = \langle \mathbf{u} - \mathbf{v} \mid \mathbf{u} - \mathbf{v} \rangle.$$

Векторы **u** и **v** называются *ортогональными*, то есть $\mathbf{u} \perp \mathbf{v}$, если их скалярное произведение равно нулю:

$$\mathbf{u} \perp \mathbf{v} \text{ означает } \langle \mathbf{u} \mid \mathbf{v} \rangle = 0.$$

Из вышесказанного понятно, что это соответствует случаю cos θ = 0, и, таким образом, угол θ прямой, а отрезки OU и OV *перпендикулярны*.

А.4. Векторные базисы, координаты и дуальность

(Конечным) базисом векторного пространства является семейство векторов ε_1, ε_2, ε_3... ε_n, обладающее следующим свойством: каждый вектор v в этом пространстве можно выразить как *линейную комбинацию*

$$\mathbf{v} = v_1\varepsilon_1 + v_2\varepsilon_2 + v_3\varepsilon_3 + \ldots + v_n\varepsilon_n$$

членов этого семейства, иными словами, комбинация ε_1, ε_2, ε_3, ..., ε_n *охватывает* все векторное пространство. Кроме того, базис требует, чтобы все векторы были *линейно независимыми*. Таким образом, чтобы охватить пространство, нам нужны *все* ε. Последнее условие эквивалентно тому, что нулевой вектор **0** (=**v**) получается лишь в случае, когда все коэффициенты в выражении v_1, v_2, v_3, ..., v_n равны нулю — или, аналогично, что вышеприведенное представление каждого **v**

---

[1] В тригонометрии cos θ, именуемый *косинусом* угла θ, определяется в евклидовом прямоугольном треугольнике ABC, где угол θ находится у вершины A, а прямой угол — у вершины B, как *отношение* AB/AC. Величина sin θ = BC/AC называется *синусом* θ, а величина tg θ = BC/AB — *тангенсом* θ. Функции, обратные трем этим, я обозначу соответственно как $\cos^{-1}$, $\sin^{-1}$ и tg^{-1} (так что $\cos(\cos^{-1} X) = X$ и т. д.).

уникально. Для каждого конкретного вектора **v** коэффициенты v_1, v_2, v_3, ..., v_n из вышеприведенного выражения являются *координатами* **v** относительно этого базиса, часто они именуются *компонентами* v в этом базисе (причем, строго говоря, «компонентами» v на самом деле должны считаться величины $v_1\varepsilon_1$, $v_2\varepsilon_2$ и т. д., но обычно под *компонентами* понимаются просто скаляры v_1, v_2, v_3 и т. д.). Число элементов в семействе базисных векторов называется *размерностью* векторного пространства, причем это число не зависит от того, какой именно базис был выбран для конкретного векторного пространства. В случае двумерного евклидова пространства любые два ненулевых и непропорциональных вектора могут выступать в качестве базиса (то есть речь может идти о любых векторах **u** и **v**, указывающих точки U и V, при условии, что эти точки не лежат на одной прямой с O). В трехмерном евклидовом пространстве в таком качестве подойдут любые три линейно независимых вектора **u**, **v** и **w** (соответствующие им точки U, V и W не должны лежать в одной плоскости с O). Направления базисных векторов в точке O в каждом из случаев дают возможные комбинации координатных осей, поэтому точка P в контексте базиса (**u**, **v**, **w**) была бы представлена как

$$\mathbf{p} = x\mathbf{u} + y\mathbf{v} + z\mathbf{w},$$

где *координаты* P — это (x, y, z). Следовательно, с алгебраической точки зрения совсем не сложно сделать обобщение от двух или трех до n измерений, где n — любое положительное целое число.

В общем случае базис не требует, чтобы координатные оси были перпендикулярны друг другу, но в стандартных *декартовых* координатах (пусть они и названы в честь Декарта, сам он не настаивал, что координатные оси должны быть перпендикулярны) они должны пересекаться под прямым углом:

$$\mathbf{u} \perp \mathbf{v},\ \mathbf{u} \perp \mathbf{w},\ \mathbf{v} \perp \mathbf{w}.$$

Более того, в геометрическом контексте обычна ситуация, когда единицы расстояния одинаковы по всем трем осям. Так возникает условие *нормировки*, согласно которому координатные базисные векторы **u**, **v** и **w** должны быть *единичными* (то есть длина такого вектора должна быть равна единице):

$$\|\mathbf{u}\| = \|\mathbf{v}\| = \|\mathbf{w}\| = 1.$$

Такой базис называется *ортонормированным*.

В n измерениях семейство из n ненулевых векторов ε_1, ε_2, ε_3, ..., ε_n образует ортогональный базис, если все эти векторы взаимно перпендикулярны:

$$\varepsilon_j \perp \varepsilon_k \text{ всякий раз, когда } j \neq k \text{ (при } j, k = 1, 2, 3, ..., n),$$

и этот базис к тому же считается ортонормированным, если все эти векторы являются единичными:

$$\|\varepsilon_i\| = 1 \text{ для всех } i = 1, 2, 3, ..., n.$$

Два этих условия зачастую объединяются в виде

$$\langle \varepsilon_i \,|\, \varepsilon_j \rangle = \delta_{ij},$$

где используется символ *Кронекера*, определяемый по формуле

$$\delta_{ij} = \begin{cases} 1 & \text{если } i = j, \\ 0 & \text{если } i \neq j. \end{cases}$$

На основании данной формулы легко показать (если скаляры являются вещественными числами), что внутреннее произведение **u** и **v** и расстояние |UV| между U и V в декартовых координатах описываются, соответственно, следующими выражениями:

$$\langle \mathbf{u} \,|\, \mathbf{v} \rangle = u_1 v_1 + u_2 v_2 + ... + u_n v_n$$

и

$$|\mathrm{UV}| = |\mathbf{u} - \mathbf{v}| = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + ... + (u_n - v_n)^2}.$$

Завершая этот раздел, рассмотрим последнюю концепцию, применимую непосредственно к любому (конечномерному) векторному пространству **V**, а именно *дуальное* ему векторное пространство **V*** с той же размерностью, что и у **V**. Дуальное пространство тесно связано с **V** и часто отождествляется с ним, но на самом деле это векторное пространство следует считать отдельным. Элемент **p**, относящийся к **V***, называется *линейным отображением* (или линейной *функцией*) **V** на систему скаляров, и это означает, что **p** — функция элементов **V**, то есть скаляр **p**(**v**), где **v** — любой вектор, относящийся к **V**. Эта функция *линейна* в том смысле, что

$$\mathbf{p}(\mathbf{u} + \mathbf{v}) = \mathbf{p}(\mathbf{u}) + \mathbf{p}(\mathbf{v}) \text{ и } \mathbf{p}(a\mathbf{u}) = a\mathbf{p}(\mathbf{u}).$$

Пространство всех таких **p** является векторным; мы называем его **V*** и определяем для него операции сложения **p** + **q** и умножения на скаляр *a***p** по формуле

$$(\mathbf{p}+\mathbf{q})(\mathbf{u}) = \mathbf{p}(\mathbf{u}) + \mathbf{q}(\mathbf{u}) \text{ и } (a\mathbf{p})(\mathbf{u}) = a\mathbf{p}(\mathbf{u})$$

для всех **u**, относящихся к **V**. Можете проверить и убедиться, что эти правила определяют **V*** как векторное пространство с такой же размерностью, что и у **V**,

а оно, будучи ассоциировано с любым базисом (ε_1, ..., ε_n) для **V**, образует дуальный базис (ϱ_1, ..., ϱ_n) для **V***, где

$$\varrho_i\left(\varepsilon_j\right) = \delta_{ij}\,.$$

Если повторить такую операцию «дуализации» для получения векторного n-пространства **V****, то мы вновь получим **V**, то есть **V**** естественным образом отождествляется с исходным пространством **V**, так что его можно описать формулой

$$\mathbf{V}^{**} = \mathbf{V},$$

действие элемента **u** из **V**, выступающего в качестве элемента **V****, определяется соотношением $\mathbf{u}(\mathbf{p}) = \mathbf{p}(\mathbf{u})$.

Как интерпретировать элементы дуального пространства **V*** в геометрическом или физическом смысле? Давайте вновь вернемся к рассуждениям в терминах евклидового трехмерного пространства ($n = 3$). Как вы помните, элемент **u** из векторного пространства **V**, отложенный относительно начала координат O, может указывать на некоторую другую точку U в нашем евклидовом пространстве (или представлять параллельный перенос, при котором U сдвигается на место O). Фактически элемент **p** из пространства **V***, иногда именуемый *ковектором*, будет ассоциирован с *плоскостью* P, проходящей через начало координат O и содержащей все точки U, для которых $\mathbf{p}(\mathbf{u}) = 0$ (рис. А.8). Плоскость P полностью характеризует ковектор с точностью до коэффициента пропорциональности, то есть не позволяет отличить **p** от $a\mathbf{p}$, где a — любой ненулевой скаляр. Однако в физическом смысле масштаб **p** можно понимать как некоторую *силу*, ассоциированную с плоскостью P. Можно считать, что эта сила сообщает плоскости

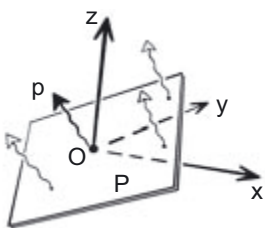


Рис. А.8. В n-мерном векторном пространстве **V** любой ненулевой элемент **p** из дуального ему пространства **V*** (так называемый *ковектор*) может интерпретироваться в **V** как гиперплоскость, проходящая через начало координат O, с сопоставленной ей некоторой «силой» (с квантово-механической точки зрения это просто частота). Здесь показан случай $n = 3$, где ковектор **p** изображен как двумерная плоскость P, повернутая относительно координатных осей **x**, **y**, **z**, исходящих из начала координат O

некоторый *импульс*, направленный прочь от плоскости P. В разделе 2.2 мы обнаружили, что в квантовой механике этот импульс позволяет приписать плоскости «частоту колебаний», которую можно отождествить с волновым числом плоской волны, распространяющейся прочь от плоскости P.

На этой картинке не используется метрическая «длина», присущая *евклидовому* трехмерному пространству. Но при помощи внутреннего произведения $\langle \cdots | \cdots \rangle$ можно «отождествить» векторное пространство $\mathbf{V}$ и дуальное ему пространство $\mathbf{V}^*$, и в этом случае ковектор $\mathbf{v}^*$, ассоциированный с вектором $\mathbf{v}$, будет «оператором» $\langle \mathbf{v} | \ \ \rangle$, в результате действия которого на произвольный вектор $\mathbf{u}$ получается скаляр $\langle \mathbf{v} | \mathbf{u} \rangle$. В терминах геометрии евклидова трехмерного пространства плоскость, ассоциированная с дуальным вектором $\mathbf{v}^*$, будет проходить через начало координат O перпендикулярно OV.

Данное описание также применимо к векторным пространствам с *любой* (конечной) размерностью n, где вместо описания ковектора в трехмерном пространстве двумерной плоскостью мы получим $(n - 1)$-плоскостное описание ковектора в n-мерном пространстве, и такая $(n - 1)$-плоскость будет проходить через начало координат O. Подобная высокоразмерная плоскость, имеющая всего на одно измерение меньше, чем окружающее пространство, часто именуется *гиперплоскостью*. Опять же, для полного описания ковектора (с точностью до коэффициента пропорциональности) гиперплоскости необходимо сопоставить определенную «силу», которую можно понимать как импульс или «частоту» (волновое число), направленную прочь от гиперплоскости.

Вышеприведенные рассуждения излагались в терминах *конечномерных* пространств. Однако можно рассматривать и векторные пространства с *бесконечным* числом измерений. Такие пространства, в базисе которых присутствует бесконечное число элементов, встречаются в квантовой механике. Большинство из вышеизложенного соблюдается и в таких пространствах, и основные различия возникают, если попытаться рассмотреть *дуальное* векторное пространство. Ведь для дуального векторного пространства обычны ограничения, связанные с линейными отображениями, из которых, как считается, состоит дуальное пространство $\mathbf{V}^*$, и такие ограничения нужны, чтобы отношение $\mathbf{V}^{**} = \mathbf{V}$ продолжало соблюдаться.

A.5. Математика многообразий

Рассмотрим более общее понятие *многообразия*. Многообразие не обязано быть плоским, как евклидово пространство, оно может быть причудливым образом искривлено и даже иметь отличную от евклидова пространства топологию.

Многообразия играют фундаментальную роль в современной физике. Отчасти дело в том, что общая теория относительности Эйнштейна описывает гравитацию в терминах искривленного пространственно-временного многообразия. Но, что еще важнее, многие другие физические концепции наиболее удобно описывать в категориях многообразий — таковы, например, конфигурационные и фазовые пространства, которые мы рассмотрим в разделе А.6. Они обладают огромным числом измерений и иногда имеют сложную топологию.

Итак, что такое многообразие? В принципе, это просто гладкое конечномерное пространство, и его можно назвать *n-многообразием*. Как же понимать прилагательное «гладкое» в данном контексте? Чтобы соблюсти математическую точность, давайте обсудим эту проблему в терминах многомерного *анализа*. В этой книге я старался не вдаваться в формализм математического анализа (не считая кратких ремарок, сделанных в конце раздела А.11), но без некоторого интуитивного представления о связанных с ним базовых концепциях обойтись не удастся.

Итак, что же такое «гладкое *n*-мерное пространство»? Рассмотрим любую точку P в пространстве. Если в точке P пространство должно быть *гладким*, то можно представить себе, как оно должно выглядеть в окрестностях P под все более и более сильным увеличением. Если пространство является гладким в окрестности P, то в *пределе* бесконечного увеличения оно будет выглядеть как *плоское n*-мерное пространство; см. рис. А.9: изображенная на рисунке вершина конуса P, например, гладкой *не будет*. В случае глобально гладкого многообразия

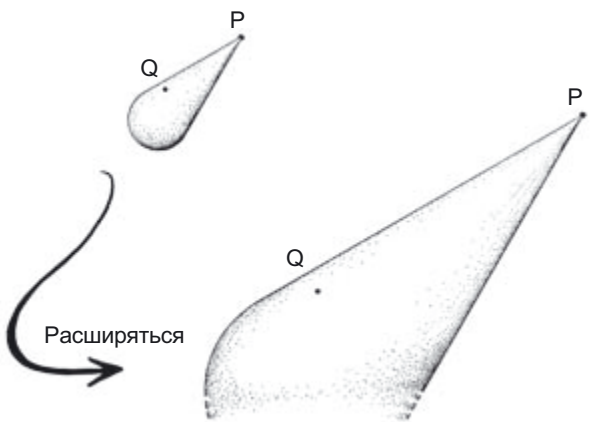


Рис. А.9. Многообразие, изображенное сверху, не будет гладким в точке P, так как независимо от того, насколько его увеличить, мы не получим в пределе плоское пространство. Однако в точке Q оно гладкое, так как при увеличении кривизна пространства уменьшается, и в данном случае пространство в пределе получается плоским

такое предельное «растянувшееся» пространство хотя и будет плоским, не может считаться евклидовым n-пространством, поскольку оно не обязательно должно иметь *метрическую* структуру (то есть в нем может не быть *расстояний*), какая есть у евклидова пространства. Однако в пределе оно должно обладать структурой *векторного* пространства, описанной в разделах А.3 и А.4, где начало координат и будет соответствовать точке P, на чем мы сейчас заостряем внимание. (Представьте себе, что выбрали точку на карте Google, а затем бесконечно расширяете эту карту.)

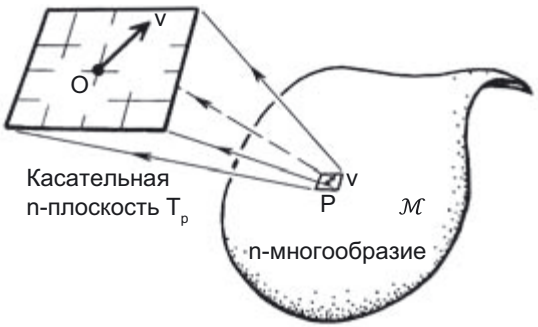


Рис. А.10. Касательный вектор **v** в точке P (гладкого) многообразия M будет элементом касательного пространства T_P, которое соприкасается с многообразием в точке P. Можно сказать, что векторное пространство T_P, вплотную смежное с P, бесконечно расширяется. Точка O — это начало координат в плоскости T_P

Такое ограничивающее векторное пространство называется *касательным* к точке P и часто обозначается как T_P. Различные элементы T_P как таковые называются *касательными векторами* в точке P (рис. А.10). Чтобы составить хорошее интуитивное представление о геометрической сущности касательного вектора, вообразим стрелку, основание которой расположено в точке P, но сама стрелка направлена от P вдоль многообразия. Различные *направления* вдоль многообразия, откладываемые из точки P, будут описываться разными ненулевыми векторами в T_P (с точностью до умножения на скаляр). Чтобы наше пространство можно было считать глобально гладким *n-многообразием*, к каждой его точке должно прилегать четко определенное касательное n-пространство.

Иногда многообразие может обладать более выраженной структурой, а не просто *гладкостью*, подразумевающей, что многообразие ограничивается касательными пространствами. Например, в *римановом многообразии* присутствует локальная мера *длины*, получаемая путем приравнивания касательных пространств к *евклидовым* векторным пространствам, — для этого каждому T_P присваивается

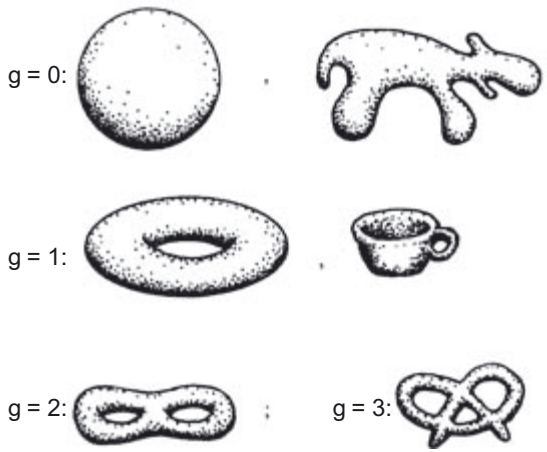


Рис. А.11. Примеры двумерных пространств с различными топологиями. Величина g соответствует роду поверхности (числу «ручек») (ср. с рис. 1.44)

внутреннее произведение $\langle \cdots | \cdots \rangle$, описанное в разделе А.3. Существуют и другие типы локальных структур, которые могут быть важны в физике, например *симплектические* структуры, применимые к *фазовым пространствам*, — о них мы поговорим далее. В случае с обычными фазовыми пространствами оказывается, что у них есть своеобразное внутреннее произведение $[\cdots | \cdots]$ для касательных векторов в некоей точке, для которых соблюдается антисимметричное соотношение $[\mathbf{u} | \mathbf{v}] = -[\mathbf{v} | \mathbf{u}]$ в противовес симметричному $\langle \mathbf{u} | \mathbf{v} \rangle = \langle \mathbf{v} | \mathbf{u} \rangle$, которое соблюдается для риманова многообразия.

В глобальном масштабе топология многообразия может быть простой, как у n-мерного евклидова пространства, либо значительно более затейливой, как у двумерных образцов, показанных на рис. А.11 и на рис. 1.44 в разделе 1.16. Но в любом случае, какова бы ни была общая топология, n-многообразие в описанном выше смысле везде можно *локально* уподобить плоскому n-мерному векторному пространству, и в нем не обязательно должны существовать локальные понятия угла или расстояния, как в евклидовом n-пространстве $\mathbb{E}^n$. Как вы помните, можно присваивать *координаты* для обозначения различных точек векторного пространства, как описано в разделе А.4. Давайте обсудим общий случай присваивания координат n-многообразию. В случае евклидова n-пространства допустимо считать, что это пространство как единое целое можно смоделировать в виде пространства $\mathbb{R}^n$, образованного n-координатными кортежами вещественных чисел $(x_1, x_2, ..., x_n)$, как декартовы координаты на рис. А.4, но любое такое представление далеко не уникально. В целом многообразие можно описать в категориях координат, но в таком случае возникает еще более сильная

произвольность, нежели при описании евклидового пространства в координатах векторного. Также еще одна проблема заключается в том, могут ли такие координаты применяться *глобально* ко всему многообразию либо только в некоторых локальных областях. Придется рассмотреть все эти вопросы.

Давайте вернемся к вышеупомянутому варианту присвоения координат (x_1, x_2, ..., x_n) в евклидовом пространстве $\mathbb{E}^n$. Если решить эту задачу только что описанным способом — присваивая координаты векторному пространству, то необходимо отметить, что в $\mathbb{R}^n$ нет конкретной точки O, выделенной в качестве «начала координат», которая в векторном пространстве получила бы координаты (0, 0, ..., 0). Она совершенно произвольна и лишь увеличивает ту произвольность, что сложилась при выборе *базиса* для векторного пространства. На уровне координат произвольность при выборе точки отсчета можно выразить как свободу «трансляции»[1] заданной координатной системы — например, системы $\mathfrak{C}$ в другую систему $\mathfrak{A}$. При этом к каждому компоненту x_i в описании $\mathfrak{C}$ прибавляется фиксированная величина A_i (обычно для каждого значения i эта величина отличается), поэтому если представить точку P в системе координат $\mathfrak{C}$ как n-координатный кортеж ($x_1, x_2, ..., x_n$), то P в системе $\mathfrak{A}$ будет представлена n-координатным кортежем ($X_1, X_2, ..., X_n$), где при

$$X_i = x_i + A_i \quad (i = 1, 2, ..., n)$$

начало координат O в $\mathfrak{C}$ будет представлено в системе $\mathfrak{A}$ n-координатным кортежем ($A_1, A_2, ..., A_n$).

Это очень простое координатное преобразование, в результате которого мы получаем просто еще одну систему координат такого же «линейного» типа, что и прежняя. Изменив базис векторного пространства, также получим другую систему координат все того же конкретного типа. Часто при изучении структур евклидовой геометрии используются более универсальные системы координат, именуемые *криволинейными*. Одна из наиболее известных подобных систем именуется полярными координатами евклидовой плоскости (рис. А.12 *а*), где стандартные декартовы координаты (x, y) меняются на (r, θ), причем

$$y = r\sin\theta, \quad x = r\cos\theta.$$

Верно и обратное:

$$r = \sqrt{x^2 + y^2}, \quad \theta = \tan^{-1}\frac{x}{y}.$$

[1] Термин «трансляция» в данном случае несет двойную смысловую нагрузку: кроме привычного значения «перенос из одной системы описания в другую», здесь есть и математический: «сдвиг без поворота».

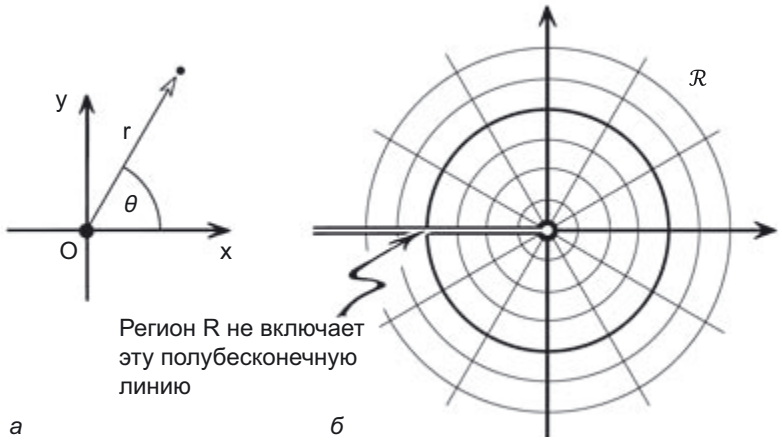


Рис. А.12. «Криволинейная» система полярных координат (r, θ): *а*) отношение к стандартной системе декартовых координат (x, y); *б*) чтобы построить корректную координатную карту R, нужно исключить из нее определенную центральную линию, здесь — луч $\theta = \pm\pi$

Как понятно из названия, координатные оси в криволинейной системе координат не обязаны быть прямыми линиями (или плоскостями и т. д., если рассматривать ситуации с высшими измерениями). Поэтому из рис. А.12 видно, что если линии, заданные уравнением $\theta = \text{const}$, прямые, то линии, заданные уравнением $r = \text{const}$, будут искривленными, то есть круговыми. Пример с полярными координатами иллюстрирует и еще одно распространенное свойство криволинейных координат, а именно: зачастую они не накрывают все пространство ровной сеткой от края до края. Центральная точка (0, 0), присутствующая в системе (x, y), в системе (r, θ) корректно не представлена (θ в данной точке не имеет единственного значения), и даже если мы обойдем эту точку по окружности, то обнаружим, что θ скачком меняет значение на 2π (то есть на 360°). Однако полярные координаты позволяют верно обозначить точки области $\mathcal{R}$ на плоскости, которая не включает центральную точку O (при $r = 0$) и луч, исходящий из O в направлении, противоположном $\theta = 0$, который бы неоднозначно описывался $\theta = \pm\pi$, то есть $\theta = \pm 180°$ (рис. А.12 *б*). (Следует отметить, что в данном случае я трактую полярную координату θ как протяженную от −180° до +180°, хотя часто используется и диапазон от 0 до 360°.)

Данная область $\mathcal{R}$ — пример так называемого открытого подмножества евклидовой плоскости $\mathbb{E}^2$. Чтобы дать интуитивно понятное определение «открытости» применительно к $\mathcal{R}$, представим, что $\mathcal{R}$ — это подмножество n-многообразия $\mathcal{M}$, область $\mathcal{M}$, имеющая полную размерность n многообразия $\mathcal{M}$ и не включающая никаких

границ или «краев», которые могут быть у $\mathcal{R}$. (В случае полярных координат на плоскости такой исключенной областью окажется часть оси абсцисс, на которой расположены неположительные значения x.) Другим примером открытого подмножества $\mathbb{E}^2$ будет *диск* (или 2-шар) — полностью лежащий в пределах единичной окружности (то есть описываемый неравенством $x^2 + y^2 < 1$). С другой стороны, ни сама единичная окружность ($x^2 + y^2 = 1$), ни область, состоящая из этого диска *вместе* с границей единичной окружности (то есть *замкнутый* единичный диск $x^2 + y^2 \leq 1$) не будут открытыми. Соответствующие утверждения верны и для более высоких измерений, так что в $\mathbb{E}^3$ «замкнутая» область $x^2 + y^2 + z^2 \leq 1$, *не будет* открытой, а 3-шар $x^2 + y^2 + z^2 < 1$ будет, и т. д. Чуть точнее ситуацию можно сформулировать так: открытая область $\mathcal{R}$ n-многообразия $\mathcal{M}$ характеризуется следующим свойством: любая точка p области $\mathcal{R}$ расположена в центре достаточно небольшого координатного n-шара, полностью лежащего в пределах $\mathcal{M}$. Эта ситуация проиллюстрирована на рис. А.13 для двумерного случая (открытый диск), любая точка которого, независимо от того, насколько она удалена от границы, располагается в небольшом круге, полностью располагающемся в пределах диска.

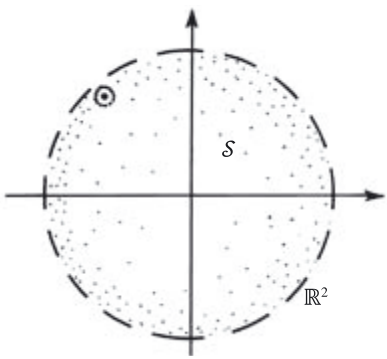


Рис. А.13. Участок $\mathcal{S}$ многообразия называется открытым множеством, если представляет собой подмножество, каждая точка которого содержится в координатном шаре, полностью расположенном внутри $\mathcal{S}$. Здесь эта ситуация проиллюстрирована на примере двух измерений, причем $\mathcal{S}$ — подмножество $\mathbb{R}^2$, задаваемое $x^2 + y^2 < 1$; при этом, как можно убедиться, любая точка в пределах $\mathcal{S}$ располагается в небольшом диске, полностью укладывающемся в $\mathcal{S}$. Область $x^2 + y^2 \leq 1$ не удовлетворяет этому условию, поскольку выражение не соблюдается для точек, расположенных на границе (которая теперь является частью множества)

Вообще, по топологическим причинам обнаруживается, что невозможно глобально присвоить координаты целому многообразию $\mathcal{M}$ в единственной системе координат $\mathfrak{C}$, причем выясняется, что любая попытка такой координатизации где-то отказывает (например, в точках северного и южного полюса, на линии

перемены дат — если говорить о координатах широты и долготы на шарообразной Земле). В таких случаях, если нам понадобится присвоить координаты $\mathcal{M}$, одной координатной системы для этого будет мало; напротив, нам потребуется покрыть все $\mathcal{M}$ лоскутным одеялом взаимно пересекающихся областей $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$, ... (см. рис. А.14), которое называется *открытым покрытием* $\mathcal{M}$, где мы присвоим координатную систему $\mathfrak{C}_i$ каждому конкретному $\mathcal{R}_i$ (i = 1, 2, 3, ...). При каждом наложении различных частей открытых множеств этого покрытия, то есть на каждом непустом пересечении

$$\mathcal{R}_i \cap \mathcal{R}_j$$

(символ «$\cap$» означает *пересечение*), у нас получатся две разные системы координат, а именно $\mathfrak{C}_i$ и $\mathfrak{C}_j$, и нам потребуется указать преобразование (например, преобразование между рассмотренными выше системами координат: декартовыми (x, y) и полярными (r, θ); см. рис. А.12). Собирая координатные системы таким образом, можно создавать пространства со сложной геометрией или топологией, как показано для двух измерений на рис. А.11 и рис. 1.44 *а* из раздела 1.16.

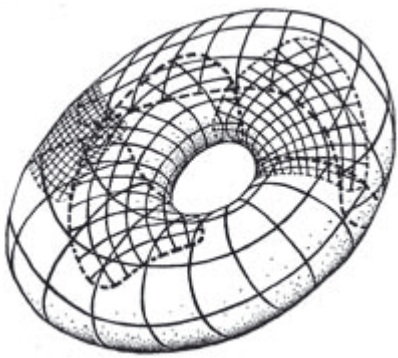


Рис. А.14. Здесь показано открытое покрытие пространства — в данном случае двумерного тора — открытыми координатными областями $\mathbb{R}^2$ (в тексте обозначены как $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$, ...)

Необходимо помнить, что координаты могут считаться просто *вспомогательными* единицами, введенными для удобства, чтобы с их помощью было проще исследовать свойства многообразия. Как правило, сами по себе координаты не несут никакого конкретного смысла; в частности, в этих координатах не имеет смысла выражать евклидовы расстояния. (Как мы помним из раздела А.4, формула для вычисления евклидовых расстояний между точкой (X, Y, Z) и точкой (x, y, z) в $\mathbb{E}^3$ в декартовых координатах имеет вид $\sqrt{(X-x)^2+(Y-y)^2+(Z-z)^2}$.)

Нас скорее заинтересовали бы такие свойства многообразия, которые *не зависят* от каких-либо координатных систем (например, в полярных координатах на плоскости расстояния вычисляются по совершенно другой формуле). Эта проблема особенно важна в эйнштейновской теории относительности, где пространство-время считается четырехмерным многообразием и никакие пространственные или временные координаты не обладают абсолютным статусом. В общей теории относительности это именуется *принципом общей ковариантности* (см. разделы 1.2, 1.7 и 2.13).

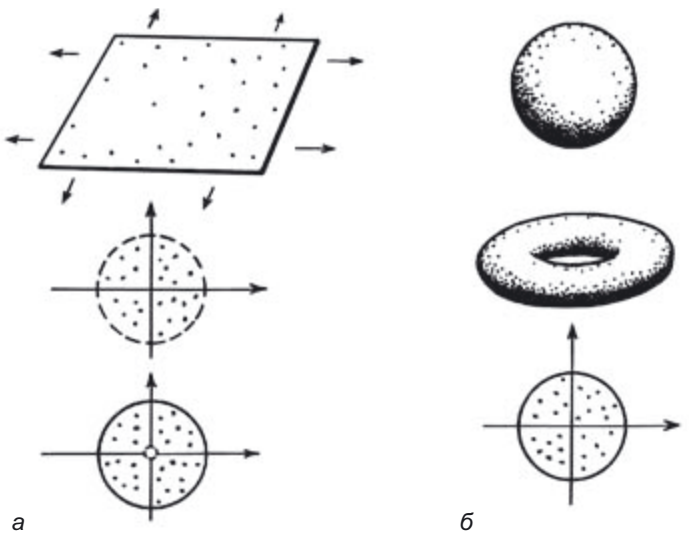


Рис. А.15. Различные примеры некомпактных двумерных многообразий: бесконечная евклидова плоскость, открытый единичный круг, замкнутый единичный круг, в котором удалено начало координат (*а*); различные примеры компактных двумерных многообразий: сфера S_2, тор $S_1 \times S_1$, замкнутый единичный круг (*б*)

Многообразие может быть, что называется, *компактным* — это, в принципе, означает, что оно смыкается само с собой как замкнутая кривая (размерность $n = 1$) или как замкнутые поверхности, изображенные на рис. А.15 *а*, или как замкнутая топологическая поверхность, показанная на рис. 1.44 *а* в разделе 1.16 (размерность $n = 2$). Либо многообразие может быть *некомпактным*, подобно евклидову n-пространству или поверхности с отверстиями, показанной на рис. 1.44 *б*. Разница между компактными и некомпактными поверхностями проиллюстрирована на рис. А.15, где некомпактные пространства можно считать «продолжающимися до бесконечности» или «имеющими разрывы», как, например, «отверстия» на рис. 1.44 *б* (где *граничные кривые* трех отверстий

не считаются частью многообразия). Если выразиться немного точнее, компактное многообразие обладает следующим свойством: в любой бесконечной последовательности точек такого многообразия есть *предельная точка*, то есть точка P, такая, что любое открытое множество, содержащее P, включает бесконечное множество элементов этой последовательности (рис. А.16). Подробнее о технических деталях и о тех проблемах, которые я здесь не затрагиваю, см. [Tu, 2010 и Lee, 2003].

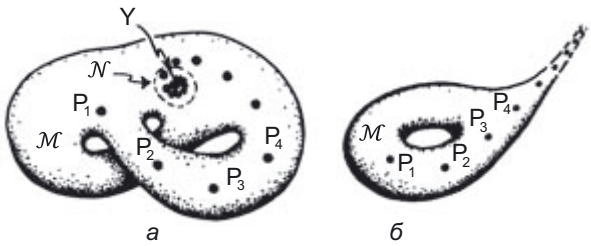


Рис. А.16. Характеристика компактности многообразия $\mathcal{M}$: *а*) в компактном $\mathcal{M}$ любая бесконечная последовательность точек P_1, P_2, P_3, ... имеет точку накопления y в $\mathcal{M}$; *б*) в некомпактном $\mathcal{M}$ некоторые бесконечные последовательности точек P_1, P_2, P_3, ... не имеют точки накопления y в $\mathcal{M}$ (точка накопления y обладает таким свойством: любое открытое множество $\mathcal{N}$, содержащее Y, также содержит бесконечное множество P_i)

Иногда мы рассматриваем такие области внутри многообразия, которые имеют *края*, но подобные области, строго говоря, не являются многообразиями в описанном здесь смысле, а относятся к более общей категории пространств под названием *многообразия с краем* (такова, например, поверхность, изображенная на рис. 1.44 *б* в разделе 1.16, но теперь границы отверстий включаются в состав такого многообразия с краем). Такие пространства могут быть компактными, не «замыкаясь при этом на себя» (см. рис. А.15 *б*). Многообразие может быть *связным*, что (выражаясь упрощенно) означает: оно состоит из цельного куска. В противном случае оно является *несвязным*. 0-многообразие состоит из одной точки, если оно связное, и из двух или более точек, если несвязное. Часто компактное многообразие (не имеющее границ) именуется «замкнутым»[1].

[1] Это еще один пример путаной математической терминологии, поскольку этот термин вступает в конфликт с рассмотренным выше топологическим понятием *замкнутого множества*. *Любое* многообразие образует замкнутое множество в топологическом смысле (то есть «замкнутое» как антоним «*открытого*», описанного выше; в замкнутом множестве содержатся все его предельные точки [Tu, 2010; Lee, 2003]), независимо от того, является ли конкретное многообразие замкнутым.

А.6. Многообразия в физике

Наиболее очевидный образец многообразия в физике — это трехмерное многообразие, соответствующее обычному евклидову трехмерному пространству. Однако, согласно общей теории относительности Эйнштейна (см. раздел 1.7), необходимо учитывать, как многообразия могут быть *искривлены*. Например, если рассматривать магнитные поля (обсуждались в разделе А.2) в искривленном трехмерном пространстве, то это будут *векторные поля*, как те, что показаны на рис. А.17. Более того, *пространства-времена* из общей теории относительности являются искривленными четырехмерными многообразиями, и зачастую приходится рассматривать в пространстве-времени такие поля, которые гораздо сложнее обычных векторных по своей природе.

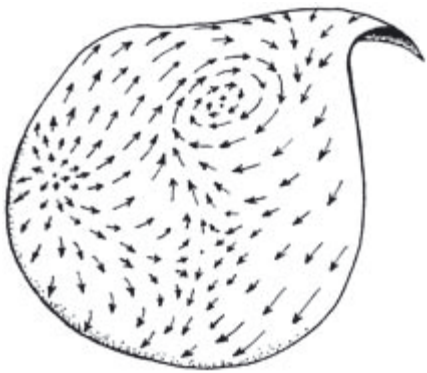


Рис. А.17. Гладкое векторное поле на многообразии. Три точки без стрелок — это три места, где векторное поле обращается в ноль

Однако в обычной физике (не связанной с теорией струн) нас зачастую интересуют многообразия, в которых больше трех или четырех измерений (при этом трехмерное многообразие используется для описания обычного пространства, а четырехмерное многообразие — для описания пространства-времени). Можно поинтересоваться: зачем, кроме как ради чисто математической забавы, заниматься столь многомерными многообразиями, топология которых может отличаться от евклидовой? Следует четко отметить, что многообразия, размерность которых значительно превышает четыре и которые могут обладать сложными топологиями, действительно играют разнообразные ключевые роли в современной теоретической физике — играют вполне безотносительно требований многих современных физических теорий (например, теории струн, рассмотренной в главе 1), предполагающих, что должно быть больше трех пространственных

измерений. Среди простейших и важнейших примеров многомерных многообразий — *конфигурационные* и *фазовые* пространства. Давайте кратко их обсудим.

Конфигурационное пространство — это математическое пространство (многообразие $\mathcal{C}$), каждая точка которого полностью описывает местоположения всех отдельно взятых элементов некоторой рассматриваемой нами физической системы (рис. А.18). Простым примером будет шестимерное конфигурационное пространство, каждая из точек которого представляет местоположение (в том числе пространственную ориентацию) некоторого жесткого тела в обычном евклидовом трехмерном пространстве (рис. А.19). Чтобы зафиксировать, скажем, центр тяжести (центр масс) G тела B, нам потребуются три координаты, а также еще три координаты, чтобы зафиксировать ориентацию B в пространстве — всего получается шесть.

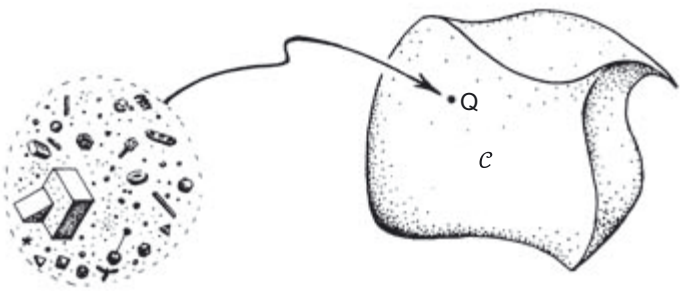


Рис. А.18. Точка Q в конфигурационном пространстве C соответствует местоположению (и ориентации, если речь идет об асимметричной фигуре) каждого из элементов рассматриваемой системы

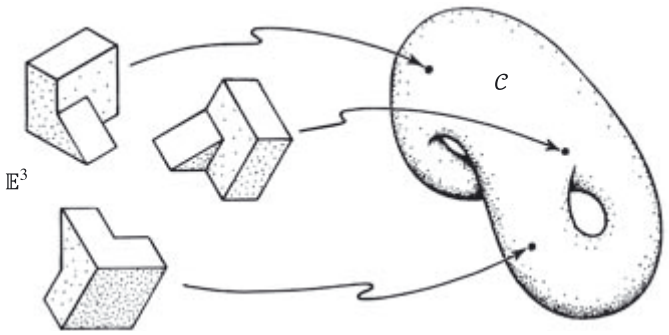


Рис. А.19. Конфигурационное пространство отдельно взятого жесткого тела неправильной формы, расположенного в евклидовом трехмерном пространстве $\mathbb{R}^3$, — это некомпактное, искривленное топологически нетривиальное шестимерное многообразие

Шестимерное пространство $\mathcal{C}$ является *некомпактным*, так как G может находиться в любой точке бесконечного евклидова трехмерного пространства; более того, $\mathcal{C}$ также обладает нетривиальной (и интересной) топологией. Оно является, так сказать, «неодносвязным», поскольку в $\mathcal{C}$ есть замкнутые кривые, которые нельзя неразрывно стянуть в точку [Tu, 2010; Lee, 2003]. Такая кривая соответствует непрерывному вращению B на 360°. Любопытно, что кривую, демонстрирующую *повтор* такого процесса, то есть непрерывное вращение на 720°, *можно* неразрывно стянуть в точку (см., например, ПкР, раздел 11.3). При этом иллюстрируется так называемое *топологическое кручение*.

В физике часто рассматриваются конфигурационные пространства, в которых насчитывается гораздо больше измерений — например, при изучении газа, когда нас может интересовать точное расположение всех его молекул. Если имеется N молекул (которые считаются отдельными точечными частицами, не имеющими внутренней структуры), то в конфигурационном пространстве будет $3N$ измерений. Разумеется, величина N может быть очень велика, но, тем не менее, общий математический аппарат для изучения многообразий, выстраиваемый нами на основе интуитивных представлений об одном, двух или трех измерениях, исключительно хорошо подходит для анализа столь сложных систем.

Фазовое пространство $\mathcal{P}$ очень похоже на конфигурационное пространство, но в нем учитываются и *движения* отдельных компонентов. Во втором примере конфигурационного пространства, рассмотренном выше, где каждая точка $3N$-многообразия $\mathcal{C}$ отражает всю совокупность местоположений всех молекул газа, соответствующее фазовое пространство представляет собой $6N$-многообразие $\mathcal{P}$, в котором также должно быть представлено движение каждой из частиц. Можно представить его себе, взяв три компонента скорости (определяющих вектор скорости) каждой частицы, но по техническим причинам оказывается, что удобнее брать три компонента *импульса* каждой частицы. Вектор импульса частицы (как минимум в ситуациях, интересующих нас здесь) — это просто произведение вектора скорости на *массу* данной частицы. Вектор скорости дает для каждой частицы еще три компонента дополнительно к тем, что были у нас ранее, то есть получается всего по шесть компонентов для каждой частицы, а фазовое пространство $\mathcal{P}$ для нашей системы, состоящей из N бесструктурных частиц, будет иметь $6N$ измерений (рис. А.20).

Если же у частиц есть какая-то внутренняя структура, то все усложняется. Ранее, как вы помните, в случае с жестким телом конфигурационное пространство уже содержит шесть измерений, поскольку должны учитываться три числа, определяющих пространственную ориентацию тела. Для описания *поворотов* тела (относительно его центра масс) нужно инкорпорировать в фазовое пространство — вдобавок к трем компонентам импульса, отражающим движение центра

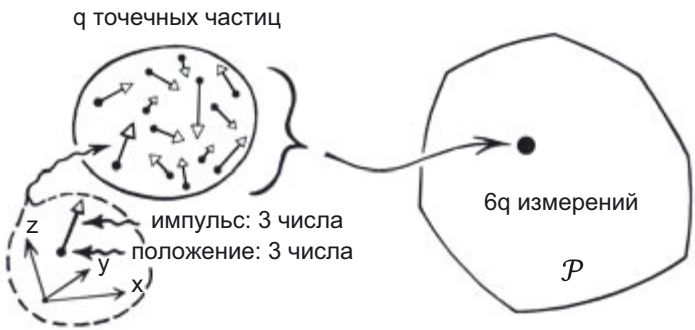


Рис. А.20. Тогда как конфигурационное пространство $\mathcal{C}$, содержащее n бесструктурных классических точечных частиц, является n-многообразием, в фазовом пространстве $\mathcal{P}$ также учитываются три импульсные степени свободы, поэтому пространство $\mathcal{P}$ является $6n$-мерным.

масс, — еще три компонента *момента импульса* вращения вокруг центра масс. Так получится двенадцатимерное фазово-пространственное многообразие $\mathcal{P}$. Следовательно, для N частиц, каждая из которых структурно подобна жесткому телу, мы получим фазовое пространство с $12N$ измерениями. Действует правило: фазовое пространство $\mathcal{P}$ физической системы должно иметь *вдвое* больше измерений, чем соответствующее ему конфигурационное пространство $\mathcal{C}$.

Фазовые пространства, будучи так называемыми *симплектическими* многообразиями, обладают многими красивыми математическими свойствами, которые особенно важны для их динамических характеристик. Как упоминалось в разделе А.5, каждое касательное пространство такого многообразия $\mathcal{P}$ имеет *антисимметричное* «внутреннее произведение» $[\mathbf{u}, \mathbf{v}] = -[\mathbf{v}, u]$, зависящее от так называемой *симплектической формы*. Отметим, что это дает нам меру для измерения модуля касательного вектора, поскольку прямо подразумевает, что $[\mathbf{u}, \mathbf{u}] = 0$ для любого касательного вектора $\mathbf{u}$. Однако симплектическая форма позволяет измерить *площадь* любого *двумерного* элемента поверхности, где $[\mathbf{u}, \mathbf{v}]$ соответствует участку площади, натянутому на два вектора: $\mathbf{u}$ и $\mathbf{v}$. В силу антисимметрии эта плоскость получается *ориентированной* в том смысле, что, если поменять порядок $\mathbf{u}$ и $\mathbf{v}$ (фактически описать площадь в противоположном направлении; см. рис. А.21), то знак площади изменится на противоположный. Доведя такую меру площади до бесконечно малой величины, можно сложить все бесконечно малые элементы (технически — *интегрировать*) и тем самым измерить любую (оговоримся: компактную, то есть имеющую конечный размер) двумерную поверхность (см. рис. А.15 *в*). Далее можно развить это представление о площади, поработав с *произведениями* таких площадей и определив, таким образом, «объем» любой четномерной (компактной) области поверхности, лежащей

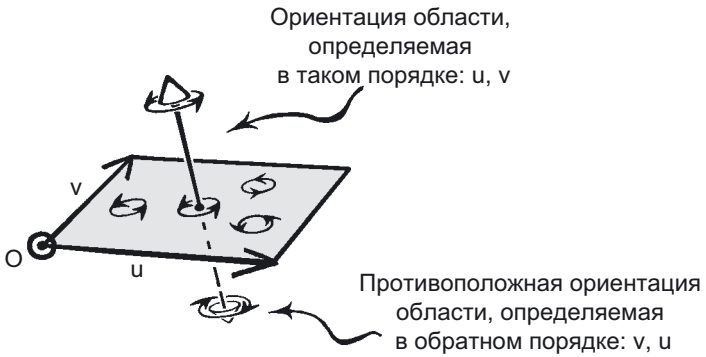


Рис. А.21. Двумерный плоский элемент, задаваемый векторами **u**, **v** в пространстве, касательном к точке многообразия, обладает ориентацией, зависящей от того, в каком порядке берутся **u** и **v**. В трехмерном окружающем пространстве можно трактовать эту ориентацию в терминах направлений в одну или другую сторону от плоскости, но лучше предположить некоторое «кручение» вокруг двумерного плоского элемента, поскольку такое понятие применимо и к более многомерному окружающему пространству. Если окружающее пространство является симплектическим многообразием, то область, которую эта симплектическая структура сопоставляет двумерной плоскости, будет иметь знак, зависящий от ориентации двумерной плоскости

в пределах $\mathcal{P}$. Это касается всего пространства $\mathcal{P}$, поскольку оно обязательно является четномерным, а также любой полномерной области в пределах $\mathcal{P}$ (где, так или иначе, *конечность* гарантируется *компактностью*). Такая мера объема называется *мерой Лиувилля*.

Хотя детальные математические характеристики симплектических многообразий не слишком важны в этой книге, здесь следует остановиться на двух особых свойствах такой геометрии. Они касаются кривых, пролегающих в $\mathcal{P}$, — так называемых *эволюционных кривых*, каждая из которых соответствует одному из возможных вариантов эволюции рассматриваемой физической системы с течением времени, причем считается, что такая эволюция согласуется с динамическими уравнениями, управляющими системой (речь может идти о простой динамике классической теории Ньютона, либо о более затейливой динамике теории относительности, либо о многих других физических теориях). Такая динамика считается *детерминированной* в том смысле, что поведение обычных классических физических систем, состоящих из отдельных частиц, полностью зависит от местоположений и импульсов всех этих частиц в любой избранный нами момент t. Если в системе есть динамические сплошные поля (например, электромагнитные), то также ожидается схожая детерминированная эволюция. Соответственно, в контексте фазового пространства $\mathcal{P}$ каждая эволюционная

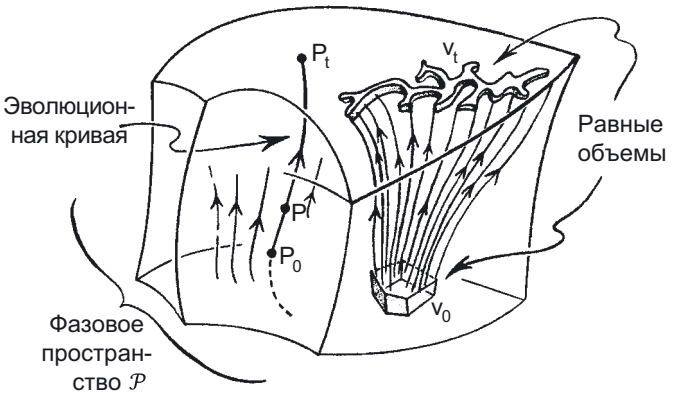


Рис. А.22. Здесь при помощи эволюционной кривой описана динамическая эволюция классической системы в терминах фазового пространства $\mathcal{P}$. Каждая точка P в пространстве $\mathcal{P}$ соответствует одномоментному местоположению и движениям всех составляющих элементов системы, а затем динамические уравнения определяют эволюцию системы. Получается эволюционная кривая, исходящая из точки P и достигающая некоторой точки P_t, которая соответствует состоянию системы на некоторый момент t в будущем. Детерминизм динамических уравнений означает, что через каждую точку P в направлении будущего проходит уникальная эволюционная кривая, которая продолжается и в прошлое до тех пор, пока не достигает некоторой исходной точки P_0, соответствующей первичному состоянию системы. Симплектическая структура P позволяет узнать объем (меру Лиувилля) любой компактной области в $\mathcal{V}$, а согласно теореме Лиувилля, этот объем сохраняется, распространяясь вдоль эволюционных кривых, независимо от того, сколь запутанной может стать эта область

кривая c, соответствующая одному из возможных вариантов целостной эволюции системы, полностью фиксируется в любой точке, выбранной на этой кривой. Все семейство эволюционных кривых образует так называемое *слоение* $\mathcal{P}$, причем через любую выбранную нами точку в $\mathcal{P}$ пролегает ровно одна эволюционная кривая.

Первая из этих особенностей, связанная с симплектической природой $\mathcal{P}$, дает нам следующее: точные положения всех этих эволюционных кривых $\mathcal{P}$ полностью фиксируются, как только становится известно значение *энергии* в системе в каждой точке $\mathcal{P}$ (эта энергетическая функция именуется *гамильтонов функционал*), однако эта примечательная и важная роль энергии не имеет непосредственного отношения к излагаемым здесь вещам. Вторая особенность, *важная* для нас в данном случае, заключается в том, что естественная мера Лиувилля, определяемая на фазовых пространствах (и зависящая от их симплектической структуры), *сохраняется* с течением времени согласно имеющимся динамическим законам. Это поразительный факт, формулировка которого называется *теоремой Лиувилля*. В $2n$-мерном фазовом пространстве $\mathcal{P}$ данный объем имеет *размер*,

выражаемый вещественным числом по формуле $L_n(\mathcal{V})$ и присущий любым (компактным) $2n$-мерным субобластям $\mathcal{V}$ пространства $\mathcal{P}$. По мере увеличения параметра t и движения точек $\mathcal{P}$ по их эволюционным кривым в пределах $\mathcal{P}$ также будет двигаться и весь регион $\mathcal{V}$ таким образом, что $2n$-объем $L_n(\mathcal{V})$ всегда будет оставаться одинаковым. Это особенно важно в контексте главы 3.

А.7. Расслоения

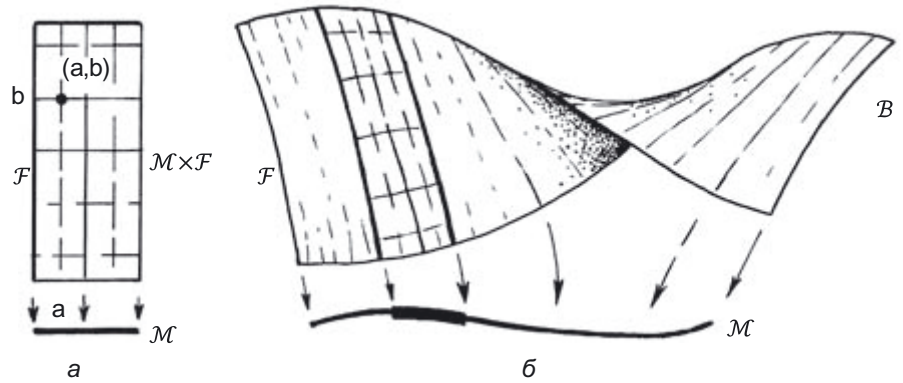
Важное математическое понятие, играющее ключевую роль в современных представлениях о типах структур, которые могут находиться на многообразиях, либо о силах природы, называется *тривиальное расслоение* или просто *расслоение* [Steenrod, 1951; ПкР, глава 15]. Расслоение можно считать таким понятием, которое позволяет вписать «поле» в универсальном физическом понимании в общую геометрическую систему многообразий, описанную в разделе А.4. Расслоения также помогут нам лучше разобраться в вопросе функциональной свободы, поднятом в разделе А.2.

С нашей нынешней точки зрения можно представить расслоение $\mathcal{B}$ как $(r + d)$-многообразие, которое плавно строится из непрерывного семейства копий r-многообразия $\mathcal{F}$, обладающего меньшим числом измерений, и эти копии называются *слоями* $\mathcal{B}$. Структурно это семейство само по себе должно было бы иметь форму иного многообразия $\mathcal{M}$ — это было бы d-многообразие, именуемое *базовым пространством*, так что каждая точка базового пространства $\mathcal{M}$ соответствовала бы конкретному экземпляру многообразия $\mathcal{F}$ в целом семействе, образующем $\mathcal{B}$. Грубо говоря, мы могли бы понимать наше расслоение $\mathcal{B}$ следующим образом:

$\mathcal{B}$ — это непрерывное $\mathcal{M}$, выраженное в $\mathcal{F}$.

$\mathcal{B}$ называется слоем $\mathcal{F}$ на $\mathcal{M}$, где все $\mathcal{B}$ само является многообразием, размерность которого равна сумме размерностей $\mathcal{M}$ и $\mathcal{F}$. Описание $\mathcal{B}$ как непрерывного $\mathcal{M}$, выраженного в $\mathcal{F}$, с технической точки зрения правильнее понимать как *проекцию* π, где $\mathcal{B}$ отображается на $\mathcal{M}$, а *прообраз* любой точки $\mathcal{M}$ (то есть целая часть $\mathcal{B}$, отображаемая в π на эту конкретную точку) оказывается одной из копий $\mathcal{F}$, составляющих $\mathcal{B}$. Это означает, что проекция π постепенно расплющивает в точку $\mathcal{M}$ каждый отдельный $\mathcal{F}$ в составе $\mathcal{B}$ (рис. А.23). Таким образом, базовое пространство $\mathcal{M}$ и пространство слоев $\mathcal{B}$ комбинируются и дают общее пространство $\mathcal{B}$ данного многообразия.

Мы хотим, чтобы это описание получилось *непрерывным*; в частности, наша проекция должна образовывать сплошное отображение (то есть не содержать



а *б*

Рис. А.23. Рисунок помогает понять, что такое тривиальное расслоение. Общее пространство $\mathcal{B}$ — это многообразие, которое можно считать «непрерывным $\mathcal{M}$, выраженным в $\mathcal{F}$», где $\mathcal{M}$ именуется *базовым пространством*, а $\mathcal{F}$ — *слоем*. Существует *проекция* π (обозначенная стрелками), отображающая каждый экземпляр $\mathcal{F}$ в $\mathcal{B}$ на слой, расположенный «над» этой точкой в $\mathcal{M}$. *а*) Над любым достаточно малым открытым подмножеством $\mathcal{M}$ будет область $\mathcal{B}$, являющаяся *произведением пространств*, а именно произведением этого подмножества и $\mathcal{F}$ (см. рис. А.25), но *б*) $\mathcal{B}$ в целом не должно быть таким произведением, поскольку в его структуре присутствует глобальная «скрученность»

скачков). Однако здесь я также требую, чтобы все наши отображения и пространства были гладкими (технически предпочтительно, чтобы это были так называемые $\mathcal{C}^\infty$ (см., например, ПкР, раздел 6.3)), чтобы их можно было по мере необходимости описывать на языке математического анализа. Однако эта книга рассчитана и на читателей, не знакомых с теорией математического анализа (некоторые базовые концепции излагаются в разделе А.11). Тем не менее некоторое интуитивное представление о дифференцировании, интегрировании, касательных векторах и т. д. читателю безусловно бы пригодилось (этих вопросов мы касаемся и в разделе А.5). Так, *дифференцирование* работает со скоростью изменения и крутизной кривых и т. д., а *интегрирование* — с площадями и объемами и т. д. Эти понятия помогли бы сориентироваться во многих местах книги (см. рис. А.44 в разделе А.11).

Два простых примера многообразий показаны на рис. А.24; в данном случае базовое пространство $\mathcal{B}$ — это *круг*, а пространство слоя $\mathcal{F}$ — отрезок. Две иные с топологической точки зрения возможности — это *цилиндр* (рис. А.24 *а*) и *лента Мёбиуса* (рис. А.24 *б*). Цилиндр — это пример так называемого *произведения пространств*, или *тривиального* расслоения, где произведение $\mathcal{M} \times \mathcal{F}$ двух пространств $\mathcal{M}$ и $\mathcal{F}$ должно мыслиться как пространство, состоящее из *пар* (a, b); в это пространстве все a относятся к $\mathcal{M}$, а все b — к $\mathcal{F}$ (рис. А.25). Можно отметить,

что такое представление о произведении применимо и к парам положительных целых чисел. Ведь в парах (*a*, *b*), где *a* принимает значения от 1, 2, 3 до *A*, а *b* — от 1, 2, 3 до *B*, произведение действительно выражается просто по формуле *AB*.

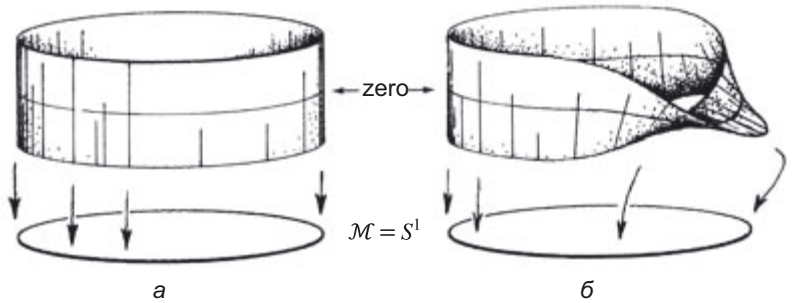


Рис. А.24. Два возможных расслоения, где $\mathcal{F}$ — это отрезок, а базовое пространство $\mathcal{M}$ — это круг S^1. Они расположены в виде цилиндра (*а*) и в виде ленты Мёбиуса (*б*)

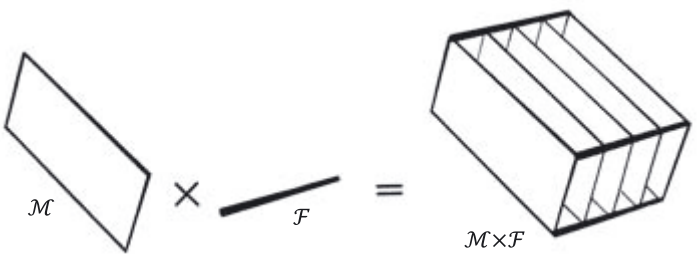


Рис. А.25. Произведение пространств $\mathcal{M} \times \mathcal{F}$, состоящее из многообразий $\mathcal{M}$ и $\mathcal{F}$, — это конкретный тип многообразия $\mathcal{F}_n$ на $\mathcal{M}$, так называемое тривиальное расслоение, состоящее из пар (*a*, *b*), где *a* — точка на $\mathcal{M}$, а *b* — точка на $\mathcal{F}$. Его также можно считать тривиальным расслоением $\mathcal{M}$ на $\mathcal{F}$.

Более общий случай иногда именуется *скрещенное произведение*; примером такого произведения является лента Мёбиуса. Этот пример показывает, что расслоение *локально* и всегда является произведением пространств — в том смысле, что если взять любую точку *a* из базового пространства $\mathcal{M}$, то в $\mathcal{M}$ найдется такой достаточно небольшой регион $\mathcal{M}_a$, для которого *часть* $\mathcal{B}_a$ из расслоения $\mathcal{B}$, лежащая *над* $\mathcal{M}_a$ (то есть часть $\mathcal{B}$, проецируемая π на $\mathcal{M}_a$), может быть выражена в форме произведения:

$$\mathcal{B}_a = \mathcal{M}_a \times \mathcal{F}.$$

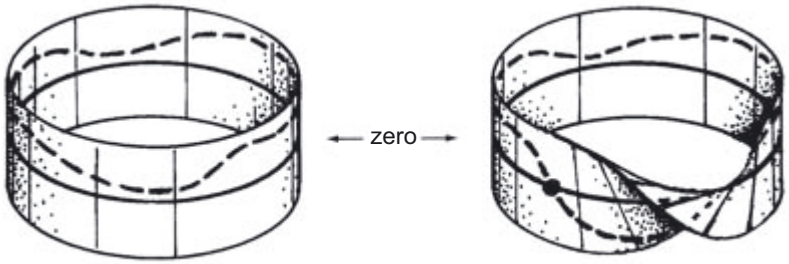


Рис. А.26. Прерывистыми линиями обозначены поперечные сечения расслоений с рис. А.24. Один из способов отличить два расслоения друг от друга: у цилиндра много таких поперечных сечений, которые нигде не обращаются в нуль (*а*), а в случае ленты Мёбиуса у каждого поперечного сечения есть нулевая точка (пересечение с нулевой линией) (*б*)

Такая *локальная* структура произведения всегда соблюдается в расслоении, даже если все расслоение целиком не удается (непрерывно) представить таким образом, — кстати, в случае с лентой Мёбиуса это как раз невозможно (см. рис. А.24 *б*).

Такое четкое топологическое различие между цилиндром и лентой Мёбиуса можно понять в контексте так называемых *поперечных сечений* расслоений. Поперечное сечение расслоения $\mathcal{B}$ — это подмногообразие $\mathcal{X}$ в составе $\mathcal{B}$ (то есть меньшее многообразие $\mathcal{X}$, гладко вписанное в $\mathcal{B}$), пересекающее каждый слой ровно в одной точке. (Иногда поперечное сечение удобно представить как образ некоторого отображения, спроецированного с базового пространства $\mathcal{M}$ обратно на многообразие $\mathcal{B}$, с сохранением вышеупомянутого свойства, поскольку $\mathcal{X}$ всегда будет топологически идентично $\mathcal{B}$.) Как и в случае любых *произведений* пространств (когда $\mathcal{F}$ содержит более одной точки), найдутся взаимно непересекающиеся поперечные сечения (как, например, a_1, b и a_2, b, где a_1 и a_2 — разные элементы $\mathcal{F}$, а b целиком охватывает $\mathcal{M}$). Это хорошо проиллюстрировано в случаях с цилиндром на рис. А.26 *а*. Но в случае с лентой Мёбиуса *любые* два поперечных сечения должны пересекаться (в чем мы уже могли убедиться; см. рис. А.26 *б*). Пример демонстрирует топологическую нетривиальность ленты Мёбиуса.

Поперечные сечения многообразий важны с физической точки зрения потому, что позволяют построить красивое геометрическое представление *физического* поля, тогда как в виде $\mathcal{M}$ мы представляем пространство или пространство-время. Вспомните магнитные поля, рассмотренные в разделе А.2. Такое поле можно уподобить поперечному сечению многообразия, базовым пространством которого является обычное евклидово трехмерное пространство, а слоем над

каждой точкой P — трехмерное векторное пространство всех магнитных полей, которые могут существовать в точке P. Вскоре, в разделе А.8, мы вновь об этом поговорим. В данном случае нас интересует феномен *гладких* поперечных сечений. Имеется в виду не только то, что гладкими должны быть все рассматриваемые пространства и отображения, но и то, что поперечное сечение $\mathcal{X}$ везде должно быть *трансверсально* слоям — в том смысле, что в точке P_0, где $\mathcal{X}$ пересекается со слоем $\mathcal{F}_0$, не будет такой касательной к $\mathcal{X}$, которая совпадала бы с касательной к $\mathcal{F}_0$. На рис. А.27 приведены примеры, в которых удовлетворяется или нарушается трансверсальность.

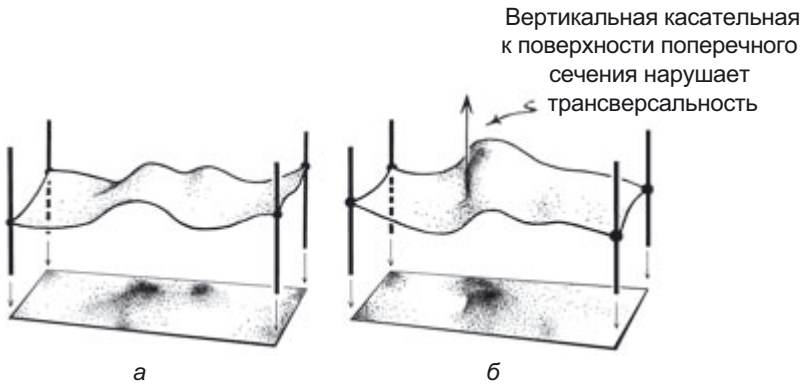


Рис. А.27. Условие трансверсальности для поперечного сечения. На данном локальном изображении базовому пространству соответствует плоская поверхность, а слоям — вертикальные линии. *а*) Условие трансверсальности удовлетворяется, волны на поперечном сечении никогда не достигают вертикальной крутизны; б) это поперечное сечение хотя и является гладким, достигает вертикальной крутизны и, следовательно, не удовлетворяет условию трансверсальности (у представленного здесь поля будет бесконечная производная)

Важно понимать: для того чтобы расслоение получалось *нетривиальным* (то есть не было произведением пространств), необходимо, чтобы пространство слоя $\mathcal{F}$ обладало той или иной *симметрией*. В случае с лентой Мёбиуса симметрия заключается в возможности перевернуть линию (то есть слой $\mathcal{F}$) из конца в конец, не меняя саму линию, — в этом и заключается нетривиальность примера. Такой механизм достаточно универсален, и если пространство $\mathcal{F}$ не обладает вообще никакой симметрией, то мы не сможем построить нетривиальное пространство, взяв $\mathcal{F}$ в качестве слоя. Этот факт важен и при рассмотрении *калибровочных теорий*, лежащих в основе современных теорий естественных взаимодействий (см. раздел 1.8). Калибровочные теории зависят от так называемой *калибровочной связности*, нетривиальность которой коренным образом зависит от того, об-

ладают ли слои $\mathcal{F}$ нетривиальной (непрерывной) симметрией, так что смежные слои в расслоении можно соотносить друг с другом немного по-разному, в зависимости от рассматриваемого варианта «связности».

Здесь важно обратить внимание на один термин. В любом расслоении $\mathcal{B}$ с базовым пространством $\mathcal{M}$ и слоем $\mathcal{F}$ пространство $\mathcal{M}$ можно назвать *факторным пространством* $\mathcal{B}$. Естественно, это касается тривиального случая с произведением расслоений, где и $\mathcal{M}$, и $\mathcal{F}$ являются элементами факторного пространства $\mathcal{M} \times \mathcal{F}$. Необходимо сравнить такое положение с совершенно иной ситуацией, где пространство $\mathcal{M}$ является *подпространством* другого пространства S, если его можно без проблем отождествить с некоторым регионом внутри S и записать:

$$\mathcal{M} \hookrightarrow \mathcal{S}.$$

Очевидное различие между двумя этими явно разными феноменами (которые, как ни странно, все равно часто путают) важно в теории струн (см. разделы 1.10, 1.11 и 1.15, а также рис. 1.32 в разделе 1.10).

Конкретный класс расслоений, которые очень важны как в физике, так и в чистой математике, — это так называемые *векторные расслоения*, для которых пространство слоя $\mathcal{F}$ является *векторным пространством* (см. раздел А.3). Примеры векторных расслоений будут актуальны и в контексте магнитных полей, рассмотренных в разделе А.2, — как мы увидим в разделе А.8. Возможные значения магнитных полей в каждой конкретной точке образуют векторное пространство. То же будет касаться и электрических, и многих других физических полей; такие поля можно складывать друг с другом или умножать на вещественные скалярные числа, получая таким образом другие поля для рассмотрения. Другой класс примеров — *фазовые пространства*, рассмотренные в разделе А.6. В данном (втором) случае речь идет о своеобразном векторном пространстве, так называемом *кокасательном расслоении* $T^*(\mathcal{C})$ конфигурационного пространства $\mathcal{C}$, которое автоматически оказывается и симплектическим многообразием, упомянутым в разделе А.6.

Как определяется кокасательное расслоение? *Касательное* расслоение $T(\mathcal{M})$ n-многообразия $\mathcal{M}$ — это расслоение, касательное к данной точке (см. раздел А.5). Каждое касательное пространство — это n-мерное векторное пространство, поэтому общее пространство $T(\mathcal{M})$ является n-многообразием (рис. А.28 *а*). Кокасательное расслоение $T^*(\mathcal{M})$, относящееся к пространству $\mathcal{M}$, создается точно таким же образом, за тем исключением, что слой в каждой точке $\mathcal{M}$ теперь является *кокасательным* пространством (*дуальным* касательному пространству; см. раздел А.4) в этой точке (рис. А.28 *б*). Когда $\mathcal{M}$ является конфигурационным

пространством $\mathcal{C}$ некоторой (классической) физической системы, кокасательные векторы можно приравнять к системе *импульсов*, тогда как кокасательное расслоение $T^*(\mathcal{C})$ будет отождествляться с *фазовым пространством* системы (см. раздел А.6). Следовательно, обычное фазовое пространство действительно будет общим пространством (обычно нетривиального) векторного расслоения над соответствующим конфигурационным пространством, где слои обеспечивают всевозможные импульсы, а проекция π является попросту отображением, в котором все эти импульсы «забываются».

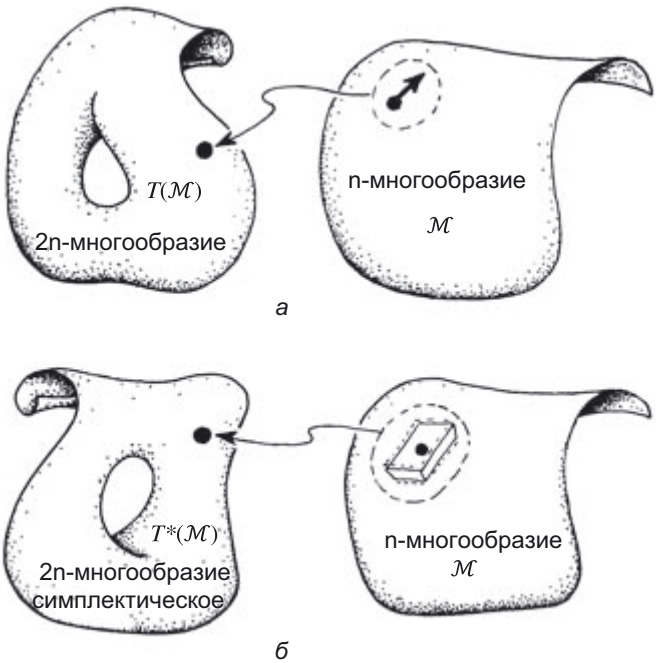


Рис. А.28. Каждая точка в $2n$-мерном касательном расслоении $T(\mathcal{M})$ n-многообразия $\mathcal{M}$ представляет точку $\mathcal{M}$ плюс касательный вектор к $\mathcal{M}$ в этой точке (*а*); каждая точка симплектического $2n$-мерного кокасательного расслоения $T^*(\mathcal{M})$ в $\mathcal{M}$ представляет точку в $\mathcal{M}$ плюс расположенный там касательный ковектор к $\mathcal{M}$(*б*)

К другим примерам расслоений, естественным образом возникающих в физике, относятся основополагающие квантово-механические расслоения, такие как, например, показанное на рис. 2.16 *б* в разделе 2.8, где комплексное n-мерное векторное пространство, известное под названием *гильбертова пространства* $\mathcal{H}^n$ (за исключением начала координат 0), считается расслоением над *проективным* гильбертовым пространством $\mathbb{P}\mathcal{H}^n$, каждый слой которого является экземпляром

плоскости Весселя (см. раздел А.10), в которой удалено начало координат. Более того, $(2n-1)$-сфера S^{2n-1}, образованная нормированными векторами гильбертова пространства, — это расслоение с круглыми слоями (S^1-расслоение) на $\mathbb{P}\mathcal{H}^n$. Другие образцы расслоений, важные с физической точки зрения, фигурируют в калибровочных теориях физических взаимодействий, и об этом шла речь выше. Особенно интересен следующий факт, упомянутый в разделе 1.8: расслоение, описывающее (вейлевскую) калибровочную теорию электромагнетизма, фактически оказывается пятимерным «пространством-временем» Калуцы — Клейна, и пятым измерением в этой теории является круг, вдоль которого выстраивается *симметрия*. Все пятимерное многообразие оказывается круглым, наложенным на четырехмерное многообразие обычного пространства-времени (см. рис. 1.12 в разделе 1.6). Направление симметрии задается так называемым *(векторным) полем Киллинга*, вдоль которого метрическая структура многообразия остается неизменной.

С этим феноменом связано и представление о стационарном *пространстве*-времени, обладающем таким глобальным вектором **k**, который везде времениподобен и остается неизменным вдоль направления времени, задаваемого вектором **k**. Если **k** ортогонален семейству пространственноподобных трехмерных поверхностей, то говорят, что пространство-время статично, как показано на рис. А.29. Однако расслоенная структура, образованная времениподобными кривыми, протяженными в направлении **k**, может показаться несколько неестественной,

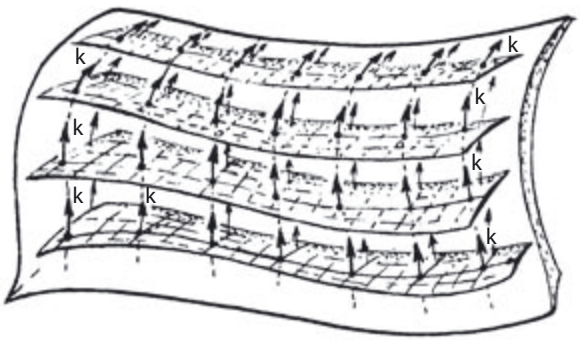



Рис. А.29. На многообразии $\mathcal{M}$, имеющем метрическую структуру (таково, например, пространство-время в общей теории относительности), может существовать векторное поле Киллинга **k**, выражающее (возможно, лишь локально) непрерывную симметрию $\mathcal{M}$. Если $\mathcal{M}$ — это пространство-время, а **k** времениподобно, то $\mathcal{M}$ называется стационарным. Если же, вдобавок, **k** ортогонально семейству метрически идентичных пространственноподобных 3-поверхностей $\mathcal{S}$, как показано здесь, то $\mathcal{M}$ называется статическим, но обычно не следует считать, что $\mathcal{M}$ обладает структурой расслоения (так или иначе), поскольку участки шкалы времени вдоль $\mathcal{S}$ могут различаться

поскольку на самом деле эти кривые в большинстве своем не эквивалентны друг другу, шкала времени на них различается.

Как упоминалось выше, чтобы расслоение было *нетривиальным*, пространство слоя $\mathcal{F}$ должно обладать некоторой *симметрией* (например, допускать сквозной оборот, как лента Мёбиуса). Иные симметричные преобразования, применимые к некоторой структуре, образуют математическую *группу*. Строго говоря, группа в абстрактных категориях — это система операций a, b, c, d и т. д., которые могут применяться последовательно и записываться (как обычное умножение) просто ab и т. д. Эти операции всегда удовлетворяют $(ab)c = a(bc)$, и существует единичный элемент e, для которого $ae = a = ea$ для всех a. Также у каждого элемента a есть обратное a^{-1}, так что $a^{-1}a = e$. Различные группы, часто используемые в физических теориях, получают собственные имена, например $\mathrm{O}(n)$, $\mathrm{SO}(n)$, $\mathrm{U}(n)$, где, в частности, $\mathrm{SO}(3)$ — это группа вращений обычной сферы в евклидовом пространстве, не допускающая отражений, а $\mathrm{O}(3)$ — аналогичная группа, но отражения в ней разрешены. $\mathrm{U}(n)$ — это группа симметрий для n-мерного гильбертова пространства, описанного в разделе 2.8; так, в частности, $\mathrm{U}(1)$ — она же $\mathrm{SO}(2)$ — это унимодулярная группа фазовых вращений в плоскости Весселя, то есть умножение на $e^{i\theta}$ (θ — вещественное число).

А.8. Функциональная свобода на уровне расслоений

Здесь нас особенно интересует концепция векторных расслоений, поскольку позволяет дополнительно разобраться в проблемах функциональной свободы, рассмотренных на «интуитивном» уровне в разделе А.2. Чтобы понять это, давайте вновь вернемся к такой проблеме: почему в физике нас особенно интересуют (гладкие) поперечные сечения расслоений? Ответ, отчасти уже данный выше, таков: *физические поля* можно интерпретировать как такие поперечные сечения, а их гладкость (и в частности трансверсальность) выражает гладкость рассматриваемого поля. Здесь мы будем считать базовое пространство $\mathcal{M}$ *физическим пространством* (обычно имеется в виду трехмерное многообразие) или *физическим пространством-временем* (обычно это четырехмерное многообразие). Условие трансверсальности означает, что *производная* (градиент, или скорость изменения в пространстве или во времени) в рассматриваемом поле всегда будет конечной.

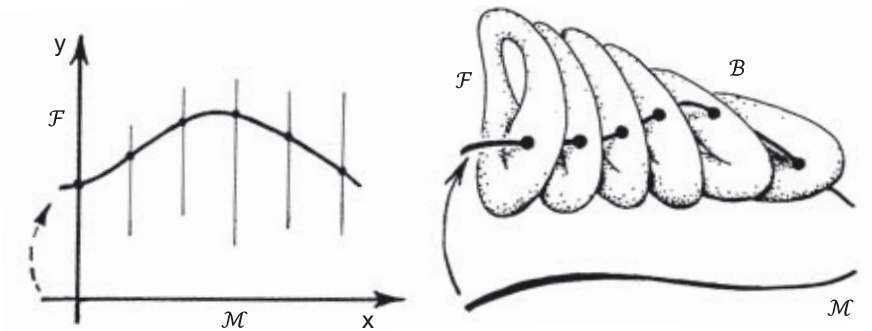
Приведем конкретный пример, иллюстрирующий эту концепцию. Есть скалярное поле, определенное в пространстве $\mathcal{M}$. В таком случае $\mathcal{F}$ будет просто копией континуума вещественных чисел $\mathbb{R}$, так как *скалярное* поле (в данном

контексте) — это всего лишь присваивание вещественного числа (*силы* поля) каждой точке $\mathcal{M}$. Соответственно, будем считать наше расслоение тривиальным:

$$\mathcal{B} = \mathcal{M} \times \mathbb{R}$$

без какой-либо «закрутки». Поперечное сечение $\mathcal{X}$ в $\mathcal{B}$ позволит нам гладко присвоить вещественное число каждой точке в $\mathcal{M}$ — это и *есть* скалярное поле. Чтобы получить простое представление о том, что в данном случае происходит, достаточно вообразить обычный график функции, где $\mathcal{M}$ также будет считаться одномерным, просто еще одной копией $\mathbb{R}$ (рис. А.30 *а*). Сам график является поперечным сечением. Трансверсальность требует, чтобы кривая никогда не достигала вертикальной крутизны. Отвесный график означает, что в этой точке функция имеет бесконечную производную, а в гладком поле это недопустимо. Это очень специфический вариант поля, представленного в качестве поперечного сечения расслоения. Другие возможные «значения», которые могло бы иметь это поле, могут образовывать отнюдь не линейное пространство, а, вполне возможно, сложное многообразие с нетривиальной топологией как на рис. А.30 *б*, где само пространство-время может быть более сложным пространством.

В качестве примера чуть более сложного, чем на рис. А.30 *а*, обратимся к магнитным полям из раздела А.2. Здесь речь идет об обычном трехмерном пространстве,



а *б*

Рис. А.30. Обычный график функции $y = f(x)$ простейшим образом иллюстрирует поперечное сечение расслоения, описывающего физическое поле. Здесь слои обозначены в виде вертикальных линий, пересекающих график, некоторые из которых втянуты, а горизонтальной осью в данном простейшем случае служит многообразие $\mathcal{M}$. По условию трансверсальности кривая не может достигать вертикальной кривизны (*а*); показан общий случай, где $\mathcal{M}$, а также слой $\mathcal{F}$ с теми значениями, которые поле может принимать в данной точке многообразия $\mathcal{M}$, могут быть обычными многообразиями. Любая конкретная конфигурация поля будет представлять собой поперечное сечение расслоения (и удовлетворять трансверсальности) (*б*)

так что $\mathcal{M}$ — это трехмерное многообразие (евклидово трехмерное пространство), а $\mathcal{F}$ — это трехмерное пространство возможных магнитных полей в точке (это опять трехмерное пространство, так как нам нужно три компонента, чтобы определить магнитное поле в каждой точке). Можно отождествить $\mathcal{F}$ с $\mathbb{R}$ (это пространство, состоящее из троек (B_1, B_2, B_3) вещественных чисел, являющихся 3-компонентами магнитного поля; см. раздел А.2). Наше расслоение $\mathcal{B}$ можно считать просто «тривиальным» произведением $\mathcal{M} \times \mathbb{R}^3$ $\mathcal{M} \times \mathbb{R}^3$. Итак, поскольку у нас имеется магнитное *поле*, а не просто значение поля в конкретной точке, мы ищем гладкое *поперечное сечение* расслоения $\mathcal{B}$ (рис. А.31). Магнитное поле — пример *векторного поля*, где каждой точке базового пространства (в данном случае $\mathcal{M}$) гладко присваивается вектор. В принципе, векторное поле является просто гладким поперечным сечением некоторого векторного расслоения, но этот термин часто используется именно в тех случаях, когда речь идет о расслоении, *касательном* к рассматриваемому пространству (см. рис. А.17 в разделе А.6).

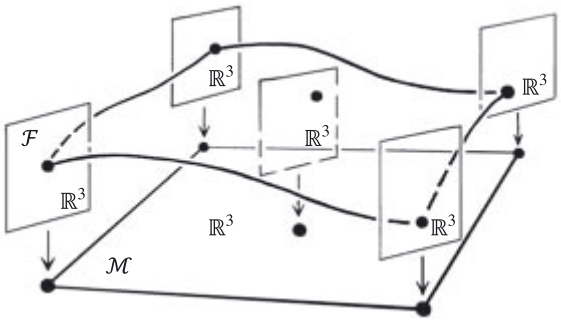


Рис. А.31. Иллюстрация помогает понять, как можно представить магнитное поле в плоском трехмерном пространстве ($\mathbb{R}^3$) в виде поперечного сечения тривиального $\mathbb{R}^3$-расслоения на $\mathbb{R}^3$ (то есть $\mathbb{R}^3 \times \mathbb{R}^3$), где все плоскости должны считаться $\mathbb{R}^3$

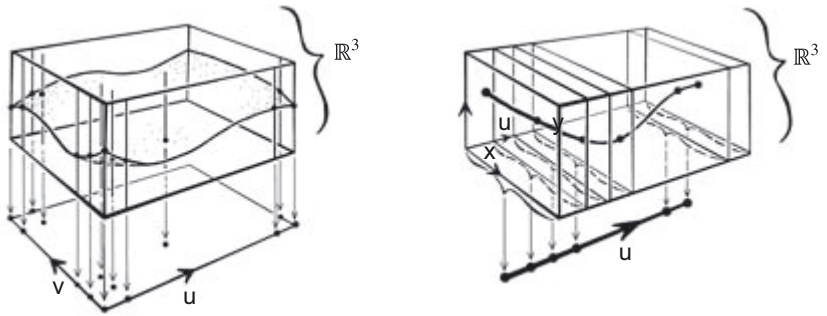
Если $\mathcal{M}$ относится к категории искривленных трехмерных пространств, одно из которых фигурирует в общей теории относительности, тождество $\mathcal{B} = \mathcal{M} \times \mathbb{R}^3$ на самом деле работать не будет, поскольку невозможно будет естественным образом взять и отождествить касательные пространства в различных точках $\mathcal{M}$. Во многих многомерных имитационных моделях касательное расслоение $\mathcal{B}$ к d-мерному $\mathcal{M}$ даже *топологически* не будет совпадать с $\mathcal{M} \times \mathbb{R}^d$ (но есть любопытное исключение — случай $d = 3$). Правда, в данном контексте такие глобальные проблемы не слишком важны, поскольку даже в общей теории относительности мы рассуждаем в совершенно локальных пространственных (или пространственно-временных) категориях. Для таких случаев вполне подходит «тривиальная» локальная структура $\mathcal{M} \times \mathbb{R}^d$.

Рассматривать ситуацию в таком ракурсе полезно именно потому, что на таких примерах проблема *функциональной свободы* особенно очевидна. Предположим, у нас есть n-компонентное поле, определенное на d-мерном многообразии $\mathcal{M}$. В таком случае нас интересует (гладкое) поперечное сечение $\mathcal{X}$ (d+n)-мерного расслоения $\mathcal{B}$. Мы обсудим лишь локальное поведение в $\mathcal{M}$, поэтому вполне можем предположить, что работаем с *тривиальным* расслоением $\mathcal{B} = \mathcal{M} \times \mathbb{R}^n$. Если поле выбрано абсолютно произвольно, то многообразие $\mathcal{X}$ окажется свободно выбранным d-мерным подмногообразием (d+n)-мерного многообразия $\mathcal{B}$. ($\mathcal{X}$ является d-многообразием, поскольку, как отмечалось в разделе А.4, оно топологически идентично $\mathcal{M}$). Правда, строго говоря, нельзя сказать, что $\mathcal{X}$ выбирается абсолютно произвольно, поскольку, во-первых, нужно гарантировать, что повсюду выполняется условие трансверсальности, а во-вторых, что $\mathcal{X}$ не «перекручено» тем или иным образом, то есть что сечение проходит через каждый слой всего один раз. Однако эти условия не так важны в контексте функциональной свободы, поскольку на *локальном* уровне типичное d-многообразие, выбранное в рамках ($d + n$)-многообразия $\mathcal{B}$, действительно будет трансверсальным, и в этом районе мы пересечем каждый слой $\mathcal{F}$ всего по одному разу. Степень (локальной) функциональной свободы в заданном d-многообразии попросту равна (локальной) свободе при выборе d-многообразия $\mathcal{X}$ в пределах ($d + n$)-многообразия $\mathcal{B}$.

Итак, ключевая проблема заключается именно в первоочередной важности значения d, и не столь важно, какова величина n (или $d + n$). Как же это «увидеть»? Как получить представление о том, «сколько» d-многообразий умещается в многообразии ($d + n$)?

Целесообразно рассмотреть случаи $d = 1$ и $d = 2$, иными словами, кривые и поверхности в окружающем трехмерном многообразии (которое может представлять собой обычное евклидово трехмерное пространство), поскольку в таком случае легко визуализировать происходящее (рис. А.32). При $d = 2$ мы рассматриваем обычные скалярные поля в двумерном пространстве, поэтому базовое пространство (локально) можно принять как $\mathbb{R}^2$, а слой — как $\mathbb{R}^1 (= \mathbb{R})$, поэтому наши поперечные сечения окажутся просто *поверхностями* (двумерными поверхностями) в $\mathbb{R}^3$ (евклидовом трехмерном пространстве). Функциональная свобода, то есть «число» скалярных полей, которые мы можем свободно выбрать, соответствует свободе выбора двумерных поверхностей в трехмерном пространстве (рис. А.32 *а*). Но в случае $d = 1$ *базовое* пространство локально является $\mathbb{R}^1$, а *слой* — $\mathbb{R}^2$, так что поперечные сечения оказываются просто *кривыми* в $\mathbb{R}^3$ (рис. А.32 *б*).

Теперь вопрос: почему поверхностей в $\mathbb{R}^3$ настолько больше, чем кривых в $\mathbb{R}^3$? Иными словами (если переформулировать эту проблему в терминах поперечных сечений расслоений, то есть в терминах функциональной свободы полей),



а *б*

Рис. А.32. *а*) Функциональная свобода ∞^{∞^2} при выборе однокомпонентных полей в двумерном пространстве точно такая же, как и при выборе двумерных поверхностей в $\mathbb{R}^3$, причем последнее рассматривается как расслоение $\mathbb{R}^1$ на $\mathbb{R}^2$ ((u, v)-плоскости). Сравним случай на рис. *а* со случаем на рис. *б*, где рассмотрена функциональная свобода $\infty^{2\infty}$ при выборе двухкомпонентных полей ((x, y)-плоскостей) в одномерном пространстве (u-координатном), точно такая же, как при выборе одномерных поверхностей (то есть кривых) в $\mathbb{R}^3$, причем последнее рассматривается как расслоение $\mathbb{R}^2$ на $\mathbb{R}^1$

почему $\infty^{\infty^2} >> \infty^{2\infty}$ (или $\infty^{1\infty^2} >> \infty^{2\infty^1}$), если пользоваться нотацией из раздела А.2? Во-первых, следует пояснить число 2 в $\infty^{2\infty}$. Если мы хотим описать кривую, для этого достаточно рассмотреть за один раз один компонент слоя $\mathcal{F}$ в $\mathbb{R}^2$. Это означает, что нужно просто рассмотреть *проекцию* нашей кривой в двух разных направлениях, то есть в двух координатных направлениях, и получить таким образом *две* кривые в каждой из двух плоскостей — это плоскости (x, u) и (y, u), где x, y — координаты слоя, а u — координата базового пространства. Такая *пара* плоских кривых эквивалентна исходной пространственной кривой. Свобода каждой плоской кривой равна ∞^{∞} (одна гладкая функция, выраженная в вещественных числах, с единственной вещественночисленной переменной). Поэтому свобода для пары кривых будет равна $\infty^{\infty} \times \infty^{\infty} = \infty^{2\infty}$.

Для того чтобы понять, почему (локальная) свобода ∞^{∞^2} двумерных поверхностей в $\mathbb{R}^3$ гораздо больше этой — на самом деле даже больше свободы любого конечного числа плоских кривых (в локальном масштабе), — можно рассмотреть любое конечное число k параллельных сечений двумерной поверхности (то есть значения нашей скалярной функции для k разных фиксированных значений двух координат u, v в базовом пространстве $\mathbb{R}^2$, ведь на рис. А.33 каждая из этих k кривых обладает локальной свободой ∞^{∞}, поэтому общая свобода составит $(\infty^{\infty})^k = \infty^{k\infty}$). (Естественно, семейство кривых k можно трактовать как всего одну кривую, если допустить, что эта кривая может быть прерывистой. Это одна из причин, по которой такие выкладки применимы лишь *локально*. Локальный

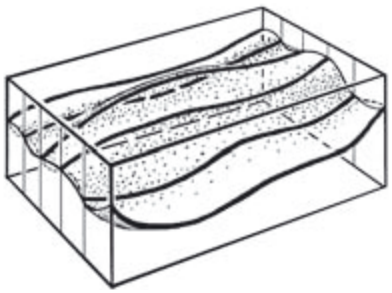


Рис. А.33. Здесь показано, почему $\infty^{\infty^2} >> \infty^{k\infty}$ независимо от того, как велико может быть положительное целое число k. Имея k кривых (здесь $k = 6$), гладких и хорошо разделенных (поскольку рассматривается локальная ситуация, где кривые не оборачиваются и не возвращаются), всегда можно найти множество поверхностей, проходящих через них, поэтому число поверхностей в $\mathbb{R}^3$ должно превышать любое конечное число k кривых в $\mathbb{R}^3$

участок прерывистой кривой ничем не будет отличаться от локального участка единственной целостной кривой, свобода которой будет меньше, чем у k отдельных локальных участков кривых.) Конечно же, независимо от того, сколь велико будет конечное число k, мы получим неизмеримо больше свободы, заполнив двумерную поверхность между двумя этими сечениями. Вот и подтверждение факта $\infty^{\infty^2} >> \infty^{k\infty}$ независимо от того, сколь велико может быть конечное число k.

Несмотря на то что я привел доказательство лишь для $\infty^{r\infty^d} >> \infty^{s\infty^f}$, только в случае $r = 1$, $d = 2$, $s = 1$ (обобщено до $s = k$) и $f = 1$ можно построить универсальное доказательство, следуя вышеприведенным рассуждениям, хотя визуализировать именно такой случай мы и не в состоянии. В принципе, нам требуется просто обобщить нашу кривую в $\mathbb{R}^3$ до f-многообразия $\mathbb{R}^{f+k}$, а поверхность в $\mathbb{R}^3$ — до d-многообразия в $\mathbb{R}^{d+r}$, где первый случай соответствует поперечным сечениям k-расслоения на f-многообразии, а второй — поперечным сечениям r-расслоения на d-многообразии. Поскольку $d > f$, второе неизмеримо больше первого независимо от величин r и s.

До сих пор мы рассматривали *свободно выбранные* поля (или поперечные сечения), но из раздела А.2 помним, что в случае с реальными магнитными полями в трехмерном пространстве существует *связь* ($\operatorname{div} \mathbf{B} = 0$). Поэтому наши магнитные поля (связанные таким образом) являются не просто *произвольными* гладкими поперечными сечениями $\mathcal{B}$, а поперечными сечениями, удовлетворяющими этому отношению. Как упоминалось в разделе А.2, фактически это означает, что один из трех компонентов магнитного поля, скажем B_3, будет зависеть от двух других, B_1 и B_2, а также от информации о действиях B_3 на двумерном подмногообразии S трехмерного многообразия $\mathcal{M}$. Что касается функциональной свободы, нас не

слишком интересует происходящее на S (поскольку двумерное S дает нам гораздо меньшую функциональную свободу, чем оставшееся трехмерное $\mathcal{M}$), так что основная функциональная свобода составит $\infty^{2\infty^3}$, и она соответствует свободно выбранным трехмерным многообразиям в пятимерном пространстве $\mathcal{B} = \mathcal{M} \times \mathbb{R}^2$.

Здесь важно обсудить еще одну проблему связи: что будет, если считать $\mathcal{M}$ четырехмерным *пространством-временем*, а не просто трехмерным пространством. В физике обычна такая ситуация: есть *уравнения поля*, описывающие *детерминированную эволюцию* физических полей в пространстве-времени — если учесть в них достаточно подробные данные на конкретный момент времени. В теории относительности, особенно в общей теории относительности Эйнштейна, мы стремимся не трактовать время как универсальный феномен, который обладал бы абсолютными свойствами глобально во всей Вселенной, а стараемся описывать происходящее в контексте некоторой произвольно выбранной координаты времени t. В таком случае некоторое исходное значение t, скажем $t = 0$, даст нам *пространственноподобную* (общепринятый термин, см. раздел 1.7) исходную трехмерную поверхность N, и нужные поля, определенные в N, в таком случае будут уникально описывать магнитные поля во всем четырехмерном пространстве-времени благодаря уравнениям поля. (В общей теории относительности возможны ситуации, в которых могут возникать так называемые *горизонты Коши*, и в таких случаях строгая уникальность немного нарушается, но эти проблемы не важны в рассматриваемом здесь «локальном» контексте.) Также часто возникают связи между полями в рамках исходной трехмерной поверхности, но, как бы то ни было, в обычной физике приходится иметь дело с функциональной свободой, присущей трехмерной поверхности N, и эта свобода равна $\infty^{N\infty^3}$ для некоторого положительного числа N. Число 3 здесь обусловлено размерностью исходной трехмерной поверхности N. Если в некоторой теории, например в теории струн (см. раздел 1.9), функциональная свобода выражается по формуле $\infty^{N'\infty^d}$, где $d > 3$, то требуется очень убедительно объяснить, почему эта избыточная свобода не должна проявляться на уровне физических процессов.

А.9. Комплексные числа

Математические выкладки из раздела А.2—А.8 касались прежде всего проблем классической физики, где физические поля, точечные частицы и само пространство-время описывались в контексте вещественночисленной системы $\mathbb{R}$ (в такой системе координаты, значения силы полей и т. д. обычно считаются вещественными числами). Однако когда в первой четверти XX века появилась квантовая механика, оказалось, что она коренным образом зависит от гораздо более обширной системы $\mathbb{C}$, образованной *комплексными* числами. Тогда выяснилось, как я под-

черкивал в разделе 1.4 и 2.5, что эти комплексные числа описывают устройство известной нам физической реальности в самых микроскопических масштабах.

Что такое комплексные числа? Это числа, для получения которых требуется сделать нечто, на первый взгляд, невозможное: извлечь квадратный корень из отрицательного числа. Как вы помните, квадратный корень из числа a — это число, соответствующее $b^2 = a$. Таким образом, квадратный корень из 4 равен 2, квадратный корень из 9 равен 3, квадратный корень из 16 равен 4, квадратный корень из 25 равен 5, а квадратный корень из 2 равен 1,414213562... и т. д. Можно допустить, что *отрицательные* числа, по модулю равные этим значениям квадратных корней (−2, −3, −4, −5, −1,414213562 ... и т. д.), также могут считаться квадратными корнями, поскольку $(-b)^2 = b^2$. Но если само a окажется отрицательным, то возникает проблема. Ведь независимо от того, положительно или отрицательно b, квадрат его всегда положителен, поэтому сложно себе представить, как вообще можно получить отрицательное число, возведя что-либо в квадрат. Можно свести эту проблему к поиску квадратного корня из −1, поскольку, если бы у нас было число (обозначим его i), удовлетворяющее $i^2 = -1$, то 2i должно было бы удовлетворять $(2i)^2 = -4$, $(3i)^2 = -9$, $(4i)^2 = -16$ и т. д., и вообще $(ib)^2 = -b^2$. Разумеется, как мы уже убедились, чем бы ни было такое i, это не может быть обычное вещественное число. Поэтому оно часто называется *мнимым*, равно как и все его вещественно-численные кратные, например 2i или 3i или −i, −2i и т. д.

Правда, такая терминология обманчива, ведь она подразумевает, что так называемые вещественные числа более «реальны», чем мнимые. Полагаю, такое ощущение обусловлено тем, что меры расстояния и времени в некотором смысле «действительно» являются такими вещественночисленными величинами. Но мы этого не знаем. Мы знаем, что вещественные числа действительно очень удобны для описания времени и расстояния, но не знаем, правомерны ли такие описания в абсолютно *любых* масштабах расстояния или времени. Мы не понимаем, как устроен физический континуум в масштабах, скажем, одной гугольной (см. раздел А.1) метра или секунды. Так называемые вещественные числа — это *математические* конструкты, которые, тем не менее, исключительно ценны при формулировании законов классической физики.

Однако вещественные числа можно считать «реальными» и в платоновском смысле — точно так же, как и любую другую непротиворечивую математическую структуру, если принять распространенную среди математиков точку зрения, согласно которой математическая непротиворечивость есть единственный критерий такой платоновской «экзистенциальности». Правда, так называемые мнимые числа образуют столь же непротиворечивую математическую структуру, как и вещественные, поэтому они не менее «реальны» все в том же платоновском

смысле. Отдельный (и действительно *нерешенный*) вопрос — насколько точно каждая из этих систем моделирует реальный мир.

Комплексные числа — элементы числовой системы $\mathbb{C}$ — получаются путем сложения друг с другом (так называемых) вещественных и мнимых чисел. Таким образом, комплексные числа имеют форму $a + \mathrm{i}b$, где a и b относятся к системе вещественных чисел $\mathbb{R}$. По-видимому, впервые такие числа обнаружил замечательный итальянский врач и математик Джероламо Кардано в 1545 году, а их алгебру описал в 1572 году другой глубоко проницательный итальянец Рафаэлло Бомбелли (см., например, Уайкс [1969]; однако, по-видимому, мнимые числа как таковые принимались в расчет гораздо раньше — возможно, это делал еще Герон Александрийский в I веке н. э.). Впоследствии удалось выявить множество волшебных свойств комплексных чисел, а их полезность с чисто математической точки зрения сегодня не вызывает никаких сомнений. Комплексным числам нашлось разнообразное применение и в физике, например в теории электрических цепей и гидродинамике. Однако вплоть до XX века их считали чисто математическими конструктами, которые удобны при вычислениях, но никак не проявляются в физическом мире.

Однако с появлением квантовой механики ситуация драматически изменилась. Система $\mathbb{C}$ играет ключевую роль в математической формулировке этой теории и проявляется в ней столь же непосредственно, как и различные аспекты $\mathbb{R}$ в классической физике. Фундаментальная квантово-механическая физическая роль $\mathbb{C}$, как было рассказано в разделах 1.4 и 2.5—2.9, зависит от разнообразных примечательных математических свойств комплексных чисел. Как говорилось выше, такое число записывается в форме $x + \mathrm{i}y$, где x и y — вещественные числа (элементы $\mathbb{R}$), а величина i удовлетворяет равенству

$$\mathrm{i}^2 = -1.$$

Обычные алгебраические правила, применимые к вещественночисленным величинам, с тем же успехом применяются и к комплексночисленным. Таким образом, операции сложения и умножения комплексных чисел определяются в категориях вещественночисленных операций в виде

$$\begin{aligned}(x+\mathrm{i}y)+(u+\mathrm{i}v)&=(x+u)+\mathrm{i}(y+v),\\(x+\mathrm{i}y)\times(u+\mathrm{i}v)&=(xu-yv)+\mathrm{i}(xv+yu),\end{aligned}$$

где x, y, u и v — вещественные числа. Кроме того, обратные операции деления и вычитания комплексных чисел (кроме деления на ноль) определяются путем получения отрицательного или обратного комплексного числа следующим образом:

$$-(x+\mathrm{i}y)=(-x)+\mathrm{i}(-y) \text{ и } (x+\mathrm{i}y)^{-1}=\frac{x}{x^2+y^2}-\mathrm{i}\frac{y}{x^2+y^2}=\frac{x-\mathrm{i}y}{x^2+y^2},$$

где x и y — вещественные (и в последнем случае оба ненулевые) числа. Правда, комплексное число часто записывается *единственным* символом; например, можно обозначить $x + \mathrm{i}y$ через z, а $u + \mathrm{i}v$ через w:

$$z = x + \mathrm{i}y \text{ и } w = u + \mathrm{i}v,$$

а потом записывать их сумму как $z + w$, произведение как zw, а отрицательное и обратное число к z просто как $-z$ и z^{-1} соответственно. Разность и частное комплексных чисел теперь определяются просто как $z-w=z+(-w)$ и $z \div w = z \times (w^{-1})$, а $-w$ и w^{-1} даются точно так же, как и в вышеприведенном случае с z.

Оказывается, можно манипулировать комплексными числами так же, как и вещественными, но во многих отношениях эти правила становятся более систематическими, чем в случае с вещественными числами. Важная иллюстрация этого принципа — так называемая фундаментальная алгебраическая переменная, согласно которой любой многочлен единственной переменной z

$$a_0 + a_1 z + a_2 z^2 + a_3 z^3 + \ldots + a_{n-1} z^{n-1} + a_n z^n$$

всегда можно разложить в произведение из n линейных членов. В качестве примера возьмем простые квадратные многочлены $1 - z^2$ и $1 + z^2$. Возможно, разложение первого из них знакомо читателю — при этом используются только вещественночисленные коэффициенты, но для второго случая понадобятся комплексные числа:

$$1-z^2=(1+z)(1-z), \quad 1+z^2=(1+\mathrm{i}z)(1-\mathrm{i}z).$$

На данном конкретном примере мы только начинаем понимать, как комплексные числа систематизируют алгебру, но здесь принцип $\mathrm{i}^2 = -1$ используется весьма прямолинейно, и магия комплексных чисел пока не очевидна. Однако она начинает просматриваться в полной теореме (где, следовательно, можно предположить, что коэффициент a_n ненулевой, и, значит, на эту величину можно делить; можно смело взять a_n равным 1). Таким образом, можно разложить на множители любой (вещественный или комплексный) многочлен:

$$a_0 + a_1 z + a_2 z^2 + \ldots + a_{n-1} z^{n-1} + z^n = (z-b_1)(z-b_2)(z-b_3)\ldots(z-b_n)$$

в контексте комплексных чисел b_1, b_2, b_3... b_n. Обратите внимание: если z принимает любое из значений b_1 ... b_n, то многочлен *обращается в нуль* (поскольку

обнуляется его правая часть). Магия заключается в том, что, просто присовокупив единственное число i к системе вещественных чисел, чтобы удалось решить очень специфическое уравнение $1 + z^2 = 0$, мы выясняем, что совершенно *даром* получаем решения для *всех* нетривиальных многочленных уравнений с одной переменной!

Резюмировав ситуацию иначе, можно сказать, что любые уравнения вида $z^\alpha = \beta$ решаются при условии, что α и β — комплексные числа. В следующем разделе мы рассмотрим еще некоторые примеры комплексночисленной магии (другие подобные примеры см., в частности, в [Nahin, 1998]).

А.10. Комплексная геометрия

Стандартное представление комплексных чисел (впервые оно было описано норвежско-датским исследователем и математиком Каспаром Весселем в отчете от 1787 года, а затем подробно изложено в статье 1799 года) предназначено для обозначения комплексными числами точек в евклидовой плоскости, где отдельное комплексное число $z = x + iy$ соответствует точке в декартовых координатах (x, y) (рис. А.34). Уважая приоритет Весселя, я называю такую плоскость в этой книге *плоскостью Весселя*, несмотря на то что более распространены наименования *плоскость Аргана* или *плоскость Гаусса* (в честь математиков, описавших ее гораздо позже — в работах от 1806 и 1831 года соответственно). Гаусс якобы утверждал, что обдумывал эту идею задолго до публикации (но, очевидно, не в десятилетнем возрасте, когда появился отчет Весселя). По-видимому, источники не сообщают, когда именно до этой идеи впервые дошли Вессель или Арган (см. [Crowe, 1967]). Как у суммы, так и у произведения двух комплексных чисел имеется простая геометрическая характеристика. *Сумма* комплексных чисел w и z вычисляется по известному закону параллелограмма (рис. А.35 *а*, ср. с разделом А.3, рис. А.6), причем линия от 0 до $w + z$ — это диагональ параллелограмма, который образован двумя этими точками и двумя исходными точками $w + z$; *произведение* вычисляется по закону *подобных треугольников* (рис. А.35 *б*), согласно которому треугольник с вершинами в точках 0, 1, w подобен (без отражения) треугольнику, образованному $0, z, wz$. (Здесь также существуют различные вырожденные случаи, где параллелограмм редуцируется до линии, но она требует особого описания.)

Геометрия плоскости Весселя поясняет многие проблемы, которые, как может показаться на первый взгляд, не имеют ничего общего с комплексными числами. Важный пример касается сходимости степенных рядов. Степенной ряд — это выражение

$$a_0 + a_1 z + a_2 z^2 + a_3 z^3 + a_4 z^4 + \ldots,$$

Рис. А.34. В (комплексной) плоскости Весселя $z = x + iy$ представлены как (x, y) в стандартном декартовом виде

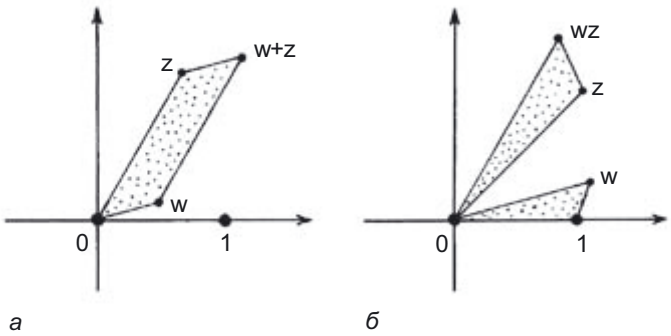


Рис. А.35. Так в плоскости Весселя можно геометрически представить сложение по закону параллелограммов (*а*) и умножение по закону подобных треугольников (*б*)

где $a_0, a_1, a_2, \ldots$ — комплексные константы, в которых (в отличие от многочленов) ряд членов продолжается бесконечно. (На самом деле многочлены попадают под определение степенных рядов, когда мы принимаем все a_r равными нулю, за исключением членов, где r имеет конкретное заданное значение.) При заданном значении z можно обнаружить, что сумма членов *сходится* к некоторому комплексному числу либо что она *расходится*, то есть не сходится. (Такой феномен трактуется как сумма постоянно увеличивающегося набора членов — *частичная сумма* ряда Σ_r, которая может сходиться, а может и не сходиться к конкретному комплексному значению S. Технически *схождение* к S означает, что для сколь угодно малого положительного числа ε найдется некоторое значение r, для которого разность $|S - \Sigma_q|$ будет меньше ε для всех x, если q больше r.)

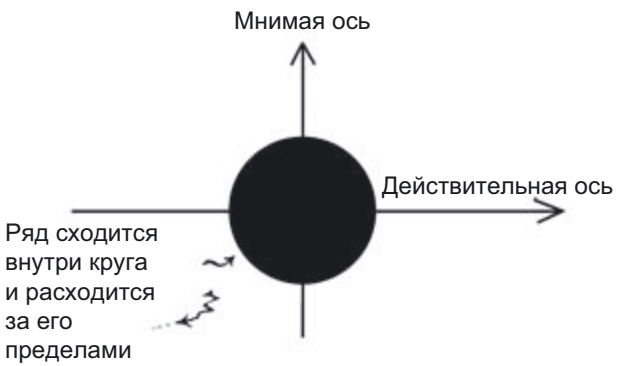


Рис. А.36. Для любого степенного ряда комплексных чисел $A_0 + A_1Z + A_2z^2 + A_3z^3 + + A_4z^4 + ...$ можно построить круг с центром в начале координат плоскости Весселя, который называется *кругом сходимости*, причем ряд сходится для любого z, расположенного строго внутри круга (черная область), и расходится для любого z, расположенного строго за пределами круга (белая область). Таким образом, радиус сходимости (радиус круга) вполне может быть нулевым (ряд сходится только при $z = 0$) или бесконечным (ряд сходится при любых z)

Итак, выясняется примечательная роль плоскости Весселя: если ряд сходится при некоторых (ненулевых) значениях z и расходится при других, то существует *круг* (именуемый «круг сходимости»), центр которого совпадает с началом координат в плоскости Весселя. Этот круг обладает следующим свойством: ряд сходится для любого комплексного числа, расположенного строго в пределах круга, и расходится до бесконечности для любого числа, расположенного строго за пределами этого круга (рис. А.36). Но вопрос о том, что происходит с рядом при значениях, расположенных именно на окружности, уже сложнее.

Этот примечательный результат позволяет разрешить многие загадочные проблемы, к которым иначе не подступиться, например почему ряд $1 - x^2 + x^4 - x^6 + x^8 - ...$ для вещественночисленной переменной x начинает расходиться, как только x оказывается больше 1 или меньше -1, тогда как алгебраическое выражение *суммы* ряда (для $-1 < x < 1$), записываемое как $1/(1 + x^2)$, ничем не примечательно при значениях $x = \pm 1$ (рис. А.37). Проблема возникает при комплексном значении $z = \mathrm{i}$ (или $z = -\mathrm{i}$), где функция $1/(1 + z^2)$ становится бесконечной. Отсюда можно сделать вывод о том, что окружность сходимости должна проходить через точки $z = \pm \mathrm{i}$. Эта окружность также проходит через $z = \pm 1$, поэтому расходимость для вещественных значений x ожидается лишь за пределами круга, где $|x| > 1$ (рис. А.38).

Кроме рассмотренного выше, есть и еще один момент, связанный с расходящимися рядами, который мне хотелось бы отметить. Можно задуматься, а есть

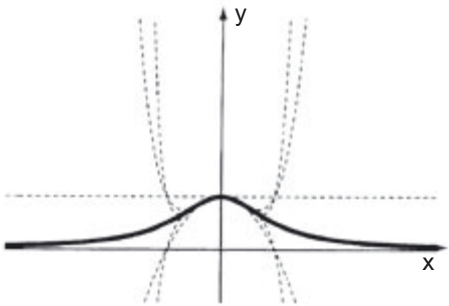


Рис. А.37. Показанная здесь функция вещественного переменного $y = f(x) = 1/(1+x^2)$ (график обозначен темной непрерывной линией) в интервале $-1 < x < 1$ может быть представлена бесконечным рядом $1 - x^2 + x^4 - x^6 + x^8 - x^{10} + ...$, но ряд расходится при $|x| > 1$. Частичные суммы $y = 1$, $y = 1 - x^2$, $y = 1 - x^2 + x^4$, $y = 1 - x^2 + x^4 - x^6$ и $y = 1 - x^2 + x^4 - x^6 + x^8$ представлены прерывистыми линиями, которые показывают точки расхождения. Если судить только по вещественным переменным, то, казалось бы, нет никаких причин, по которым функция вдруг должна расходиться именно в тех точках, где $|x|$ выходит за единицу, поскольку кривая $y = f(x)$ в этих точках ничем не примечательна — здесь она столь же гладкая, как и в других местах, где мы, возможно, хотели бы увидеть начало расхождения

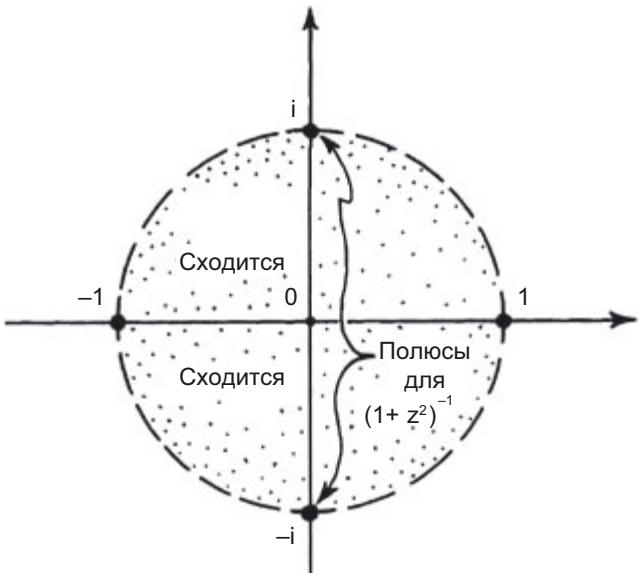


Рис. А.38. В плоскости Весселя видно, какова проблема с $f(x) = 1/(1+x^2)$. В комплексной форме $f(z) = 1/(1 + z^2)$, где $z = x + iy$, мы видим, что функция становится бесконечной на «полюсах» $z = \pm i$, и круг сходимости не может расшириться за пределы этих точек. Следовательно, вещественные ряды для $f(x)$ также должны расходиться на $|x| > 1$

ли смысл присваивать ряду ответ $1/(1 + x^2)$, если x больше 1. В частности, при $x = 2$ имеем

$$1-4+16-64+256-\ldots=\frac{1}{5},$$

и это, разумеется, абсурд, достаточно всего лишь попытаться сложить все члены по порядку. В левой части у нас только целые числа — так почему же в правой части *дробь*? Однако ответ $\frac{1}{5}$ в чем-то «верен», поскольку, если «обозначить сумму» ряда через Σ и приплюсовать к ней 4Σ, то мы, по-видимому, получим

$$\begin{aligned}\Sigma+4\Sigma=1-4+16-64+256-1024+\ldots+\\ +4-16+64-256+1024-\ldots=1,\end{aligned}$$

то есть $5\Sigma = 1$, и мы действительно приходим к ответу $\sum=\frac{1}{5}$. Воспользовавшись схожей аргументацией, можно «доказать» и еще более причудливое тождество (выведенное Леонардом Эйлером в XVIII веке):

$$1+2+3+4+5+6+\ldots=-\frac{1}{12},$$

которое, как ни странно, играет существенную роль в теории струн (см. раздел 3.8 и уравнение (1.3.32) у Полчински [1998]).

Рассуждая логически, можно счесть «шулерством» почленное сложение безнадежно расходящихся рядов для получения таких ответов; однако в них кроется глубокая истина, которая проявляется при так называемом *аналитическом продолжении*. Иногда это может потребоваться для обоснования особых манипуляций с расходящимися рядами, благодаря которым некий диапазон функции, корректно определяемый рядом в одном регионе плоскости Весселя, можно продолжить и в другие регионы, где исходный ряд расходится (рис. А.38). Следует отметить, что при такой процедуре функции захватывают другие точки, но не начало координат. Таким образом, приходится рассмотреть ряд вида $a_0+a_1(z-Q)+a_2(z-Q)^2+a_3(z-Q)^3+\ldots$, чтобы представить расширение функции в окрестностях точки $z = Q$ (см., например, рис. 3.36 в разделе 3.8).

Для аналитического продолжения характерна примечательная *жесткость*, которой обладают голоморфные функции. Их можно как угодно «сгибать», как и гладкие вещественночисленные функции. Полностью детализированная структура голоморфной функции в сколь угодно малом регионе ограничивает возможности этой функции в «отдаленных» пределах. Странным образом голоморфная функция «гнет свою линию», и отклонить ее с этого пути невозможно. Это свойство сыграло важную роль в разделах 3.8 и 4.1.

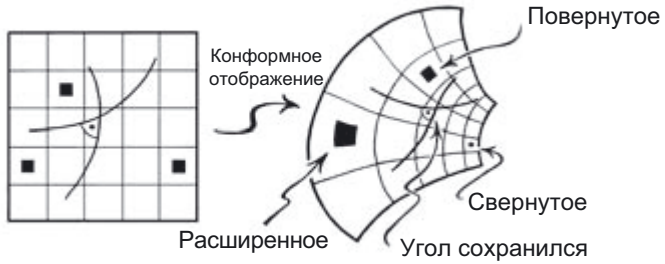


Рис. А.39. Голоморфное отображение одного участка плоскости Весселя на другой имеет две характеристики: оно конформное и без отражения. «Конформное» с геометрической точки зрения означает, что при отображении углы между пересекающимися кривыми не изменяются; аналогично сохраняются бесконечно малые фигуры: они могут оказаться больше, меньше или повернуться, но форма их не изменится в пределе малых размеров

Зачастую бывает полезно подумать о *преобразованиях* в плоскости Весселя. Два простейших таковы: складываем фиксированное комплексное число A с координатой z в плоскости либо умножаем координату z на фиксированное комплексное число B:

$$z \mapsto A + z \text{ или } z \mapsto Bz.$$

Первое из этих действий соответствует *трансляции* плоскости (жесткому движению без вращения), а второе — вращению и/или равномерному расширению/сжатию. Это такие преобразования в плоскости, при которых контуры фигур сохраняются (без отражения), а размеры могут и не сохраняться.

Трансформации (отображения), задаваемые этими функциями, — такие функции называются *голоморфными* — выстраиваются на основе z путем сложения, умножения, с учетом постоянных комплексных чисел и взятием пределов, так что их можно описать в категориях степенных рядов. С геометрической точки зрения для этих функций характерна так называемая *конформность* (и нерефлексивность). У этих отображений есть следующее свойство: бесконечно малые фигуры при преобразовании сохраняются (хотя они и могут вращаться и/или изотропно расширяться или сжиматься). Свойство конформности можно описать и иначе: при таких преобразованиях сохраняются *углы* между кривыми (рис. А.39). Феномен конформной геометрии также очень важен и в высших измерениях (см. разделы 1.15, 3.1, 3.5, 4.1 и 4.3).

Пример *не*голоморфной функции комплексного числа z — это величина $\overline{z}$, определяемая в виде

$$\overline{z} = x - \mathrm{i}y,$$

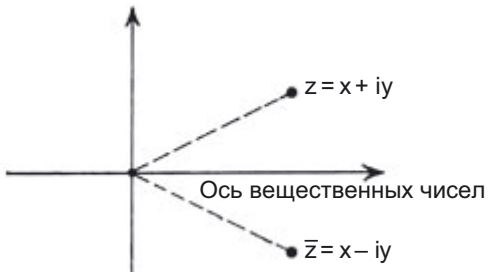


Рис. А.40. Операция комплексного сопряжения ($z \mapsto \bar{z}$) отражения по оси вещественных чисел в плоскости Весселя не является голоморфной. Хотя она и явно конформна, ее ориентация обратна ориентации плоскости Весселя

где $z = x + iy$, при условии, что x и y — вещественные числа. Отображение $z \mapsto \bar{z}$ конформно в том смысле, что малые углы в нем сохраняются, но оно не считается голоморфным, поскольку ориентация меняется на обратную, и отображение определяется *отражением* плоскости Весселя по оси вещественных чисел (рис. А.40). Это пример *анти*голоморфной функции, которая представляет собой комплексное сопряжение голоморфной (см. раздел 1.9). Пусть антиголоморфные функции тоже конформны, они обладают обратной ориентацией в том смысле, что привносят в локальную структуру отражение. Кстати, именно поэтому мы не считаем функцию $\bar{z}$ голоморфной, поскольку в противном случае терялся бы весь смысл. Так, например, вещественная и мнимая части функции z должны были бы считаться голоморфными, так как $x = \frac{1}{2}(z + \bar{z})$ и $y = \frac{1}{2}(z - \bar{z})$. Более того, это также касалось бы величины $|z|$, так называемого *модуля z*, который определяется следующим образом:

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

Отметим, что (по теореме Пифагора) $|z|$ — это просто расстояние от начала координат 0 до точки z в плоскости Весселя. Естественно, отображение $z \mapsto z\bar{z}$ далеко не конформное, поскольку оно расплющивает всю плоскость вдоль неотрицательной части оси вещественных чисел (то есть $z = \bar{z} \geq 0$), поэтому она определенно не голоморфна. Удобно трактовать голоморфную функцию z как такую, которая не включает $\bar{z}$. Соответственно, z^2 голоморфна, а $z\bar{z}$ — нет.

Голоморфные функции — центральный элемент комплексночисленного анализа. Они аналогичны *гладким* функциям при вещественночисленном анализе. Но у комплексночисленного анализа есть важная магическая деталь, отсутствующая у вещественночисленного. Вещественночисленные функции могут быть сколь угодно гладкими. Например, функция $x \times |x|$, равная x^2 при положительном x

и $-x^2$ при отрицательном x, обладает всего одной степенью гладкости (технически, C^1), тогда как функция x^3, график которой выглядит обманчиво просто, обладает бесконечным множеством степеней гладкости (технически, C^∞ или C^ω). Другой пример — функция $x^2 \times |x|$ (x^3 при $x \geq 0$ и $-x^3$ при $x < 0$), которая обладает двумя степенями гладкости (технически, C^2), тогда как у очень похожей на нее функции x^4 этих степеней бесконечно много, и т. д. (рис. А.41). Однако с комплексными функциями все значительно упрощается, поскольку даже минимальная степень комплексной гладкости (C^1) также подразумевает максимальную (C^∞) и, более того, подразумевает, что ее можно расширить в виде степенного ряда (C^ω), так что любая комплексно-гладкая функция автоматически оказывается и голоморфной (подробнее см. Rudin [1986] и в книге ПкР (главы 6 и 7)).

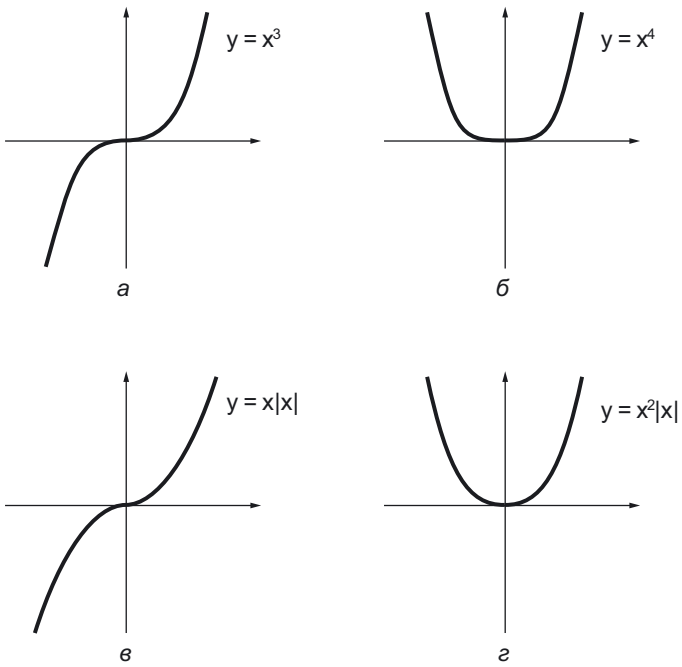


Рис. А.41. Функции действительного переменного могут иметь различный порядок гладкости. Функции $y = x^3$ и $y = x^4$, представленные графиками (*а*) и (*б*), являются бесконечно-гладкими. Такие функции называются аналитическими, обозначаются как принадлежащие множеству $C^\omega(\Omega)$ и допускают расширение в комплексную область. В то же время функция $y = x|x|$, представленная графиком (в) и совпадающая с функцией $y = x^2$ для положительных x и с функцией $y = -x^2$ для отрицательных x, имеет порядок гладкости 1 (обозначается как принадлежащая множеству $C^1(\Omega)$), а функция $y = x^2|x|$, представленная графиком (*г*) и совпадающая с функцией $y = x^3$ для положительных x и с функцией $y = -x^3$ для отрицательных x, имеет порядок гладкости 2, несмотря на то, что визуально графики функций (*в*) и (*г*) очень похожи на графики функций (*а*) и (*б*)

Особенно интересна одна голоморфная функция, а именно *степенная функция* ez (обычно записываемая как exp(*z*)). Мы уже сталкивались с этой функцией для вещественных чисел, когда обсуждали в разделе А.1 ряд

$$e^z = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} + \ldots$$

($n! = 1 \times 2 \times 3 \times \ldots \times n$). На самом деле этот ряд сходится для *всех* значений *z* (то есть у него бесконечный круг сходимости). Если *z* лежит в пределах единичной окружности в плоскости Весселя, а именно в круге, центр которого совпадает с началом координат 0, а радиус равен единице (рис. А.42), то возникает магическая формула (Котса — Муавра — Эйлера)

$$e^{i\theta} = \cos\theta + i\sin\theta,$$

где θ — угол (измеряемый против часовой стрелки), образуемый радиусом до *z* и осью положительных вещественных чисел. Также можно отметить расширение этой формулы с учетом таких точек *z*, которые не обязательно относятся к единичной окружности в плоскости Весселя:

$$z = re^{i\theta} = r\cos\theta + ir\sin\theta,$$

где модуль *z* — это $r = |z|$, как отмечалось выше, а θ называется *аргументом z* (рис. А.42).

Вся теория вещественных многообразий (которую мы вкратце обозначили в разделе А.5) также распространяется и на *комплексные* многообразия, где вещественночисленные координаты из вещественных многообразий меняются на

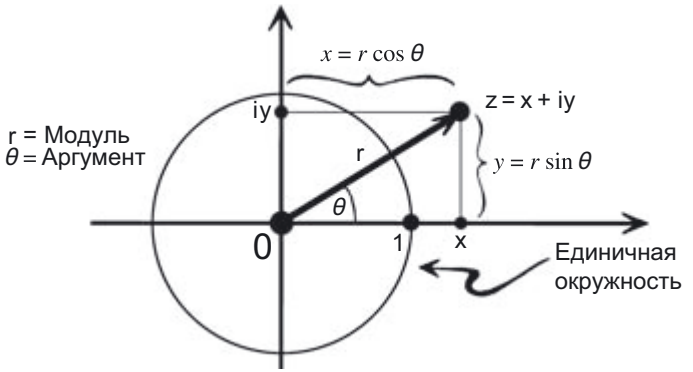


Рис. А.42. Отношение между полярными и декартовыми координатами в плоскости Весселя описывается формулой $z = re^{i\theta} = r(\cos\theta + ir\sin\theta)$. Величина *r* называется модулем, а θ — аргументом комплексного числа *z*

комплексночисленные. Однако всегда остается вариант трактовать комплексное число $z = x + iy$ как пару вещественных чисел (x, y). В таком случае можно повторно выразить комплексное n-многообразие как вещественное $2n$-многообразие (с использованием так называемых *комплексных структур*, обусловленных голоморфными свойствами комплексных координат). Здесь можно отметить, что вещественное многообразие в данном случае переинтерпретируется как комплексное, но число измерений в таком комплексном многообразии обязательно должно быть четным. Тем не менее само по себе это условие далеко не является достаточным, чтобы можно было присвоить настоящему вещественному $2n$-многообразию комплексную структуру, а ведь такое присваивание является необходимым условием подобной переинтерпретации. Такая возможность — редкая удача, особенно когда значения n достаточно велики.

В случае однокомплексномерного многообразия все эти проблемы гораздо понятнее. Если выразить *комплексную кривую* в контексте вещественных чисел, то получаются определенные типы вещественных двумерных поверхностей, так называемые *римановы поверхности*. Если выразить риманову поверхность в вещественных числах, то получится обычная вещественная поверхность, обладающая так называемой *конформной* структурой (как было отмечено выше, конформность связана с определением *угла* между кривыми на поверхности) и *ориентацией* (это попросту означает, что в рамках всей поверхности понятие «против часовой стрелки» не изменяется; см. рис. А.21). Римановы поверхности могут обладать разнообразными топологиями, и некоторые примеры таких топологий показаны на рис. А.13 в разделе А.5. Они играют ключевую роль в теории струн (см. раздел 1.6). Обычно римановы поверхности считаются *замкнутыми*, то есть не имеющими границ, но можно считать, что в таких поверхностях есть *отверстия*, или *проколы* (см. рис. 1.44), играющие определенную роль в *теории струн* (см. раздел 1.6).

Для нас особенно важна простейшая риманова поверхность с топологией обычной сферы, так называемая *риманова сфера*, которая, как отмечалось в разделе 2.7, играет особую роль в контексте квантово-механического спина. Можно легко построить риманову сферу, просто добавив к целой плоскости Весселя одну смежную точку (которую можно обозначить ∞). Чтобы убедиться в том, что *вся* риманова сфера может считаться подлинным (однокомплексномерным) комплексным многообразием, можно накрыть эту сферу двумя координатными лоскутами, причем один из них соответствует исходной плоскости Весселя, где координаты откладываются от z, а другой — копии плоскости Весселя, где координаты откладываются от $w(=z^{-1})$. Теперь сюда включается наша новая точка $z = \infty$, служащая просто началом координат w, а вот точка начала координат z не включается. Две эти плоскости Весселя объединяются лоскутом $z = w^{-1}$, и получается целая риманова сфера (рис. А.43).

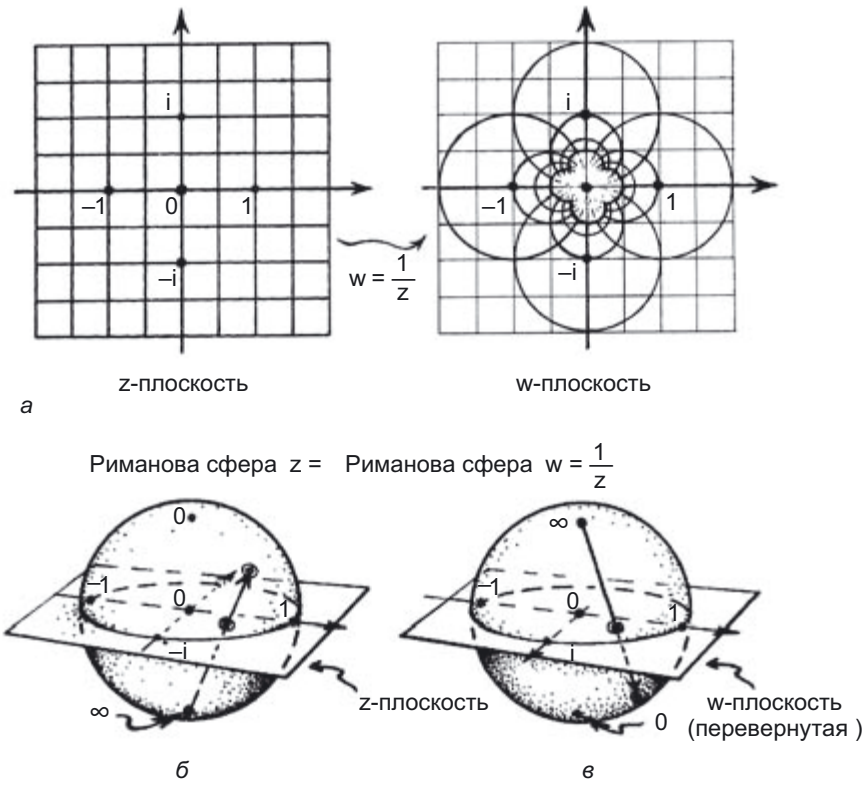


Рис. А.43. Риманова сфера — это многообразие, которое можно собрать из двух координатных лоскутов, каждый из которых копирует плоскость Весселя, в данном случае z-плоскость и w-плоскость, относящиеся друг к другу как $w = z^{-1}$: *а*) линии постоянных частей z (вещественной и мнимой), отображенные на w-плоскость; *б*) стереографическая проекция, сделанная с южного полюса римановой сферы, дает z-плоскость; *в*) стереографическая проекция, сделанная с северного полюса римановой сферы, дает w-плоскость, то есть перевернутое отображение

А.11. Гармонический анализ

Часто для решения уравнений, связанных с физическими задачами, физики берут на вооружение мощную процедуру гармонического анализа. Как правило, в таких случаях речь идет о *дифференциальных* уравнениях (зачастую относящихся к типу *частичных* дифференциальных уравнений, пример которого ($\operatorname{div} \mathbf{B} = 0$) упомянут в разделе в А.2). Дифференциальные уравнения решаются при помощи *дифференциального исчисления*, а я намеренно не затрагивал в книге подробностей этой темы, и здесь лишь в самых общих чертах опишу основные алгебраические свойства дифференциальных операторов.

Что же это за операция — дифференцирование? Для функции $f(x)$ с *одной* переменной такая операция (обозначим ее D), примененная к f, заменяет ее на новую функцию $f'(x)$, так называемую *производную* f, значение которой $f'(x)$ в точке x соответствует уклону исходной функции f в точке x, поэтому можно записать $Df = f'$ (рис. А.44). Также можно рассмотреть *вторую* производную f'' функции f, значение которой $f''(x)$ в точке x характеризует уклон f' в точке x. Оказывается, таким образом можно измерить, насколько исходная функция f «изгибается» в точке x (и так мы измерили бы *ускорение*, если бы величина x соответствовала времени). Можно записать

$$f'' = D\left(Df\right) = D^2 f$$

и т. д., чтобы получить k-ю производную $D^k f$ от f для любого положительного целого числа k. *Обратная* операция применительно к D (иногда обозначаемая D^{-1}, а чаще — символом интеграла $\int$) приводит нас к *интегральному анализу площадей и объемов*.

Когда переменных больше ($u, v, ...$) и они могут быть (локальными) координатами в n-мерном пространстве, производная может применяться к каждой координате отдельно. Производную u можно обозначить как D_u (она будет называться *частичной*, когда все остальные переменные не изменяются), производную v — как D_v и т. д. Опять же, ее можно возводить в разные степени, а также складывать в различных комбинациях. В качестве показательного примера можно привести один хорошо изученный дифференциальный оператор — так называемый *лапласиан* (впервые примененный выдающимся французским математиком Пьером-Симоном де Лапласом в конце XVIII века и описанный в его классическом труде «Небесная механика» [Laplace, 1829—1839]). Обычно лапласиан обозначается через ∇^2 (или Δ), и в трехмерном евклидовом пространстве с декартовыми координатами u, v, w он описывается формулой:

$$\nabla^2 = D_u^2 + D_v^2 + D_w^2.$$

Это означает, что, применительно к некоторой функции f (с тремя переменными u, v, w) величина $\Delta^2 f$ представляет собой сумму вторых производных f относительно u, относительно v и относительно w:

$$\nabla^2 f = D_u^2 f + D_v^2 f + D_w^2 f.$$

Уравнения с ∇^2 широчайшим образом применяются как в физике, так и в математике, начиная с лапласовского $\nabla^2 \varphi = 0$, при помощи которого он описывал ньютоновское гравитационное поле в контексте скалярной величины, именуемой *потенциалом* φ гравитационного поля. (Вектор, описывающий силу и на-

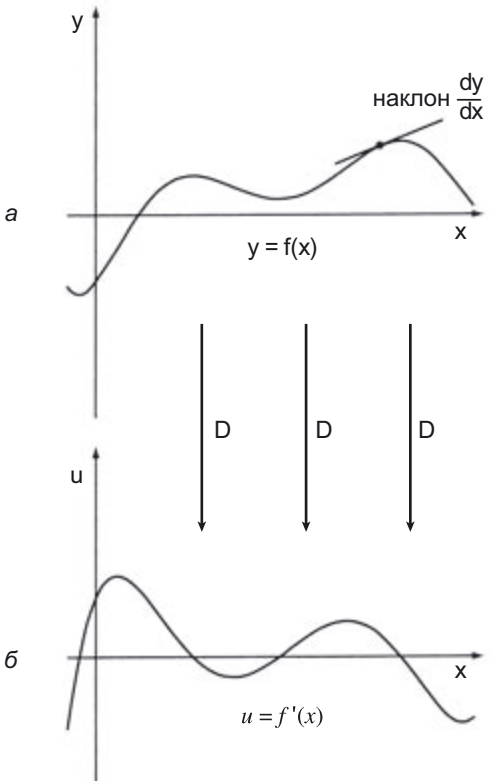


Рис. А.44. Операция дифференцирования, обозначенная здесь через D, меняет функцию $f(x)$ на новую функцию $f'(x)$, где значение $f'(x)$ для каждого x — это наклон $f(x)$ в точке x. Обратная операция интегрирования, связанная с областями под нижней кривой, будет описываться при помощи стрелок с противоположным направлением

правление гравитационного поля, имел бы три следующих компонента: $-D_u\varphi$, $-D_r\varphi$ и $-D_w\varphi$.) Другой важный пример связан с двумерным евклидовым пространством (декартовы координаты x и y), где оператор равен $\nabla^2 = D_x^2 + D_y^2$, а плоскость считается плоскостью Весселя для комплексного числа $z = x + iy$. Далее находим, что любая голоморфная функция ψ от z (см. раздел А.10) имеет вещественную и мнимую части f и g:

$$\psi = f + ig,$$

причем каждая из частей удовлетворяет уравнению Лапласа:

$$\nabla^2 f = 0, \ \nabla^2 g = 0 .$$

Тождество Лапласа — пример *линейного* дифференциального уравнения; это означает, что если у нас есть любое из двух решений: $\nabla^2\phi = 0$ и $\nabla^2\chi = 0$, то любая линейная комбинация

$$\lambda = A\phi + B\chi,$$

где A и B — константы, также будет иметь решение

$$\nabla^2\lambda = 0.$$

Хотя линейность в целом не характерна для дифференциальных уравнений, оказывается, что дифференциальные уравнения в самом деле играют ключевую роль в теоретической физике. Выше я уже приводил пример с ньютоновской теорией тяготения, выраженной в контексте лапласовой потенциальной функции φ. Другими важными примерами такого рода являются максвелловы линейные уравнения электромагнитного поля (см. разделы 1.2, 1.6, 1.8, 2.6 и 4.1), а также основное шрёдингеровское уравнение квантовой механики (см. разделы 2.4—2.7 и 2.11).

При работе с такими линейными уравнениями гармонический анализ обладает огромной силой. Термин «гармонический» взят из музыки: музыкальные звуки можно анализировать на уровне различных «чистых тонов». Например, скрипичная струна может вибрировать по-разному. У *основного* тона есть конкретная частота v, при которой вся струна вибрирует наиболее простым образом (без узлов), но струна также может вибрировать и с различными *гармониками* с частотами $2v$, $3v$, $4v$, $5v$ и т. д., где контуры вибрирующей струны (с 1 узлом, 2 узлами, 3 узлами, 4 узлами и т. д.) соответствуют по форме контурам волны, дающей чистый гармонический фон (рис. А.45). Базовое дифференциальное уравнение, описывающее вибрацию струны, является линейным, поэтому общее вибрационное состояние можно выстроить в виде линейной комбинации таких *мод*, то есть отдельных чистых тонов вибрации (то есть сначала фундаментальных, в затем — со всеми гармониками). Чтобы дать общее решение дифференциального уравнения для струны, нужно просто указать последовательность чисел, где каждое число в известном смысле будет соответствовать магнитуде вклада каждой моды. Любая форма волны при условии, что периодичность этой волны соответствует частоте основного тона, в таком случае уникально выражается в виде суммы синусоидальных компонентов (термин *синусоидальная* означает характерный контур синусоиды, присущей функции $y = \sin x$, как на рис. А.46). Такое представление периодической функции в виде гармоник называется *анализ Фурье* в честь французского математика Жозефа Фурье, впервые исследовавшего такое представление периодических волновых колебаний в контексте синусоидальных гармоник. Ниже в этом разделе я покажу красивый способ построения таких представлений.

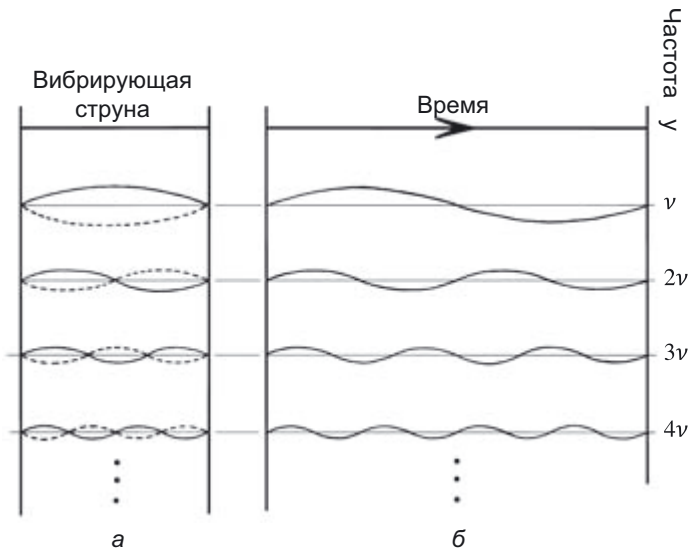


Рис. А.45. Различные моды вибрации (скрипичной) струны: *а*) форма каждой моды вибрации самой струны; *б*) свойства вибрации во времени, когда частота является неким целочисленным кратным основной частоты v

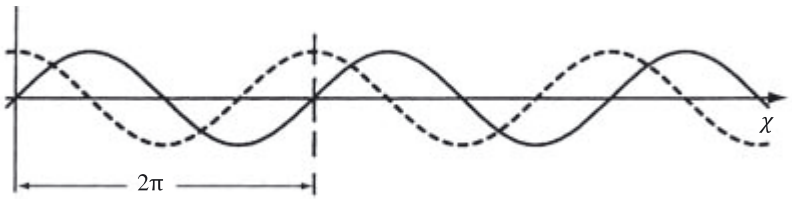


Рис. А.46. Сплошная кривая — график функции $\sin x$; пунктирная— график функции $\cos x$

Такая общая процедура универсально применима к линейным дифференциальным уравнениям, причем отдельные моды в данном случае — это особые простые решения уравнения, которые легко получить и в терминах которых можно выразить все прочие решения: для этого берем линейные комбинации мод (обычно линейные). Рассмотрим конкретный случай: лапласово уравнение в двумерном евклидовом пространстве. Это особенно простой случай, поскольку можно прибегнуть непосредственно к алгебре и анализу комплексных чисел, где необходимые нам моды выписываются напрямую. Не хочется вводить читателя в заблуждение; в более общих ситуациях так быстро управиться не получается. Тем не менее основные моменты, которые я хочу продемонстрировать, отлично иллюстрируются на этом примере, если воспользоваться комплексночисленными описаниями.

Как упоминалось выше, любое решение лапласовского уравнения $\nabla^2 f = 0$ в двух измерениях можно трактовать как вещественную часть f голоморфной функции ψ (или, что равноценно, как мнимую часть; выбор не имеет значения, поскольку мнимая часть ψ — это вещественная часть слегка иной функции $-i\psi$). Можно выразить общее решение нашего дифференциального уравнения $\nabla^2 f = 0$ как линейную комбинацию основных *мод* (это аналоги различных гармоник скрипичной струны), а для того, чтобы определить таким образом эти моды, можно перейти к соответствующим голоморфным величинам ψ. Поскольку это голоморфные функции комплексного числа z, их можно выразить как *степенные ряды*:

$$\psi = a_0 + a_1 z + a_1 z^2 + a_1 z^3 + \ldots,$$

где $z = x + iy$, и если все время брать вещественную часть, то можно выразить f в терминах x и y. Отдельные моды будут различными членами в степенных рядах, то есть вещественными и мнимыми частями отдельных степеней:

$$z^k = (x + iy)^k$$

(умноженными на некоторое подходящее постоянное число, зависящее от k, — и теперь нам нужна не только вещественная часть, но и мнимая, поскольку коэффициенты здесь комплексные). Следовательно, в терминах x и y мода будет состоять из вещественной и мнимой частей этого выражения (например, $x^3 - 3xy^2$ и $3x^2y - y^3$, где $k = 3$).

Для того чтобы точнее обрисовать ситуацию, нужно понимать, какой регион плоскости нас интересует. Сначала предположим, что этот регион — вся плоскость Весселя, поэтому нас интересуют решения уравнения Лапласа, накрывающие всю эту плоскость. В случае голоморфной функции ψ нам понадобится степенной ряд с *бесконечным* радиусом сходимости; конкретный пример такого рода — степенная функция e^z. В таком случае коэффициенты $1/k!$ стремятся к нулю, по мере того как k стремится к бесконечности, и таким образом обеспечивается сходимость для всех z данного степенного ряда (пример A):

$$e^z = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} + \ldots,$$

который мы уже рассматривали в разделе А.10 (и в разделе А.1). С другой стороны, в разделе А.10 мы рассматривали и второй ряд (пример B):

$$(1 + z^2)^{-1} = 1 - z^2 + z^4 - z^6 + z^8 - \ldots,$$

сходящийся в пределах единичной окружности $|z| = 1$ и расходящийся за ее пределами. Промежуточный случай (пример C):

$$\left(1+\frac{z^2}{4}\right)^{-1} = 1-\frac{z^2}{4}+\frac{z^4}{16}-\frac{z^6}{64}+\frac{z^8}{256}-\ldots$$

сходится в пределах круга $|z| = 2$.

Следовательно, хотя мы и можем представлять наши решения лапласовских уравнений просто как последовательности коэффициентов, то есть (1, 1, 1/2, 1/6, 1/24, ...) для примера A, (1, 0, –1, 0, 1, 0, –1, 0, 1, ...) для примера B и (1, 0, –1/4, 0, 1/16, 0, –1/64, 0, 1/256, ...) для примера C, необходимо внимательно присматриваться к свойствам этих чисел по мере движения последовательностей к бесконечности, для того чтобы знать, на самом ли деле конкретная последовательность является решением нашего дифференциального уравнения в интересующем нас регионе. Крайний пример такого рода возникает в случае, если данный регион оказывается римановой сферой (см. раздел A.10) и получается путем добавления единственной дополнительной точки ∞ к плоскости Весселя. Существует теорема, согласно которой в глобальном масштабе на римановой сфере существуют лишь такие голоморфные функции, которые по факту являются *константами*, поэтому все последовательности чисел, сответствующие решениям лапласовских уравнений, на самом деле имеют форму (K, 0, 0, 0, 0, 0, 0, ...)!

Эти примеры иллюстрируют и другой аспект гармонического анализа. Часто требуется указать *граничные значения* для решений дифференциальных уравнений. Например, нам может понадобиться найти решение лапласовского уравнения $\nabla^2 f = 0$ в n-мерном евклидовом пространстве, которое соблюдается как в пределах единичной $(n-1)$-сферы S, так и на поверхности этой сферы. На самом деле существует следующая теорема [Evans, 2010; Strauss, 1992]: если задать f как произвольную вещественночисленную функцию на S (допустим, она будет гладкой), то найдется единственное решение $\nabla^2 f = 0$ *внутри* S, которое достигает заданных значений *на* S. Теперь можно задуматься о том, что происходит с каждой отдельной модой при гармоническом разложении решений уравнения Лапласа.

Сначала давайте исследуем случай $n = 2$ и будем считать S единичной окружностью в плоскости Весселя. Будем искать решения лапласовских уравнений на единичном диске. Если рассмотреть *моду*, определяемую конкретной степенью z^k, то, воспользовавшись полярным представлением z из раздела A.10, а именно

$$z = re^{i\theta} = r\cos\theta + ir\sin\theta,$$

получим на единичной окружности S ($r = 1$):

$$z^k = e^{ik\theta} = \cos k\theta + i\sin k\theta.$$

Вокруг единичной окружности для каждой такой моды вещественная и мнимая части z будут изменяться по синусоиде, как и рассмотренная выше k-я гармоника, которую можно извлечь из скрипичной струны (то есть $\cos k\theta$ и $\sin k\theta$), где координата θ теперь отсчитывает время и может бесконечно описывать круг с течением времени (см. рис. А.45). Для *общего* решения лапласовского уравнения на этом единичном диске значение f (функция угловой координаты θ) является произвольным, поскольку периодичность определяется по кругу, то есть период равен 2π. (Разумеется, аналогичным образом можно рассмотреть и любую другую периодичность, просто откалибровав величину окружности — уменьшая или увеличивая ее по желанию.) Это просто вышеупомянутое разложение периодической функции по Фурье, причем функция связана с вибрирующей скрипичной струной.

Выше я рассматривал значение f в окрестностях граничной окружности S таким образом, словно оно меняется, подобно гладкой функции, но на самом деле такая процедура гораздо универсальнее. Например, даже в вышеприведенном примере B граничная функция далеко не гладкая. Она обращается в сингулярность в двух точках $\theta = \pm\pi/2$, соответствующих $\pm i$ в плоскости Весселя. С другой стороны, в примере C (а на самом деле и в примере A), если рассмотреть лишь часть решения на единичном диске, то выяснится, что функция f является совершенно гладкой на ограничивающей единичной окружности S. Однако минимальные требования для граничной функции f нас здесь не интересуют.

В высших измерениях ($n > 2$) применим такой же анализ. Решения уравнения Лапласа в недрах гиперсферы S — (n–1)-мерной сферы — можно разделить на гармоники, которые, как и в случае двух измерений, соответствуют различным степеням радиальной координаты r. В случае $n = 3$ S — это обычная двумерная сфера, и хотя простое описание в контексте комплексных функций здесь не сработает, мы по-прежнему можем отличать наши «моды» друг от друга по их зависимости от степени k, в которую вводится радиальная координата r. На каждой сфере принято откладывать координаты θ, ϕ с центром в начале координат ($r = R$, где R *константа*). Такие координаты именуются *сферическими полярными*, они очень похожи на координаты широты и долготы на земном шаре. Детали для нас не существенны, но они показаны на рис. А.47.

Обычные моды имеют форму

$$r^k Y_{k,m}(\theta,\ \phi),$$

где $Y_{k,m}(\theta,\phi)$ — *сферические гармоники* (предложенные Лапласом в 1782 году), то есть особые явные функции θ и ϕ, подробные графики которых нас не интересуют [Riley et al., 2006]. Значение k (в стандартной нотации обычно используется l) пробегает все целые неотрицательные значения: $k = 0, 1, 2, 3, 4, 5, ...$, а m может

принимать любые целые значения (в том числе и отрицательные), такие, что $|m| \le k$. Соответственно, допустимые значения для (k, m) будут следующими:

$$(0,0),(1,-1),(1,0),(1,1),(2,-2),(2,-1),$$
$$(2,0),(2,1),(2,2),(3,-3),(3,-2),\ldots$$

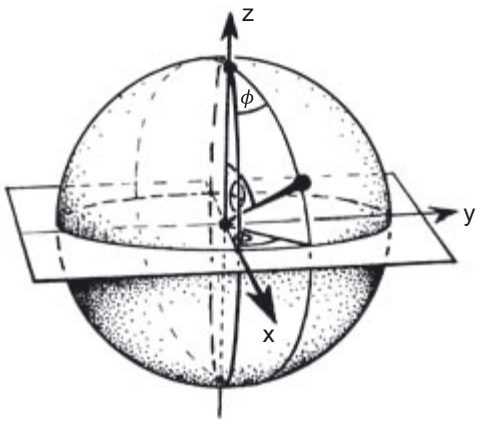


Рис. А.47. Условные полярные углы θ и ϕ на сфере S^2, стандартным образом внедренной в $\mathbb{R}^3$.

Для того чтобы указать конкретное решение лапласовского уравнения в твердом шаре, содержащемся в S (то есть $1 \ge r \ge 0$), потребуется узнать вклад каждой из этих мод — бесконечную последовательность чисел

$$f_{0,0}, f_{1,-1}, f_{1,0}, f_{1,1}, f_{2,-2}, f_{2,-1}, f_{2,0}, f_{2,1}, f_{2,2}, f_{3,-3}, f_{3,-2}, \ldots$$

сообщающую точный вклад каждой. Такая последовательность указывает f на граничной сфере S или, что эквивалентно, соответствующее решение лапласовского уравнения в недрах S. (Проблемы непрерывности/гладкости, касающиеся f на S, приводят нас к сложным вопросам о том, как последовательность $f_{k,m}$ продолжается до бесконечности.)

Здесь я хотел отметить, что при всем потенциале этих методов при изучении отдельных решений, особенно что касается вычислений, они затушевывают важную проблему, а именно проблему *функциональной свободы*, которая особенно интересовала нас в разделах А.2 и А.8, а также играет ключевую роль в дискуссии, развернутой в главе 1. Если указывать решения дифференциальных уравнений таким образом, работая, например, с уравнением Лапласа или другими, более

сложными системами, то гармонический анализ в итоге приводит к решению в виде бесконечной последовательности чисел. Сама размерность того пространства, в котором определяется решение, не говоря уже о его размере и форме, зачастую скрывается за каким-нибудь сложным асимптотическим свойством этой последовательности, и проблему функциональной свободы можно совершенно упустить из виду.

Даже в простейшем примере с вибрирующей струной — речь о ситуации со скрипичной струной, о которой говорилось выше в этой главе, — оказывается, что если действовать невнимательно, даже простой анализ мод может привести к ошибочным выводам о функциональной свободе. Рассмотрим две разные ситуации: в первом случае струна может вибрировать лишь в одной плоскости, как скрипичная струна, по которой аккуратно проводят смычком. Во втором случае, например при пиццикато, при вибрации струна может сдвигаться в двух направлениях в стороны от линии натяжения струны (здесь я не учитываю смещения *вдоль* самой струны, которые можно было бы вызвать, поглаживая струну пальцами). Моды вибрации струны можно разделить на те, что происходят в двух плоскостях, перпендикулярных направлению струны, а все остальные вибрации можно было бы считать состоящими из вибраций в двух этих плоскостях (рис. А.48). Поскольку обе плоскости равноценны друг другу, мы просто получим в каждой из плоскостей одни и те же моды с одинаковой частотой вибрации. Следовательно, единственное различие между вибрациями, возникающими при игре смычком (это вибрации в одной плоскости) и при пиццикато (неограниченные вибрации), заключается в том, что в последнем случае каждая мода успевает произойти дважды. В первом случае функциональная свобода равна $\infty^{2\infty^1}$, а во втором случае она значительно больше и составляет $\infty^{4\infty^1}$.

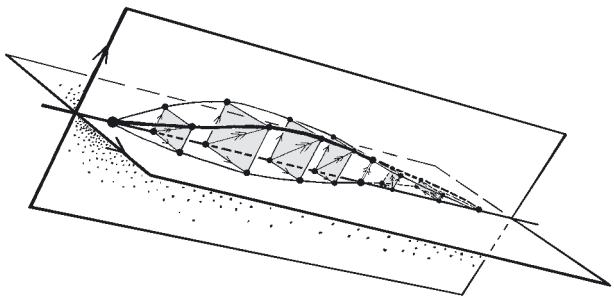


Рис. А.48. Небольшие вибрации струны в трех измерениях могут разрешаться в вибрации в двух ортогональных плоскостях, причем в каждой точке струны вектор смещения разрешается в два компонента, расположенных в этих плоскостях под прямым углом друг к другу

Значения 2 и 4 связаны с магнитудой и скоростью сдвига в каждой точке струны, причем во втором случае такая величина должна быть вдвое больше. Верхняя единица означает, что струна одномерна, и показателем степени здесь было бы большее число, если бы мы заменили струну на «*n*-брану» (см. раздел 1.15) — браны играют важную роль в модной ныне теории струн. О том, насколько важна проблема функциональной свободы, рассказано в главе 1.

В данном случае также интересно рассмотреть вибрации двумерной поверхности, например *барабана*. Такие явления часто рассматриваются в контексте анализа мод, причем основные способы вибрации барабана можно выразить как вклад различных составляющих в это движение от различных мод и записать в виде бесконечной последовательности чисел $p_0, p_1, p_2, p_3, ...$, где отдельное число означает вклад отдельной моды. На первый взгляд эта запись может показаться весьма похожей на вибрации скрипичной струны при игре смычком, которые описываются последовательностью q_0, q_1, q_2, q_3, ..., числа из которой соответствуют вкладу каждой отдельной моды вибрации струны. Однако в случае со смещениями поверхности барабана функциональная свобода будет неизмеримо выше — конкретно $\infty^{2\infty^2}$ — по сравнению с приведенной выше функциональной свободой струны $\infty^{2\infty^1}$. Можно оценить эту разницу, если представить себе *квадратный* барабан, покрытый сеткой декартовых координат (x, y), где все x и y лежат между 0 и 1. В таком случае можно было бы (весьма оригинально) попытаться представить смещения поверхности барабана в виде таких мод, которые являются произведениями $F_{ij}(x,y) = g_i(x)h_j(y)$ мод $g_i(x)$, возникающих по оси x, и мод $h_j(y)$, возникающих по оси y. Такой анализ мод позволил бы описать общие смещения поверхности барабана в контексте последовательности чисел $f_{0,0}, f_{0,1}, f_{1,0}, f_{0,2}, f_{1,1}, f_{2,0}, f_{0,3}, f_{2,1}, f_{1,2}$ и т. д., означающих магнитуду вклада от каждой $F_{ij}(x,y)$. Такая процедура вполне допустима, но она не позволяет непосредственно оценить огромную разницу между функциональной свободой ∞^{∞^2} при смещениях кожи барабана в двух измерениях и гораздо меньшей функциональной свободой ∞^{∞^1} при каждом из одномерных смещений координат x и y соответственно (или функциональной свободой $\infty^{2\infty^1}$), которая соответствовала бы смещениям x вместе со смещениями y, давая нам в итоге гораздо меньшую свободу при «произведениях смещений» вида $g(x)h(y)$.