А.М. ЯГЛОМ, И.М. ЯГЛОМ

ВЕРОЯТНОСТЬ И ИНФОРМАЦИЯ

А. М. ЯГЛОМ и И. М. ЯГЛОМ

ВЕРОЯТНОСТЬ И ИНФОРМАЦИЯ

ИЗДАНИЕ ТРЕТЬЕ,
ПЕРЕРАБОТАННОЕ И ДОПОЛНЕННОЕ

ИЗДАТЕЛЬСТВО «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ

Москва 1973

517.8
Я 29
УДК 519.21

Вероятность и информация. А. М. Я г л о м и И. М. Я г л о м, Главная редакция физико-математической литературы издательства «Наука», 1973.

Книга является общедоступным введением в новую область математики — *теорию информации*, тесно связанную с кибернетикой и имеющую ряд приложений в технике связи, лингвистике, биологии и т. д. В третьем издании подвергся тщательному просмотру весь текст и внесены многочисленные улучшения в изложение. Данные о теоретико-информационных характеристиках конкретных видов сообщений (письменная и устная речь, фототелеграммы, телевидение и пр.) пополнены результатами, полученными в разных странах на протяжении 60-х годов нашего века. В качестве одного из примеров, иллюстрирующих общее понятие «линии (или канала) связи», рассмотрена «генетическая линия связи» и отвечающий ей «генетический код». Книга пополнена двумя новыми параграфами, один из которых дает представление о *теории кодирования* — большом направлении, выделившемся из теории информации и сегодня иногда рассматриваемом как самостоятельная научная дисциплина.

Для чтения книги достаточно математической подготовки в объеме школьного курса. Книга рассчитана на студентов вузов и втузов (а частично — даже и на учащихся старших классов средней школы), преподавателей средней и высшей школы, инженеров-связистов, специалистов в области физики, биологии, лингвистики.

Я $\frac{0223—1772}{042\ (02)-73}$ 78-72

ОГЛАВЛЕНИЕ

ИЗ ПРЕДИСЛОВИЯ К ПЕРВОМУ ИЗДАНИЮ

За долгие годы сложилось такое положение, когда почти никакие сведения об интенсивной научной работе, ведущейся в области теоретической математики, не проникают за рамки узкого круга математиков-профессионалов; это обстоятельство вызывает даже иногда у неспециалистов совершенно неправильное представление об определенной «завершенности» математики, делающей исследовательскую работу в этой области почти невозможной или, во всяком случае, очень трудной. Причина такого положения кроется в том, что подавляющее большинство работ, печатающихся в математических журналах, относится к достаточно развитым разделам этой науки, с которыми трудно ознакомить лиц, не имеющих специальной подготовки; что же касается более элементарных частей математики, вроде элементарной геометрии, то трудно рассчитывать, чтобы за многовековую историю науки здесь были не замечены какие-либо факты или теоремы, имеющие действительно большое принципиальное значение [1]). Естественно, что и новые большие направления, возникшие в математике за последние десятилетия, как правило, оперируют с достаточно сложными понятиями и представлениями, мало доступными для популяризации. Тем более значительной представляется заслуга замечательного американского математика и инженера Клода Шеннона, который в 1947—1948 гг. сумел указать новую важную область математики, истоки которой связаны с совсем элементарными соображениями.

---

[1]) Однако даже в этих начальных разделах математики остаются нерешенными некоторые серьезные вопросы и появляются иногда интересные и глубокие работы (см., например, брошюру В. Г. Болтянского «Равновеликие и равносоставленные фигуры», М., Гостехиздат, 1956, излагающую, в основном, исследования последних лет).

Основные задачи, которые ставил перед собой Шеннон при создании того направления, которое в последующие годы получило название «теория информации», были связаны с чисто техническими вопросами электросвязи и радиосвязи [1]). Вообще говоря, новые применения математики в технике и естествознании обычно бывают связаны с использованием сложного математического аппарата и, кроме того, чаще всего не могут быть объяснены без глубокого проникновения в суть запутанных проблем современной науки и техники; поэтому возможности популяризации практических достижений математики сегодняшнего дня также являются весьма скромными. Именно поэтому представления неспециалистов о прикладном значении математики зачастую ограничиваются заимствованными из школьного курса сведениями о том, что геометрия еще в древнем Египте использовалась для восстановления границ земельных участков после разливов Нила, и некоторыми другими того же рода. И в этом отношении изложение круга идей, связанных с теорией информации, представляется крайне заманчивым, так как простейшие практические приложения этих идей к современным техническим вопросам вполне могут быть объяснены читателям, обладающим минимальной математической и технической подготовкой.

Настоящая книжка, рассчитанная на широкий круг читателей (для понимания всего ее содержания достаточно знакомства с математикой в объеме курса средней школы), разумеется, ни в какой мере не претендует на то, чтобы служить хотя бы только элементарным введением в теорию информации как научную дисциплину. Мы могли дать здесь лишь поверхностное представление о важных практических приложениях этой теории; также и глу-

---

[1]) Благодаря своему общему характеру работы Шеннона оказали большое стимулирующее влияние на все исследования, относящиеся к передаче и сохранению какой бы то ни было информации в природе и технике; линиями, по которым передается эта информация, могут являться не только телеграфные и телефонные провода или среда, передающая радиосигналы, но и нервы, по которым передаются сигналы от органов чувств к мозгу и от мозга к мускулам, или те почти совсем еще не исследованные пути, какими передаются от зародышевой клетки указания о дальнейшем плане построения живого организма.

бокие чисто математические проблемы, связанные с теорией информации, никак не могли быть здесь раскрыты. Основная цель, которую поставили перед собой авторы, гораздо проще — она состоит в том, чтобы ознакомить читателя с некоторыми несложными, но весьма важными, новыми математическими понятиями и на примере этих понятий показать один из возможных путей использования математических методов в современной технике.

Первая глава книги посвящена разъяснению старого (введенного еще в XVII веке) понятия *вероятности*, знакомство с которым необходимо для понимания всего дальнейшего содержания. Во второй главе рассматриваются введенные Шенноном понятия *энтропии* и *информации*, общетеоретическое значение которых было оценено математиками лишь в самое последнее время. Третья и четвертая главы посвящены примерам и приложениям; в отличие от первых двух глав строгие доказательства приводимых утверждений здесь зачастую лишь намечены или вовсе опущены, а в некоторых местах и сами утверждения сформулированы лишь в форме весьма правдоподобных предположений. При этом в третьей главе польза понятий энтропии и информации иллюстрируется на примерах с загаданными числами, фальшивыми монетами и т. п., кое в чем напоминающих те «игрушечные» задачи с игральными костями и картами, на которых в XVII в. зародилась теория вероятностей; более содержательные приложения технического порядка сосредоточены в четвертой главе. Мы рассчитываем, что ознакомление с третьей главой поможет читателю лучше почувствовать смысл основных понятий, введенных в главе II, и тем самым подготовиться к изучению наиболее сложной четвертой главы, использующей к тому же некоторые результаты третьей.

Книга предназначена для всех любителей математики и в первую очередь для тех, кто ее в настоящее время преподает или изучает; наряду с этим мы рассчитываем, что она может быть небезынтересной и для многих читателей, имеющих по своей специальности дело с техникой связи, но не обладающих солидной математической подготовкой. В основу книги положена лекция, прочитанная одним из авторов московским школьникам — участникам школьного математического кружка при Московском

государственном университете; содержание лекции здесь значительно расширено.

Авторы выражают искреннюю признательность А. Н. Колмогорову, ценные советы которого способствовали значительному улучшению книги. Они благодарны также редактору книги М. М. Горичей, замечания которой помогли устранить некоторые дефекты изложения.

А. М. Яглом,
И. М. Яглом

Москва, май 1956 г.

ИЗ ПРЕДИСЛОВИЯ КО ВТОРОМУ ИЗДАНИЮ

Второе издание книги «Вероятность и информация» по структуре почти не отличается от первого издания; читатель, вздумавший сравнить оглавления двух изданий книги, отметит, что различия здесь весьма незначительны. Не изменился также и характер книги, предполагающей у читателя весьма скромные математические знания (недостаток которых, впрочем, должен компенсироваться известной настойчивостью). При всем том частные различия между двумя изданиями настолько значительны, что смело можно говорить о новой книге.

Столь большие изменения частично связаны с тем, что эта книга посвящена очень молодой и бурно развивающейся отрасли науки, для которой два года, прошедшие со дня выхода в свет первого издания,— это большой срок. Кое-что стало за эти два года яснее авторам книги; многое удалось почерпнуть из многочисленных новых книг и статей — в последнее время количественный рост литературы по теории информации происходит с нарастающей интенсивностью. Но особенно необходимой стала переработка первого издания из-за одного просчета авторов.

Эта книга родилась из лекции, прочитанной московским школьникам,— и авторы твердо запомнили ее происхождение, на которое читатели, по-видимому, обратили мало внимания. Соответственно этому в предисловии к книге было указано, что она «предназначается для всех любителей математики и в первую очередь для тех, кто ее в настоящее время преподает или изучает». При этом

мы, однако, просмотрели еще одну, весьма многочисленную категорию читателей — лиц, серьезно интересующихся именно теорией информации (а не математикой вообще), но не желающих начинать ее изучение со специальной литературы, овладение которой требует и времени и труда. Больше всего замечаний об этой книге мы получили от математиков и от инженеров-связистов — и наши уверения, что книга не рассчитана ни на тот, ни на другой круг читателей, не производили на них никакого впечатления. Удивившая авторов быстрота, с которой первое издание книги исчезло из магазинов, появление переводов на несколько иностранных языков (венгерский, немецкий, французский, японский) — все это вынуждало считать, что книга ответила какой-то насущной потребности и заставило внимательно продумать вопрос о том, как эту потребность лучше удовлетворить.

Мы и теперь склонны полагать, что наша книга мало подходит для изучения предмета специалистами по теории вероятностей или по теории связи — первым естественно рекомендовать небольшую, но тщательно написанную книгу А. Файнстейна [5] [1]); для читателей же второй категории наиболее подходящей книгой явится, по-видимому, интересная книга Ф. М. Вудворда [4]. Также и физикам или биологам, заинтересовавшимся идеями Шеннона, естественно обратиться не к нашей книге, а к книгам Л. Бриллюэна [2] (физика) и У. Р. Эшби [11] (биология). Однако всем этим категориям читателей, возможно, будет небезынтересно ознакомиться и с настоящей, значительно более элементарной книгой. Лишь только филологам, которые уже на сегодняшний день представляют собой довольно значительную группу «потребителей» теории информации, нам нечего порекомендовать, что заставило нас при подготовке второго издания книги отнестись с большим вниманием к их возможным запросам. И если при подготовке нового издания мы по-прежнему отвергали любой материал, включение которого повысило бы уровень математической подготовки, необходимый для чтения книги, то при этом теперь уже имелись в виду не только учащиеся средней школы, но

---

[1]) Цифры в квадратных скобках указывают номера в списке литературы в конце книги.

и биологи или филологи, но знакомые с высшей математикой.

Новая точка зрения на круг читателей книги обусловила ряд существенных изменений в ее тексте. Так, например, из нового издания исчезли русские буквы *Э* (энтропия) и *И* (информация), которые, быть может, и облегчали чтение книги некоторым совсем неопытным читателям, но зато, наверно, были неудобны для всех тех, кто имел (или пожелал бы иметь в дальнейшем) дело также и с другой литературой по этому предмету, использующей иные обозначения. Естественно также было уже в главе II уделить достаточно внимания статистическому толкованию понятия энтропии, делающему его столь плодотворным для всех практических приложений теории информации. Заметно расширена последняя глава книги, имеющая наибольшее прикладное значение; объем книги увеличился также за счет напечатанного мелким шрифтом (и могущего быть опущенным при первом чтении) материала, где, учитывая интересы математиков, мы, в частности, привели строгие доказательства некоторых предложений, лишь сформулированных в основном тексте. Изменился и характер иллюстрирующих текст задач: в новом издании реже встречаются упражнения на урновую схему и математические развлечения, зато чаще — вопросы, в которых реально может быть использована теория информации. Однако мы не стали менять принятую в первом издании терминологию, в некоторых случаях отличающуюся от используемой в научной литературе заменой специальных терминов более «обыкновенными» словами (например «линия связи» вместо «канала связи», «энтропия опыта» вместо «энтропии распределения вероятностей» и т. п.). Мы сохранили также в книге целую главу, специально посвященную «задачам на смекалку», поскольку, по существу, в этих задачах в новой (и довольно привлекательной) форме рассматриваются достаточно серьезные вопросы, непосредственно связанные с задачами наиболее экономной передачи сообщений. Эту связь, которую, как оказалось, просмотрели некоторые из читателей первого издания книги, мы теперь осветили несколько подробнее, чем раньше.

К новому изданию книги приложена библиография, отсутствующая в первом издании. Убедившись (в част-

ности, и на опыте нашей работы над книгой) в удобстве, которое представляет для любых расчетов, связанных с теорией информации, наличие таблицы значений функции — $p \log p$ (где $0 \leqslant p \leqslant 1$), мы поместили такую таблицу, заимствованную из сборника [46], в качестве третьего приложения к книге. Мы сохранили в этой таблице двоичную систему логарифмов; в книге, однако, используются более привычные большинству читателей десятичные логарифмы (тем более, что нам хотелось разрушить имеющееся у некоторых инженеров представление о том, что основой теории информации является использование именно двоичных логарифмов).

Наконец, самым значительным изменением является добавление к главе IV специального § 3, содержащего сводку данных об информации, содержащейся в конкретных типах сообщений (письменная и устная речь, музыка, телевизионные и фототелеграфные изображения); в конце этого параграфа кратко указаны также некоторые данные о пропускной способности различных линий связи. Этот параграф является самым большим в книге; он мало связан с последующим текстом и вполне может быть опущен читателем, интересующимся лишь математической стороной теории информации. Нам, однако, кажется, что значительно больше будет таких читателей, для которых этот параграф окажется как раз наиболее интересным. По своему характеру § 3 гл. IV несколько отличается от остальной книги — фактически он представляет собой обзор большого числа сравнительно специальных работ, опубликованных за последние годы в различных научных и научно-технических журналах. Для удобства читателей, специально интересующихся той или иной областью приложений теории информации, мы во всех случаях точно указали источники, содержащие более подробное изложение упоминаемых нами результатов (основная часть приложенной к книге библиографии относится именно к этому параграфу), и постарались сделать наш обзор по возможности более полным (в той мере, в какой это было возможно без уменьшения степени элементарности книги). Однако надо иметь в виду, что при той интенсивности, с которой в настоящее время во всем мире ведется работа по изучению статистических свойств сообщений и линий связи, можно опасаться, что

уже к моменту выхода книги в свет приведенный в ней обзор не сможет претендовать на полноту, а еще через некоторое время собранные в нем данные и вовсе устареют. Нам тем не менее кажется, что и тогда § 3 гл. IV не будет бесполезным: ведь основная его цель — дать представление о порядке величин количества информации, встречающихся в науке и технике, и проиллюстрировать общее направление вдохновленных теорией информации технических, филологических и биологических исследований, а вовсе не служить основой для дальнейшей научной работы специалистов.

В заключение нам хочется искренне поблагодарить всех читателей нашей книги, поделившихся с нами своими соображениями, способствовавшими улучшению нового издания, в том числе С. Г. Гиндикина, А. Н. Колмогорова, В. И. Левенштейна, П. С. Новикова, И. А. Овсеевича, С. М. Рытова, В. А. Успенского, Г. А. Шестопал, М. И. Эйдельманта и особенно — Р. Л. Добрушина и А. А. Харкевича. Мы признательны также В. А. Гарманну, Л. Р. Зиндеру, Д. С. Лебедеву и Т. Н. Молошной за полезные беседы, которые мы имели с ними по вопросам, связанным с содержанием § 3 гл. IV книги.

А. М. Яглом,
И. М. Яглом

Москва, март 1959 г.

ПРЕДИСЛОВИЕ К ТРЕТЬЕМУ ИЗДАНИЮ

Первое издание настоящей книги вышло в свет в 1957 г., а второе — в 1960 г.; настоящее же третье издание по времени отстоит от второго на 13 лет. В таком большом перерыве между изданиями виноваты в первую очередь мы сами. Второе издание этой книги давно превратилось в библиографическую редкость, запросы же от читателей продолжали поступать — и издательство неоднократно обращалось к нам с предложением об ее переиздании; мы, однако, никак не могли на это решиться. Нам было ясно, что книгу нельзя оставить в том виде, который она имела во втором издании, ибо необходимо было как-то откликнуться на существенные изменения, происшедшие за эти годы в теории информации; коренная же перера-

ботка книги (сопровождаемая даже изменением ее названия, как многие нам советовали) явно требовала слишком большого труда и была нам, пожалуй, не под силу.

В конце концов мы пошли по тому пути, который почти всегда избирают люди, поставленные в затруднительное положение,— по пути компромисса. Настоящее, третье издание книги сохранило прежнее название и многое из первоначального ее облика; так, например, мы по-прежнему не предполагаем у читателя никаких знаний, выходящих за пределы школьного курса математики. Таким образом, книга эта все еще остается более простой, чем все имеющиеся учебные и монографические изложения теории информации. Мы, однако, не могли игнорировать и то обстоятельство, что, к нашему удивлению, второе издание «Вероятности и информации» как в нашей стране, так и за рубежом в ряде случаев использовалось в качестве основного пособия при чтении курсов лекций в высшей школе — и при переработке и пополнении текста стремились сделать книгу более подходящей для такого, ранее не предвиденного нами, ее употребления. В частности, мы отказались, наконец, от использования в книге десятичных логарифмов и непривычных десятичных единиц измерения количества информации (дитов), уничтожив тем самым последнее прямое свидетельство происхождения этой книги из лекции, прочитанной много лет назад учащимся средней школы [1]).

Наибольшей переработке подверглась последняя глава книги, являющаяся в ней самой важной, так как фактически главы I—III представляют собой лишь введение в основное содержание книги, сосредоточенное в главе IV. Имея в виду читателей, желающих ознакомиться по книге с основами математической теории информации, мы включили в § 2 гл. IV изложение оптимальных кодов Хафмана (более важных теоретически, чем рассматривавшиеся и в предыдущих изданиях коды Шеннона — Фано) и существенно переработали доказательство основной теоремы

[1]) В литературе, обращенной к школьникам, использование двоичных логарифмов производит впечатление некоторой вычурности; однако в книге по теории информации, рассчитанной на более опытных читателей, такое впечатление, напротив, может вызвать употребление десятичных логарифмов вместо общеупотребительных двоичных.

о кодировании при отсутствии помех, сделав его более кратким и математически четким; еще более изменен § 4, где, в частности, приведены два новых доказательства основной теоремы о кодировании при наличии помех вместе с простым доказательством обратной теоремы о кодировании. Той же цели служит и включение в первую главу книги закона больших чисел, позволившее сделать некоторые последующие выводы более строгими, а также заметное увеличение числа ссылок на серьезную научную литературу, к изучению которой естественно перейти после ознакомления с нашей книгой.

Однако наиболее существенным обстоятельством, которое нужно было учесть при подготовке книги к переизданию, было то, что за последние два десятилетия сама проблематика теории информации существенно изменилась: в настоящее время теория информации — это, в первую очередь, *теория кодирования*, бурное развитие которой невозможно было даже предсказать в период подготовки к печати предыдущего издания. Поэтому сегодня даже популярная книга по теории информации, полностью игнорирующая то ее направление, которое вызывает наибольший интерес и у теоретиков, и у инженеров-практиков, и на котором сосредоточена львиная доля усилий специалистов по теории информации во всем мире, представляется в чем-то неуместной. С другой стороны, общий характер теории кодирования и математические средства и методы, применяемые в этой важной и изящной области прикладной математики, очень существенно отличаются от основного содержания нашей книги; переориентация книги в сторону теории кодирования вызвала бы необходимость всю ее переписать заново [1]). Поэтому мы и здесь ограничились полумерами: добавленный к главе IV совсем новый заключительный параграф дает лишь самое первое представление о задачах и методах теории кодирования; с другой стороны, и в своем настоящем виде этот параграф заметно отличается от остального содержания книги. Это различие побудило нас пополнить книгу новым Приложением II, посвященным некоторым чисто алгебраическим понятиям

[1]) Мы этого не сделали — однако не можем не выразить сожаления по поводу отсутствия в русской учебной и научно-популярной литературе доступной для начинающих книги по теории кодирования.

и предложениям (но зато мы смогли исключить имевшееся в старых изданиях Приложение II, ставшее ненужным после внесения некоторого усовершенствования в изложение теоремы о кодировании при отсутствии помех). Строго говоря, новое Приложение II не необходимо для понимания содержания § 5 гл. IV, посвященного теории кодирования; однако читатель, просмотревший его до ознакомления с содержанием указанного параграфа, будет, по-видимому, лучше представлять себе возможности дальнейшего развития и обобщения результатов этого параграфа.

Особое место в книге занимает § 3 последней главы — об этом достаточно подробно говорилось в предисловии ко второму изданию. Содержащаяся в нем сводка данных, касающихся конкретных типов сообщений, является единственной известной нам сводкой такого рода в литературе — последнее обстоятельство побудило нас постараться расширить и этот параграф, включив в него обзор большого числа более новых работ. Разумеется, несмотря на существенное увеличение относящейся сюда библиографии, мы не можем претендовать на то, что охватили в с ю литературу по рассматриваемым темам — бесспорно в ней упущено большое число работ, разбросанных по огромному числу журналов самого разного профиля. Мы должны также предупредить читателя, что проверка имеющихся в отдельных исследованиях числовых данных и анализ степени их статистической достоверности никак не входили в наши задачи — в этом отношении, как нам кажется, вообще очень многое еще только предстоит сделать. Однако несмотря на то, что не все приведенные в § 3 данные вызывают полное доверие, включение всего этого материала в книгу является оправданным — оно позволяет читателю получить достаточно полное представление о достигнутых к настоящему времени результатах в области конкретно-информационных исследований и об общей направленности ведущихся здесь работ.

Разумеется, большое число связанных с теорией информации направлений оказалось совсем не затронутым в нашей книге. Помимо естественной невозможности «объять необъятное», последнее отчасти связано со стремлением в какой-то мере сохранить в настоящем издании

тот облик, который имела эта книга раньше. Так, например, мы по-прежнему почти полностью игнорируем в ней задачи, связанные с оценками энтропии и информации опытов с *бесконечным* множеством возможных исходов (по поводу относящихся сюда общих понятий и определений см., например, [17]). Мы совсем не касаемся также так называемого «алгоритмического» подхода к понятию количества информации (о нем см., например, важные работы [15] и [16]) и лишь совсем вкратце упоминаем (в § 3 гл. IV) о комбинаторном определении соответствующих понятий. Наконец, целиком вне рамок этой книги остаются все, пока еще сугубо предварительные, попытки расширительного толкования понятия информации, выходящего за рамки теории Шеннона (типа «семантической информации» или «тезаруса»; см., например, [18]—[19а]).

Главным достоинством предисловий является, как известно, то, что в них можно выразить благодарность всем, кто помог авторам в их работе. А. Н. Колмогоров любезно предоставил нам свою рукопись, на основе которой было составлено описание принадлежащего ему существенного уточнения шенноновского метода определения энтропии письменного текста при помощи опытов по отгадыванию; некоторые относящиеся сюда материалы передал нам также А. В. Прохоров. В. В. Иванов, И. А. Овсеевич, Н. В. Петрова, Б. С. Цыбаков и В. Эндрес (Дармштадт, ФРГ) обратили наше внимание на некоторые литературные источники, использованные при пополнении § 3 гл. IV. На содержании ряда мест книги отразились наши многочисленные беседы с Р. Л. Добрушиным на темы теории информации. Редактор третьего издания С. З. Стамблер внимательно прочел весь текст и способствовал его улучшению; он также передал нам большой список дополнительной литературы, частично использованный в работе над книгой. Нам приятно выразить всем перечисленным здесь лицам нашу искреннюю признательность.

А. М. Яглом,
И. М. Яглом

Москва, май 1972 г.

Глава I

ВЕРОЯТНОСТЬ

§ 1. Определение вероятности. Случайные события и случайные величины

На практике очень часто приходится сталкиваться с опытами (иначе — испытаниями, наблюдениями, процессами), могущими давать различные результаты в зависимости от обстоятельств, которых мы не знаем или не умеем учесть. Так, например, при бросании игральной кости (однородного кубика, грани которого занумерованы цифрами от 1 до 6) мы не можем знать заранее, какая из граней окажется сверху, так как это зависит от очень многих неизвестных нам обстоятельств (деталей движения руки, бросающей кость, положения игральной кости в момент броска, особенностей поверхности, на которую падает кость и т. д.). Нельзя также предсказать заранее, сколько выпускников средней школы подаст в определенный год заявления в тот или в иной институт, сколько бракованных изделий окажется в выпущенной партии или сколько дождливых дней будет в будущем году; нельзя знать, сколько ошибок сделает школьник в предстоящей ему контрольной работе или на какой билет выпадет главный выигрыш в предстоящем тираже лотереи (номера выигравших билетов определяются путем извлечения нескольких бумажек с номерами из сосуда, в котором лежат тщательно перемешанные бумажки с номерами всех билетов) и т. д. Число подобных примеров, разумеется, можно было бы значительно увеличить.

Применение математики к изучению явлений такого рода опирается на то, что во многих случаях при многократном повторении одного и того же опыта в одних и тех же условиях *частота появления рассматриваемого результата* (т. е. отношение числа опытов, в которых этот результат наблюдался, к общему числу производимых опытов) *остается все время примерно одинаковой*, близкой к

некоторому постоянному числу p. Так, например, известно, что частота попадания в цель для данного стрелка в данных условиях стрельбы, как правило, почти всегда бывает примерно одинаковой, лишь изредка уклоняясь сколько-нибудь значительно от некоторой средней цифры (с течением времени эта средняя цифра может, разумеется, измениться — в таких случаях говорят, что стрелок совершенствуется в стрельбе, или, наоборот, разучивается стрелять). Также и частота выпадения шестерки на игральной кости или процент бракованных изделий при данных условиях производства обыкновенно мало меняются при массовом повторении соответствующих «опытов» (бросания кости или изготовления данных изделий). Исходя из этого заключают, что в каждом случае существует определенное постоянное число, объективно характеризующее сам процесс стрельбы, бросания кости, производства изделий и т. д., около которого и колеблется все время (не отклоняясь от него сколько-нибудь значительно) средняя частота соответствующего результата (попадания в цель, выпадения шестерки, появления бракованного изделия) в длинном ряду «опытов». Это постоянное число называют *вероятностью* рассматриваемого события. Аналогично определяют вероятность и в ряде других вопросов, относящихся к самым различным областям математики, механики, физики, техники, биологии. Наука, изучающая свойства вероятностей и применения этого понятия, называется *теорией вероятностей*.

Согласно сказанному выше вероятность того или иного события можно приближенно оценить по результатам длинной серии опытов. Однако само существование вероятности, разумеется, нисколько не зависит от того, производим ли мы опыты или нет. В связи с этим возникает весьма естественный вопрос о методах, позволяющих находить вероятности различных событий без предварительного проведения соответствующих опытов; владея такими методами, мы можем заранее делать определенные предсказания о результатах последующих опытов, что открывает большие возможности для естественнонаучных применений понятия вероятности. Мы не будем здесь рассматривать этот вопрос во всей полноте, а ограничимся лишь одним простейшим примером, к которому,

однако, можно свести сравнительно широкий круг задач об определении вероятности [1]).

Пусть мы имеем какой-то ящик (или, как чаще говорят в подобных случаях, урну), в котором лежат тщательно перемешанные 10 шаров, отличающиеся друг от друга только цветом: 5 шаров белых, 3 черных и 2 красных. Вытащим не глядя из урны один шар; спрашивается, какова вероятность, что он будет того или другого цвета? Совершенно ясно, что здесь мы имеем 5 шансов из 10 вытянуть белый шар, 3 из 10 — вытянуть черный шар и 2 из 10 — вытянуть красный шар; другими словами, вероятности вытянуть белый, черный и красный шар равны соответственно $\frac{5}{10} = \frac{1}{2}$, $\frac{3}{10}$ и $\frac{2}{10} = \frac{1}{5}$. И действительно, если мы попробуем много раз осуществить соответствующий опыт (каждый раз после осуществления опыта возвращая вынутый шар снова в урну и тщательно перемешивая все шары), мы убедимся, что примерно в 50% всех извлечений будет вынут белый шар, в 30% — черный шар и в 20% — красный шар. Естественно, что столь же просто решается задача о нахождении вероятностей и при любом другом количестве перемешанных шаров различных цветов в урне.

Рассмотрим теперь еще несколько задач на определение вероятности, сводящихся к той же «задаче об урне».

З а д а ч а 1. *Какова вероятность того, что при бросании наугад монеты сверху окажется герб*?

Очевидно, что эта задача равносильна следующей. Пусть мы имеем урну с двумя шарами, на одном из которых написано «герб», а на втором — «цифра» (разумеется, вместо надписанных шаров можно рассматривать шары двух разных цветов, например, белого и черного). Какова

---

[1]) Читателю, желающему более основательно ознакомиться с теорией вероятностей и с путями ее применения к естествознанию и технике, можно порекомендовать рассчитанную на широкого читателя книжку Б. В. Гнеденко и А. Я. Хинчина [25] или заметно бо́льшую по объему, но также доступно составленную книгу Ф. Мостеллера, Р. Рурке и Дж. Томаса [26] (к последней книге приложен обстоятельный список литературы, сопровождаемый краткими аннотациями указанных в нем книг и статей). См. также несколько более трудные статьи А. Н. Колмогорова [33] и М. Каца [34] и другую литературу, указанную в списке литературы в конце настоящей книги.

вероятность, что при извлечении наудачу из урны одного шара мы вынем шар с надписью «герб»? Ясно, что искомая вероятность здесь равна $\frac{1}{2}$.

Задача 2. *Какова вероятность того, что при бросании игральной кости у нас выпадет число очков, делящееся на 3?*

Вместо бросания кости мы можем говорить об извлечении из урны одного из имеющихся там шести шаров, занумерованных цифрами 1, 2, 3, 4, 5 и 6. Если теперь закрасить черной краской 3-й и 6-й шар, оставив остальные шары белыми, то мы придем к задаче об определении вероятности извлечения черного шара (числа 3 и 6 делятся на три, а остальные — не делятся). Очевидно, что искомая вероятность здесь равна $\frac{2}{6} = \frac{1}{3}$.

Задача 3. *Известно, что на студенческом вечере присутствуют двести студентов из одного института, двести пятьдесят — из второго и триста — из третьего. Какова вероятность того, что студент, с которым Вы случайно заговорили, учится во втором институте?*

Очевидно, что эта задача равносильна следующей. Пусть мы имеем урну с 750 шарами; 200 из этих шаров — белые, 250 — черные и 300 — красные. Какова вероятность того, что при извлечении наудачу одного шара из урны мы вынем черный шар. Ясно, что эта вероятность равна $\frac{250}{750} = \frac{1}{3}$.

Постараемся теперь уловить общие принципы решения всех этих задач. В разобранном перед задачами примере условие, что шары в урне тщательно перемешаны и вынимаются не глядя, означает, что мы с равными основаниями можем ожидать появления любого из заключенных в урне шаров или, другими словами, что извлечения всех шаров р а в н о в е р о я т н ы. А так как шаров всего у нас было 10, то естественно заключить, что для каждого из имеющихся шаров вероятность его извлечения равна $\frac{1}{10}$. Далее, белых шаров у нас имеется пять; поэтому вероятность вынуть белый шар равна $\frac{5}{10}$.

Совершенно такие же рассуждения приводили к ответу в каждой из трех других задач. Так, например,

в случае с бросанием игральной кости мы считали, что равновероятны выпадения любой из шести граней куба; именно поэтому мы могли заменить задачу о бросании кости задачей об извлечении из урны одного из шести шаров. Но из шести граней имеются ровно две такие, что их выпадение удовлетворяет условиям задачи; вероятность выпадения какой-нибудь одной из этих двух граней равна $\frac{2}{6}$.

Если предположить, что рассматриваемый опыт (извлечение шара из урны, бросание монеты или игральной кости, разговор с одним из участников студенческого вечера и т. д.) может иметь n равновероятных исходов, то вероятность каждого из этих исходов следует считать равной $\frac{1}{n}$. Рассмотрим теперь какое-либо событие (извлечение белого шара из урны, выпадение «герба» при бросании монеты или четной цифры при бросании игральной кости, разговор со студентом, учащимся во втором институте и т. п.), определяемое результатами опыта. Если это событие осуществляется при m из n возможных равновероятных исходов опыта и не осуществляется при остальных $n - m$ исходах, то вероятность его принимается равной $\frac{m}{n}$. Другими словами, *вероятность некоторого события равна отношению числа равновероятных исходов, благоприятных для данного события, к общему числу равновероятных исходов.* Набранное курсивом предложение можно принять за о п р е д е л е н и е понятия вероятности; при этом равновероятность отдельных исходов должна оговариваться в описании производимого опыта (именно эту цель преследует указание на то, что игральная кость имеет строго форму куба и изготовлена из однородного материала, или что шары перемешаны и не отличаются ничем, кроме цвета). Хотя такое определение и не охватывает некоторые важные случаи вычисления вероятностей (см., например, статьи [33] и [34], книги [27], [28] и [29], а также напечатанный мелким шрифтом § 5 этой главы), для нас оно в большинстве случаев будет достаточным.

Условимся теперь о терминологии, которую мы будем далее применять. События, которые могут произойти или

не произойти в результате произведенного опыта, мы будем называть *случайными событиями*; в том же смысле мы будем говорить об *исходах* данного опыта. Случайные события мы будем обозначать большими латинскими буквами, а вероятности случайных событий (или определенных исходов опыта) буквой p; вероятность события A часто будет записываться как $p(A)$. Значительную роль у нас будут играть опыты, которые могут иметь *несколько* различных исходов; в таком случае мы будем обозначать все эти исходы одной буквой с разными номерами (а сами опыты — чаще всего греческими буквами).

Каждому опыту такого рода отвечает определенная *таблица вероятностей*:

исходы опыта	A_1	A_2	...	A_k
вероятности	$p(A_1)$	$p(A_2)$	...	$p(A_k)$

Так, например, опыту, рассматриваемому в разобранном на стр. 19 примере, отвечает таблица

A_1	A_2	A_3
$\frac{1}{2}$	$\frac{3}{10}$	$\frac{1}{5}$

(здесь A_1 — извлечение белого шара, A_2 — черного и A_3 — красного), а опыту, рассматриваемому в задаче 1,— простая таблица

B_1	B_2
$\frac{1}{2}$	$\frac{1}{2}$

(здесь B_1 — выпадение «герба», а B_2 — выпадение «цифры»); бросание игральной кости связано со следующей таблицей вероятностей:

выпавшее число очков	1	2	3	4	5	6
вероятности	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

и т. д.

Следует отметить одно существенное отличие последней таблицы от первых двух. Здесь результаты опыта можно записать с помощью определенных чисел (1, 2, 3, 4, 5 и 6) — возможность, которой мы не имели в предшествующих примерах. В этом случае мы можем сказать, что число очков, выпадающих при бросании кости, является случайной величиной, могущей принимать одно из шести возможных значений в зависимости от случая (т. е. в зависимости от не поддающихся учету обстоятельств). Другими примерами случайных величин могут служить число бракованных изделий на сотню, число рождений в каком-либо городе за год, число очков, выбиваемых каким-либо стрелком при определенных условиях стрельбы с одного выстрела (мишень, на которой указаны числа очков, засчитываемых при попадании в каждую из ее частей, изображена на рис. 1) и т. д. [1]).

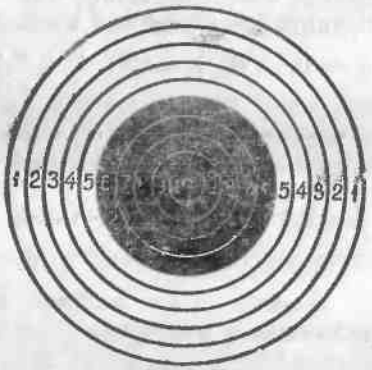


Рис. 1.

Само название «случайная величина» обязывает нас как-то оценивать ее значение. Нетрудно понять, как следует это сделать. Рассмотрим, например, первую из перечисленных выше случайных величин (число бракованных изделий на сотню); пусть это число в определенных условиях производства не превосходит 6, причем соответствующая таблица вероятностей имеет вид:

число бракованных изделий	0	1	2	3	4	5	6
вероятности	0,1	0,15	0,2	0,25	0,15	0,1	0,05

[1]) Понятие случайной величины стоит в стороне от основной темы этой книжки, но в теории вероятностей оно является одним из центральных. По этому поводу см., например, вторую часть книги Б. В. Гнеденко и А. Я. Хинчина [25].

В таком случае из большого числа N сотен изделий примерно $0{,}1N$ не будут содержать бракованных изделий, $0{,}15N$ будут содержать по одному бракованному изделию, $0{,}2N$ — по два, $0{,}25N$ — по три, $0{,}15N$ — по четыре, $0{,}1N$ — по пяти и $0{,}05N$ — по шести бракованных изделий. Следовательно, при большом N общее число a бракованных изделий можно считать равным

$$a = 0{,}1N\cdot 0 + 0{,}15N\cdot 1 + 0{,}2N\cdot 2 + 0{,}25N\cdot 3 + \\ + 0{,}15N\cdot 4 + 0{,}1N\cdot 5 + 0{,}05N\cdot 6$$

и значит, среднее значение числа бракованных изделий на сотню (средний процент брака) будет равно

$$\frac{a}{N} = 0{,}1\cdot 0 + 0{,}15\cdot 1 + 0{,}2\cdot 2 + 0{,}25\cdot 3 + 0{,}15\cdot 4 + 0{,}1\cdot 5 + \\ + 0{,}05\cdot 6 = 2{,}7.$$

Вообще, если таблица вероятностей для случайной величины α имеет вид

значения случайной величины	a_1	a_2	a_3	...	a_k
вероятности	p_1	p_2	p_3	...	p_k

то среднее значение этой величины определяется формулой

$$\text{ср. зн. } \alpha = p_1a_1 + p_2a_2 + p_3a_3 + \ldots + p_ka_k.$$

Из этой формулы, в частности, следует, что *среднее значение случайной величины* является именно средним, т. е. что оно всегда *не превосходит наибольшего из возможных значений случайной величины и не меньше наименьшего из ее значений*. В самом деле, если a_1 — наибольшее значение случайной величины α (т. е. $a_1 \geqslant a_2$, $a_1 \geqslant a_3$, ..., $a_1 \geqslant a_k$), а a_k — наименьшее ее значение (т. е. $a_k \leqslant a_1$, $a_k \leqslant a_2$, ..., $a_k \leqslant a_{k-1}$), то

$$\text{ср. зн. } \alpha = p_1a_1 + p_2a_2 + \ldots + p_ka_k \leqslant p_1a_1 + p_2a_1 + \ldots \\ \ldots + p_ka_1 = (p_1 + p_2 + \ldots + p_k)a_1 = a_1$$

и

$$\text{ср. зн. } \alpha = p_1a_1 + p_2a_2 + \ldots + p_ka_k \geqslant p_1a_k + p_2a_k + \ldots \\ \ldots + p_ka_k = (p_1 + p_2 + \ldots + p_k)a_k = a_k$$

(ибо $p_1 + p_2 + \ldots + p_k = 1$).

Задача 4. *Пусть таблицы вероятностей, указывающие частоту попаданий в мишень для двух стрелков A и B, имеют вид:*

для стрелка A

число очков	0	1	2	3	4	5	6	7	8	9	10
вероятности	0,02	0,03	0,05	0,1	0,15	0,2	0,2	0,1	0,07	0,05	0,03

для стрелка B

число очков	0	1	2	3	4	5	6	7	8	9	10
вероятности	0,01	0,01	0,04	0,1	0,25	0,3	0,18	0,05	0,03	0,02	0,01

Кого из стрелков следует считать более метким?

Здесь среднее число очков, выбиваемых одним выстрелом, для стрелка A равно

$$0{,}02 \cdot 0 + 0{,}03 \cdot 1 + 0{,}05 \cdot 2 + 0{,}1 \cdot 3 + 0{,}15 \cdot 4 + 0{,}2 \cdot 5 + \\ + 0{,}2 \cdot 6 + 0{,}1 \cdot 7 + 0{,}07 \cdot 8 + 0{,}05 \cdot 9 + 0{,}03 \cdot 10 = 5{,}24,$$

а для стрелка B оно меньше:

$$0{,}01 \cdot 0 + 0{,}01 \cdot 1 + 0{,}04 \cdot 2 + 0{,}1 \cdot 3 + 0{,}25 \cdot 4 + 0{,}3 \cdot 5 + \\ + 0{,}18 \cdot 6 + 0{,}05 \cdot 7 + 0{,}03 \cdot 8 + 0{,}02 \cdot 9 + 0{,}01 \cdot 10 = \\ = 4{,}84 < 5{,}24.$$

Поэтому более метким следует считать первого стрелка.

§ 2. Свойства вероятности. Сложение и умножение событий. Несовместимые и независимые события

Из приведенного в предыдущем параграфе определения вероятности следует, что вероятность $p(A)$ любого события A есть правильная дробь:

$$0 \leqslant p(A) \leqslant 1.$$

При этом вероятность может равняться 1: это будет означать, что событие A осуществляется при л ю б о м исходе рассматриваемого опыта, т. е. что событие A д о с т о в е р н о (так, например, равна 1 вероятность вынуть белый шар из урны, в которой заключены только белые шары). Вероятность может также равняться 0: это означает, что

событие не осуществляется ни при каком исходе опыта, т. е. оно невозможно (нулю равна вероятность вынуть из урны с белыми шарами черный шар).

Пусть теперь рассматриваемый опыт может иметь лишь два взаимно исключающих друг друга исхода A и B. В таком случае мы будем называть событие B противоположным событию A и обозначать его через $\bar{A}$ (читать эту запись можно как «не A»). Если событие A реализуется при m из n равновероятных исходов опыта, то событие $\bar{A}$ реализуется при остальных $n - m$ исходах; поэтому $p(A) = \frac{m}{n}$, $p(\bar{A}) = \frac{n-m}{n} = 1 - \frac{m}{n}$ и, следовательно,

$$p(\bar{A}) = 1 - p(A).$$

Таким образом, таблица вероятностей для опыта, имеющего только два исхода, имеет простой вид:

A	$\bar{A}$
$p(A)$	$1 - p(A)$

Рассмотрим теперь такие два события A и A_1, что выполнение события A с необходимостью влечет за собой и выполнение события A_1 (например, A есть выпадение шестерки при бросании игральной кости, а A_1 — выпадение цифры, делящейся на 3). В таком случае событие A_1 заведомо должно выполняться при всех тех исходах опыта, при которых выполняется событие A; поэтому вероятность события A_1 не может быть меньше вероятности события A. То обстоятельство, что выполнение A влечет за собой выполнение A_1, мы будем записывать в виде $A \subset A_1$ (читается «A влечет A_1»). Таким образом мы имеем следующее важное свойство вероятностей:

$$\text{если } A \subset A_1, \text{ то } p(A) \leqslant p(A_1).$$

Рассмотрим далее событие, которое состоит в том, что выполняется хоть одно из каких-то двух событий A и B; это событие мы будем называть суммой событий A и B и обозначать через $A + B$. При этом могут иметь место два существенно различных случая. Если события A и B несовместимы, т. е. сразу оба они не могут иметь места, то событие A выполняется при каких-то m_1

из n равновероятных исходов опыта, а B — при m_2 других исходах; в таком случае

$$p(A)=\frac{m_1}{n},\ p(B)=\frac{m_2}{n}\text{ и }p(A+B)=\frac{m_1+m_2}{n}=\frac{m_1}{n}+\frac{m_2}{n},$$

т. е.

$$p(A+B)=p(A)+p(B)$$

(правило сложения вероятностей). Так в рассмотренном на стр. 19 примере вероятность того, что будет извлечен белый или черный шар, в силу правила сложения, равна

$$\frac{1}{2}+\frac{3}{10}=\frac{4}{5}.$$

Свойство вероятностей, составляющее содержание правила сложения, может быть обобщено следующим образом. Пусть мы имеем k событий $A_1, A_2, \ldots, A_k$, никакие два из которых несовместимы между собой; обозначим через $A_1+A_2+\ldots+A_k$ событие, которое состоит в том, что выполняется хоть одно из рассматриваемых k событий. В таком случае, очевидно,

$$p(A_1+A_2+\ldots+A_k)=\\=p(A_1)+p(A_2)+\ldots+p(A_k);$$

этот более общий результат также иногда называют правилом сложения вероятностей. В частности, если опыт может иметь k (и только k) различных взаимоисключающих друг друга исходов, то ему отвечает таблица вероятностей

A_1	A_2	...	A_k
$p(A_1)$	$p(A_2)$	...	$p(A_k)$

в которой *сумма стоящих в нижней строке чисел равна единице*:

$$p(A_1)+p(A_2)+\ldots+p(A_k)=1;$$

это следует из того, что $p(A_1)+p(A_2)+\ldots+p(A_k)=p(A_1+A_2+\ldots+A_k)$, а событие $A_1+A_2+\ldots+A_k$ достоверно (ибо какой-то один исход опыта осуществляется наверное).

Предположим теперь, что события A и B совместимы, т. е. могут реализоваться одновременно. В таком случае уже нельзя утверждать, что $p(A+B)=p(A)+p(B)$. Действительно, пусть событие A выполняется при m_1 из n равновозможных исходов опыта, а событие B — при m_2 из этих n исходов. Событие $A+B$ выполняется, если имеет место один из m_1 первых или один из m_2 вторых исходов; однако, так как эти исходы уже не обязательно все различны, то общее число их может оказаться меньшим, чем m_1+m_2. Таким образом, в общем случае можно лишь утверждать, что *вероятность суммы двух событий* всегда *не превосходит сумму их вероятностей*:

$$p(A+B)\leqslant p(A)+p(B)$$

(но $p(A+B)\geqslant p(A)$ и $p(A+B)\geqslant p(B)$, ибо в силу определения суммы событий $A\subset A+B$ и $B\subset A+B$). Аналогично и для любого числа k (не обязательно взаимоисключающих друг друга) событий имеем

$$p(A_1+A_2+\ldots+A_k)\leqslant p(A_1)+p(A_2)+\ldots+p(A_k).$$

Неравенство $p(A+B)\leqslant p(A)+p(B)$ можно несколько уточнить. Назовем произведением двух событий A и B событие, которое состоит в том, что выполняются оба события; обозначим его через AB. Рассмотрим m_1 равновероятных исходов опыта, при которых выполняется событие A, и m_2 исходов, при которых выполняется событие B; предположим, что имеется точно l исходов, которые входят и в число m_1 первых исходов и в число m_2 вторых. Очевидно, что если имеет место один из этих l исходов (и только в этом случае), то выполняются сразу оба события A и B; поэтому $p(AB)=\frac{l}{n}$. С другой стороны, если среди m_1 первых исходов и m_2 вторых исходов имеется ровно l одинаковых, то всего мы имеем m_1+m_2-l исходов (в сумме m_1+m_2 имеется l исходов, которые засчитываются дважды). Таким образом, здесь

$$p(A+B)=\frac{m_1+m_2-l}{n}=\frac{m_1}{n}+\frac{m_2}{n}-\frac{l}{n}$$

и, следовательно,

$$p(A+B)=p(A)+p(B)-p(AB).$$

Мы видим, что задача определения вероятности с у м м ы $A + B$ событий A и B сводится к нахождению вероятности п р о и з в е д е н и я AB этих событий. Последняя задача, в общем случае не очень простая, будет рассмотрена в следующем параграфе. Однако имеется один частный случай, когда нахождение вероятности события AB не составляет труда. Это — случай, когда события A и B являются н е з а в и с и м ы м и, т. е. когда результат опыта, с которым связано выполнение или невыполнение события A, никак не отражается на условиях опыта, с результатом которого связано событие B. Так, например, независимы события, состоящие в извлечении черного шара из д в у х р а з л и ч н ы х урн, содержащих белые и черные шары; однако два последовательных извлечения черного шара из о д н о й урны (без возвращения вынутого шара обратно в урну) не представляют собой независимых событий (поскольку результат первого извлечения влияет на число оставшихся в урне черных шаров и, следовательно, отражается на условиях второго опыта).

Пусть событие A реализуется при m_1 из n_1 равновероятных исходов первого опыта, а независимое от него событие B — при m_2 из n_2 равновероятных исходов второго опыта; в этом случае вероятность события A равна $\frac{m_1}{n_1}$, а вероятность B равна $\frac{m_2}{n_2}$. Рассмотрим теперь сложный опыт, состоящий в том, что производятся о б а наши опыта. Очевидно, что этот сложный опыт может иметь $n_1 n_2$ различных равновероятных исходов, поскольку каждому из n_1 исходов первого опыта могут отвечать n_2 различных исходов второго опыта. Из этих $n_1 n_2$ равновероятных исходов событию AB будут благоприятствовать $m_1 m_2$ исходов, которые получаются, если комбинировать m_1 исходов первого опыта, благоприятствующих событию A, с m_2 исходами второго опыта, благоприятствующими B. Таким образом, вероятность события AB будет равна

$$\frac{m_1 m_2}{n_1 n_2} = \frac{m_1}{n_1} \cdot \frac{m_2}{n_2}$$

и, значит,

$$p(AB) = p(A)p(B)$$

(п р а в и л о у м н о ж е н и я в е р о я т н о с т е й).

Это правило можно обобщить следующим образом. Пусть $A_1, A_2, \ldots, A_k$ — какие-то k взаимно независимых событий, т. е. условия опыта, с результатом которого связано какое-либо одно из этих событий, никак не зависят от выполнения или невыполнения остальных событий. В таком случае

$$p(A_1A_2 \ldots A_k) = p(A_1)\, p(A_2) \ldots p(A_k).$$

Доказательство этого соотношения совершенно аналогично выводу формулы $p(AB) = p(A)p(B)$, составляющей его частный случай.

Если события A и B не независимы, то правило умножения $p(AB) = p(A)\, p(B)$ уже не обязано выполняться; так, например, если $B \subset A$ (скажем, A — выпадение четной цифры при бросании игральной кости, а B — выпадение двойки), то событие AB совпадает с событием B и, следовательно, $p(AB) = p(B)$. Пока мы можем лишь утверждать, что $p(AB) \leqslant p(A)$ и $p(AB) \leqslant p(B)$ (так как из определения произведения событий вытекает, что $AB \subset B$ и $AB \subset A$). Более подробно на вопросе о вероятности произведения двух событий мы остановимся в следующем параграфе.

Для того чтобы пояснить применения выведенных простейших свойств вероятностей, рассмотрим несколько задач.

Задача 5. *Какова вероятность того, что при двух бросаниях монеты оба раза сверху окажется герб?*

Здесь ищется вероятность события AB, где A есть выпадение герба при первом бросании, а B — выпадение герба при втором бросании. События A и B, очевидно, независимы; поэтому

$$p(AB) = p(A)\, p(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

(см. задачу 1 на стр. 19).

Задача 6. *Какова вероятность того, что взятое наудачу целое положительное число, не превосходящее тысячи, окажется целой степенью* (с показателем, бо́льшим единицы) *другого целого числа?*

Слово «наудачу» в условии этой задачи означает, что мы считаем появление любого числа от 1 до 1000

равновероятным. Далее, так как

$$2^9 < 1000 < 2^{10},\ 3^6 < 1000 < 3^7,\ 5^4 < 1000 < 5^5,$$
$$6^3 < 1000 < 6^4,\ 7^3 < 1000 < 7^4,\ 10^3 = 1000 < 10^4,$$
$$11^2 < 1000 < 11^3,\ 12^2 < 1000 < 12^3, \ldots$$
$$\ldots,\ 31^2 < 1000 < 31^3,\ 32^2 > 1000,$$

то вероятность того, что число будет являться целой степенью 2, равна $\frac{8}{1000}$ (среди 1000 чисел от 1 до 1000 имеется 8 степеней двойки: $2^2 = 4$, $2^3 = 8$, 2^4, 2^5, 2^6, 2^7, 2^8 и 2^9); точно так же вероятность того, что наше число будет целой степенью 3, 5, 6, 7, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 26, 28, 29, 30, 31 соответственно равна $\frac{5}{1000}, \frac{3}{1000}, \frac{2}{1000}, \frac{2}{1000}, \frac{2}{1000}, \frac{1}{1000}, \frac{1}{1000}, \ldots, \frac{1}{1000}$ (если число является целой степенью 4, 8, 9, 16, 25 или 27, то оно одновременно является и целой степенью меньшего числа; поэтому эти случаи мы из рассмотрения исключили). Так как все соответствующие события попарно несовместимы, то искомая вероятность равна

$$\frac{8}{1000} + \frac{5}{1000} + \frac{3}{1000} + \frac{2}{1000} + \frac{2}{1000} + \frac{2}{1000} + \underbrace{\frac{1}{1000} + \frac{1}{1000} + \ldots + \frac{1}{1000}}_{18 \text{ раз}} = \frac{40}{1000} = \frac{1}{25}.$$

Задача 7. *В колоде* 52 *карты; одна из четырех мастей объявляется «козырной». Какова вероятность того, что взятая наудачу карта является тузом или козырем?*

Пусть событие A заключается в том, что выбранная карта является тузом, а событие B в том, что она является козырем; в таком случае событие AB состоит в том, что эта карта является козырным тузом и $p(A) = \frac{1}{13}$ (в колоде имеются карты 13-ти наименований: двойки, тройки, ..., тузы), $p(B) = \frac{1}{4}$, $p(AB) = \frac{1}{52}$. Отсюда вытекает, что искомая вероятность равна

$$p(A + B) = p(A) + p(B) - p(AB) = \frac{1}{13} + \frac{1}{4} - \frac{1}{52} = \frac{4}{13}.$$

З а д а ч а 8. *Шесть охотников увидели лису и одновременно выстрелили в нее. Предположим, что каждый из охотников на таком расстоянии обычно попадает в лису и убивает ее в одном случае из трех. Какова вероятность того, что лиса будет убита?*

Пусть события A_1, A_2, . . . ,A_6 означают поражение лисы 1-м, 2-м, . . . , 6-м охотником. В условии задачи указывается, что $p(A_1) = p(A_2) = \ldots = p(A_6) = \frac{1}{3}$; требуется найти $p(S)$, где $S = A_1 + A_2 + \ldots + A_6$. События A_1, A_2, . . .,A_6, очевидно, независимы; это дает возможность при решении этой задачи воспользоваться формулой

$$p(A + B) = p(A) + p(B) - p(AB) = \\ = p(A) + p(B) - p(A)p(B)$$

(см. ниже текст, напечатанный мелким шрифтом). Однако такое решение не очень просто, так как формула, выражающая вероятность суммы м н о г и х (совместимых) событий довольно сложна.

Более удобно иное решение этой задачи. Будем искать вероятность $p(\bar{S})$ того, что лиса у ц е л е е т. Промах 1-го, 2-го, . . ., 6-го охотника естественно обозначить через $\bar{A}_1$, $\bar{A}_2$, . . .,$\bar{A}_6$; в силу формулы $p(\bar{A}) = 1 - p(A)$ имеем $p(\bar{A}_1) = p(\bar{A}_2) = \ldots = p(\bar{A}_6) = \frac{2}{3}$. Для того чтобы лиса уцелела, необходимо, чтобы в с е охотники промахнулись, т. е. здесь речь идет о вероятности события $\bar{A}_1\bar{A}_2 \ldots \bar{A}_6$, где события $\bar{A}_1, \bar{A}_2, \ldots, \bar{A}_6$ — взаимно независимы. Итак,

$$p(\bar{S}) = p(\bar{A}_1\bar{A}_2 \ldots \bar{A}_6) = p(\bar{A}_1) \cdot \ldots \cdot p(\bar{A}_6) = \\ = \frac{2}{3} \cdot \frac{2}{3} \cdot \ldots \cdot \frac{2}{3} = \frac{2^6}{3^6} = \frac{64}{729},$$

—и в силу той же формулы $p(\bar{A}) = 1 - p(A)$,

$$p(S) = 1 - \frac{64}{729} = \frac{665}{729}.$$

Формулу $p(A + B) = p(A) + p(B) - p(AB)$ можно обобщить и на случай отыскания вероятности суммы произвольного числа k (возможно — совместимых!) событий $A_1, A_2, \ldots, A_k$. Имеем

$$p(A_1 + A_2 + A_3) = p\{(A_1 + A_2) + A_3\} = \\ = p(A_1 + A_2) + p(A_3) - p\{(A_1 + A_2)A_3\}.$$

Здесь $p(A_1 + A_2) = p(A_1) + p(A_2) - p(A_1A_2)$. Сложнее раскрыть смысл выражения $p\{(A_1 + A_2)A_3\}$. Согласно определению суммы и произведения событий событие $(A_1 + A_2)A_3$ состоит в том, что выполняется х о т ь о д н о из событий A_1 и A_2 и, к р о м е т о г о, событие A_3. Но это означает, что выполняется по крайней мере одно из событий A_1A_3 и A_2A_3, состоящих в выполнении A_1 и A_3, соответственно A_2 и A_3. Таким образом, имеем

$$(A_1 + A_2)A_3 = A_1A_3 + A_2A_3$$

и, следовательно,

$$p\{(A_1 + A_2)A_3\} = p(A_1A_3 + A_2A_3) = \\ = p(A_1A_3) + p(A_2A_3) - p\{(A_1A_3)(A_2A_3)\}.$$

Далее, событие $(A_1A_3)(A_2A_3)$ состоит в том, что выполняются сразу о б а события A_1A_3 (т. е. и A_1, и A_3) и A_2A_3 (и A_2, и A_3). Другими словами, событие $(A_1A_3)(A_2A_3)$ состоит в том, что выполняются т р и события A_1, A_2 и A_3,—т. е. оно не отличается от события $A_1A_2A_3$.

Таким образом, окончательно получаем

$$p(A_1 + A_2 + A_3) = \\ = p(A_1) + p(A_2) - p(A_1A_2) + p(A_3) - p(A_1A_3) - p(A_2A_3) + \\ + p(A_1A_2A_3)$$

или, в другом порядке,

$$p(A_1 + A_2 + A_3) = \\ = p(A_1) + p(A_2) - p(A_1A_2) + p(A_3) - p(A_1A_3) - p(A_2A_3) + \\ + p(A_1A_2A_3).$$

Аналогично этому для произвольного k имеем

$$p(A_1 + A_2 + \ldots + A_k) = \\ = p(A_1) + p(A_2) + \ldots + p(A_k) - \\ - p(A_1A_2) - p(A_1A_3) - \ldots - p(A_{k-1}A_k) + \\ + p(A_1A_2A_3) + p(A_1A_2A_4) + \ldots + p(A_{k-2}A_{k-1}A_k) - \\ - p(A_1A_2A_3A_4) - \ldots + (-1)^{k-1} p(A_1A_2 \ldots A_k).$$

Доказать эту формулу можно с помощью метода математической индукции подобно тому, как мы доказали ее для $k = 3$.

Покажем, как с помощью приведенной формулы решить задачу 8. При $k = 6$ имеем:

$$p(A_1 + A_2 + \ldots + A_6) = \\ = p(A_1) + p(A_2) + \ldots + p(A_6) - p(A_1A_2) - \\ - p(A_1A_3) - \ldots - p(A_5A_6) + \\ + p(A_1A_2A_3) + p(A_1A_2A_4) + \ldots + p(A_4A_5A_6) - \ldots \\ \ldots - p(A_1A_2A_3A_4A_5A_6).$$

Но (события $A_1, A_2, \ldots, A_k$ взаимно независимы!)

$$p(A_1) = p(A_2) = \ldots = p(A_6) = \frac{1}{3},$$

$$p(A_1A_2) = p(A_1A_3) = \ldots = p(A_5A_6) = p(A_1)\,p(A_2) = \left(\frac{1}{3}\right)^2,$$

$$p(A_1A_2A_3) = \ldots = p(A_4A_5A_6) = p(A_1)\,p(A_2)\,p(A_3) = \left(\frac{1}{3}\right)^3, \ldots$$

$$\ldots, p(A_1A_2 \ldots A_6) = p(A_1)\,p(A_2) \ldots p(A_6) = \left(\frac{1}{3}\right)^6,$$

откуда получаем:

$$p(A_1 + A_2 + \ldots + A_6) =$$

$$= 6 \cdot \frac{1}{3} - C_6^2\left(\frac{1}{3}\right)^2 + C_6^3\left(\frac{1}{3}\right)^3 - C_6^4\left(\frac{1}{3}\right)^4 + C_6^5\left(\frac{1}{3}\right)^5 - C_6^6\left(\frac{1}{3}\right)^6 =$$

$$= 1 - \left(1 - \frac{1}{3}\right)^6 = 1 - \left(\frac{2}{3}\right)^6 = \frac{665}{729},$$

— т. е. тот же результат, что и выше.

Другие примеры применения этой общей формулы могут быть найдены, например, в книге [37].

Обратимся теперь к понятиям суммы и произведения *случайных величин*, которые также будут полезны нам в дальнейшем. В качестве примера, иллюстрирующего первое из этих понятий, рассмотрим следующую задачу:

Задача 9. *В цехе установлены два различных станка, производящих одинаковые изделия. Из опыта известно, что* 1-*й* (*более старый*) *станок может произвести в сутки до трех бракованных изделий, причем вероятности числа бракованных изделий здесь таковы:*

число браков. изделий (в сутки)	0	1	2	3
вероятности	0,3	0,4	0,2	0,1

2-й же (*новый*) *станок производит не больше одного бракованного изделия в сутки, причем вероятность того, что хоть одно из произведенных за сутки изделий окажется бракованным, равна здесь всего* 0,1:

число браков. изделий (в сутки)	0	1
вероятности	0,9	0,1

Спрашивается, каково среднее число произведенных цехом в сутки бракованных изделий?

В этой задаче одновременно рассматриваются д в е случайные величины α и β, первая из которых принимает значения a_0, a_1, a_2 и a_3 (а именно, 0, 1, 2 и 3) с вероятностями p_0, p_1, p_2 и p_3 (в данном случае равными 0,3, 0,4, 0,2 и 0,1; ясно, что $p_0 + p_1 + p_2 + p_3 = 1$), а вторая принимает всего два значения b_0 и b_1 (а именно, 0 и 1) с вероятностями q_0 и q_1 (равными 0,9 и 0,1; ясно, что $q_0 + q_1 = 1$). Средние значения этих случайных величин (среднее значение числа α бракованных изделий, производимых за сутки 1-м станком, и среднее значение числа β бракованных изделий, производимых за сутки 2-м станком) соответственно равны

$$\text{ср. зн. } \alpha = p_0a_0 + p_1a_1 + p_2a_2 + p_3a_3 = \\ = 0{,}3 \cdot 0 + 0{,}4 \cdot 1 + 0{,}2 \cdot 2 + 0{,}1 \cdot 3 = 1{,}1$$

и

$$\text{ср. зн. } \beta = q_0b_0 + q_1b_1 = 0{,}9 \cdot 0 + 0{,}1 \cdot 1 = 0{,}1.$$

Нас же интересует случайная величина $\alpha + \beta$ — число бракованных изделий, производимых за сутки о б о и м и станками; эта величина может принимать значения $a_0 + b_0$, $a_0 + b_1$; $a_1 + b_0$, $a_1 + b_1$; $a_2 + b_0$, $a_2 + b_1$; $a_3 + b_0$ и $a_3 + b_1$ (в нашем случае — значения 0, 1, 2, 3 и 4). Будем (пока!) считать, что случайные величины α и β *независимы*, т. е. что, скажем, величина α принимает значения 0, 1, 2 и 3 с вероятностями p_0, p_1, p_2 и p_3 (т. е. 0,3, 0,4, 0,2 и 0,1) независимо от того, какое значение принимает (для тех же суток) величина β. В таком случае и события $\alpha = a_i$ (где $i = 0$, 1, 2 или 3) и $\beta = b_j$ (где $j = 0$ или 1) будут н е з а в и с и м ы, — и значит,

$$p\,(\alpha = a_i \text{ и } \beta = b_j) = p(\alpha = a_i) \cdot p\,(\beta = b_j) = p_iq_j.$$

Исходя отсюда, мы приходим к следующей (детализированной) таблице значений случайной величины $\alpha + \beta$:

значения	$a_0 + b_0\,(=0)$	$a_0 + b_1\,(=1)$	$a_1 + b_0\,(=1)$	$a_1 + b_1\,(=2)$
вероятности	$p_0q_0\,(=0{,}27)$	$p_0q_1\,(=0{,}03)$	$p_1q_0\,(=0{,}36)$	$p_1q_1\,(=0{,}04)$
$a_2 + b_0\,(=2)$	$a_2 + b_1\,(=0)$	$a_3 + b_0\,(=3)$	$a_3 + b_1\,(=4)$	
$p_2q_0\,(=0{,}18)$	$p_2q_1\,(=0{,}02)$	$p_3q_0\,(=0{,}09)$	$p_3q_1\,(=0{,}01)$	

Теперь по основной формуле для среднего значения случайной величины имеем:

$$\begin{aligned}
\text{ср. зн. } (\alpha+\beta) &= p_0q_0(a_0+b_0)+p_0q_1(a_0+b_1)+\\
&+p_1q_0(a_1+b_0)+p_1q_1(a_1+b_1)+p_2q_0(a_2+b_0)+\\
&+p_2q_1(a_2+b_1)+p_3q_0(a_3+b_0)+p_3q_1(a_3+b_1)=\\
&=a_0(p_0q_0+p_0q_1)+a_1(p_1q_0+p_1q_1)+a_2(p_2q_0+p_2q_1)+\\
&+a_3(p_3q_0+p_1q_1)+b_0(p_0q_0+p_1q_0+p_2q_0+p_3q_0)+\\
&+b_1(p_0q_1+p_1q_1+p_2q_1+p_2q_1)=\\
&=a_0p_0(q_0+q_1)+a_1p_1(q_0+q_1)+a_2p_2(q_0+q_1)+\\
&+a_3p_3(q_0+q_1)+b_0q_0(p_0+p_1+p_2+p_3)+\\
&+b_1q_1(p_0+p_1+p_2+p_3)=\\
&=(a_0p_0+a_1p_1+a_2p_2+a_3p_3)+(b_0q_0+b_1q_1)=\\
&=\text{ср. зн. } \alpha+\text{ср. зн. } \beta=1{,}2 \text{ (браков. изделий/сутки).}
\end{aligned}$$

Таким образом мы видим, что *среднее значение суммы двух случайных величин равно сумме их средних значений.*

Впрочем, надо заметить, что последнее заключение, полученное нами с помощью довольно утомительных преобразований выражения для ср.зн.$(\alpha+\beta)$, не является особенно глубоким. В самом деле, пусть в какой-то день, который мы назовем первым, 1-й станок произвел $a^{(1)}$ бракованных изделий (где $a^{(1)}$ равно 0, 1, 2 или 3), а 2-й станок — $b^{(1)}$ бракованных изделий (где $b^{(1)}$ равно 0 или 1). Аналогично этому пусть во второй, в третий, . . ., в n-й день 1-й станок производит $a^{(2)}, a^{(3)}, \ldots, a^{(n)}$ бракованных изделий, а 2-й станок — $b^{(2)}, b^{(3)}, \ldots, b^{(n)}$ бракованных изделий. Тогда общее число выпускаемых цехом бракованных изделий в первый, во второй, в третий, ... , в n-й день равно

$$a^{(1)}+b^{(1)},\ a^{(2)}+b^{(2)},\ a^{(3)}+b^{(3)},\ \ldots\ldots a^{(n)}+b^{(n)},$$

а среднее число выпускаемых за сутки бракованных изделий будет равно

$$\frac{(a^{(1)}+b^{(1)})+(a^{(2)}+b^{(2)})+(a^{(3)}+b^{(3)})+\ldots+(a^{(n)}+b^{(n)})}{n}=$$

$$=\frac{a^{(1)}+a^{(2)}+a^{(3)}+\ldots+a^{(n)}}{n}+\frac{b^{(1)}+b^{(2)}+b^{(3)}+\ldots+b^{(n)}}{n}.$$

Но при большом n величина

$$\frac{(a^{(1)}+b^{(1)})+(a^{(2)}+b^{(2)})+(a^{(3)}+b^{(3)})+\ldots+(a^{(n)}+b^{(n)})}{n}$$

будет очень близка к ср. зн. $(\alpha+\beta)$, а величины

$$\frac{a^{(1)}+a^{(2)}+a^{(3)}+\ldots+a^{(n)}}{n}$$

и

$$\frac{b^{(1)}+b^{(2)}+b^{(3)}+\ldots+b^{(n)}}{n}$$

— к ср. зн. α и ср. зн. β, откуда и следует, что

$$\text{ср. зн. } (\alpha+\beta)=\text{ср. зн. } \alpha+\text{ср. зн. } \beta.$$

Замечательно, что устанавливаемый последним простым рассуждением результат является более общим, чем тот, который был доказан ранее! В самом деле, в этом рассуждении мы никак не использовали независимость величин α и β (которая ведь и на самом деле может не иметь места, ибо на работе обоих станков могут сказываться некоторые общие факторы, связанные, например, с тем, что оба станка используют одно и то же сырье). Но в этом последнем случае мы уже не сможем утверждать, что

$$p\,(\alpha=a_i \text{ и } \beta=b_j)=p\,(\alpha=a_i)\cdot p\,(\beta=b_j)=p_iq_j;$$

поэтому вместо величин p_0q_0, p_0q_1 и т. д. в таблице значений случайной величины $\alpha+\beta$ будут стоять какие-то вероятности p_{00} (вероятность того, что $\alpha=a_0$ и $\beta=b_0$), p_{01} (вероятность того, что $\alpha=a_0$ и $\beta=b_1$) и т. д., численные значения которых зависят от неизвестной нам во всех деталях связи между величинами α и β.

Впрочем, это обстоятельство почти не отразится на проведенной выше выкладке. В самом деле, теперь мы будем иметь

$$\begin{aligned}\text{ср. зн. } (\alpha+\beta) &= p_{00}(a_0+b_0)+p_{01}(a_0+b_1)+\\ &+p_{10}(a_1+b_0)+p_{11}(a_1+b_1)+p_{20}(a_2+b_0)+p_{21}(a_2+b_1)+\\ &+p_{30}(a_3+b_0)+p_{31}(a_3+b_1)=a_0(p_{00}+p_{01})+\\ &+a_1(p_{10}+p_{11})+a_2(p_{20}+p_{21})+p_3(p_{30}+p_{31})+\\ &+b_0(p_{00}+p_{10}+p_{20}+p_{30})+b_1(p_{01}+p_{11}+p_{21}+p_{31}).\end{aligned}$$

Но

$$p_{00} + p_{01} = p\,(\alpha = a_0 \text{ и } \beta = b_0) + p\,(\alpha = a_0 \text{ и } \beta = b_1) = \\ = p\,(\alpha = a_0 \text{ и } \beta = b_0 \text{ или } b_1).$$

Однако b_0 и b_1 — это в с е возможные значения случайной величины β, так что $p\,(\alpha = a_0$ и $\beta = b_0$ или $b_1)$ есть не что иное, как просто $p\,(\alpha = a_0) = p_0$! Точно так же устанавливается, что

$$p_{10} + p_{11} = p_1,\ p_{20} + p_{21} = p_2,\ p_{30} + p_{31} = p_3.$$

Далее,

$$p_{00} + p_{10} + p_{20} + p_{30} = p\,(\alpha = a_0 \text{ и } \beta = b_0) + \\ + p\,(\alpha = a_1 \text{ и } \beta = b_0) + p\,(\alpha = a_2 \text{ и } \beta = b_0) + \\ + p\,(\alpha = a_3 \text{ и } \beta = b_0) = \\ = p\,(\alpha = a_0, \text{ или } a_1, \text{ или } a_2, \text{ или } a_3 \text{ и } \beta = b_0) = \\ = p\,(\beta = b_0) = q_0$$

и аналогично

$$p_{01} + p_{11} + p_{21} + p_{31} = q_1.$$

Таким образом, в этом случае мы по-прежнему имеем

$$\text{ср. зн. } (\alpha + \beta) = \\ = (a_0p_0 + a_1p_1 + a_2p_2 + a_3p_3) + (b_0q_0 + b_1q_1) = \\ = \text{ср. зн. } \alpha + \text{ср. зн. } \beta.$$

Разумеется, полученный нами результат можно распространить и на л ю б о е число случайных величин, для которых тоже *среднее значение их суммы равно сумме их средних значений.*

Обратимся теперь к ситуации, в которой возникает понятие п р о и з в е д е н и я двух случайных величин:

З а д а ч а 10. *Фермер ежегодно отправляет на рынок a_0, a_1, a_2 или a_3 телят, причем вероятности (частоты) отдельных значений числа проданных телят здесь таковы:*

число телят	a_0	a_1	a_2	a_3
вероятности	p_0	p_1	p_2	p_3

(разумеется, $p_0 + p_1 + p_2 + p_3 = 1$). *С другой стороны, цена одного теленка в разные годы может равняться или b_0 или b_1, причем вероятности этих цен равны соответственно q_0 и q_1 $(= 1 - q_0)$:*

цена теленка	b_0	b_1
вероятность	q_0	q_1

Спрашивается, какова средняя годовая выручка фермера от продажи телят?

Здесь мы снова имеем дело с двумя случайными величинами α и β, причем для сохранения аналогии с задачей 9 мы обозначили возможные значения этих величин и соответствующие этим значениям вероятности теми же символами a_0, a_1, a_2, a_3; b_0, b_1, и p_0, p_1, p_2, p_3; q_0, q_1, как и выше. Интересует же нас *произведение* $\alpha\beta$ этих двух величин (произведение числа проданных телят на цену одного теленка), которое может иметь 8 значений a_0b_0, a_0b_1; a_1b_0, a_1b_1; a_2b_0, a_2b_1; a_3b_0, a_3b_1. При этом если считать величины α и β независимыми, то таблица вероятностей отдельных значений величины $\alpha\beta$ будет иметь вид

значения	a_0b_0	a_0b_1	a_1b_0	a_1b_1	a_2b_0	a_2b_1	a_3b_0	a_3b_1
вероятности	p_0q_0	p_0q_1	p_1q_0	p_1q_1	p_2q_0	p_2q_1	p_3q_0	p_3q_1

Поэтому среднее значение величины $\alpha\beta$ в этом случае равно

$$\begin{aligned}\text{ср.зн.}(\alpha\beta) &= p_0q_0a_0b_0 + p_0q_1a_0b_1 + p_1q_0a_1b_0 + p_1q_1a_1b_1 + \\ &+ p_2q_0a_2b_0 + p_2q_1a_2b_1 + p_3q_0a_3b_0 + p_3q_1a_3b_1 = \\ &= p_0a_0(q_0b_0 + q_1b_1) + p_1a_1(q_0b_0 + q_1b_1) + \\ &+ p_2a_2(q_0b_0 + q_1b_1) + p_3a_3(q_0b_0 + q_1b_1) = \\ &= (p_0a_0 + p_1a_1 + p_2a_2 + p_3a_3)(q_0b_0 + q_1b_1) = \\ &= (\text{ср.зн.}\alpha)\cdot(\text{ср.зн.}\beta).\end{aligned}$$

Таким образом мы видим, что *для независимых случайных величин* α *и* β *среднее значение их произведения равно произведению средних значений этих величин.* Так же и для большего числа взаимно независимых случайных величин *среднее значение их произведения* всегда *равно произведению их средних значений.*

Заметим, однако, что, в противоположность случаю суммы двух случайных величин, для их произведения независимость величин является существенным условием, без которого полученный нами результат может оказаться уже неверным. Для иллюстрации этого достаточно рассмотреть случай, когда $\alpha_1 = \alpha_2 = \alpha$, где α характеризуется следующей таблицей вероятностей:

значения величины α	$+1$	-1
вероятности	0,5	0,5

В этом случае, очевидно,

$$\text{ср. зн. } \alpha_1 = \text{ср. зн. } \alpha_2 = 0{,}5\,(+1) + 0{,}5\,(-1) = 0,$$

так что

$$(\text{ср. зн. } \alpha_1)\cdot(\text{ср. зн. } \alpha_2) = 0\cdot 0 = 0,$$

в то время как величина $\alpha_1\cdot\alpha_2 = \alpha^2$ всегда равна $+1$ (ибо $+1)^2 = (-1)^2 = +1$), так что

$$\text{ср. зн. } (\alpha_1\alpha_2) = 1 > 0 = (\text{ср. зн. } \alpha_1)\cdot(\text{ср. зн. } \alpha_2).$$

С установленным на этом примере неравенством

$$\text{ср. зн. } (\alpha^2) > (\text{ср. зн. } \alpha)^2$$

мы снова встретимся в § 4 этой главы.

§ 3. Условные вероятности

Два события A и B мы назвали *независимыми*, если результат опыта, с которым связано событие A, не влияет на условия опыта, с которым связано B. Однако это обстоятельство вовсе не всегда имеет место. Соответствующий пример мы уже приводили выше; повторим его здесь подробнее. Пусть A — событие, состоящее в извлечении черного шара из урны, содержащей m черных и $n - m$ белых шаров, B — событие, состоящее в извлечении черного шара из т о й ж е урны п о с л е т о г о, как из нее уже вынут один шар. Очевидно, что если первый вынутый шар был черным, т. е. если событие A имело место, то в урне после первого извлечения остается $m - 1$ черных и $n - m$ белых шаров и поэтому вероятность события B будет равна $\frac{m-1}{n-1}$. Если же первый извлеченный шар был белым (имело место событие $\bar{A}$), то в урне останется m черных и $n - m - 1$ белых шаров, и искомая вероятность станет равной $\frac{m}{n-1}$. Таким образом, вероятность события B меняется в зависимости от того, осуществляется или не осуществляется A, т. е. вероятность события B здесь может принимать два различных значения $\left(\frac{m-1}{n-1}$ и $\frac{m}{n-1}\right)$, для которых следует иметь и различные обозначения.

Вероятность, которую имеет событие B в том случае, когда известно, что событие A имело место, мы будем называть условной вероятностью события B при условии A и обозначать через $p_A(B)$. Таким образом, в нашем случае $p_A(B) = \frac{m-1}{n-1}$. Аналогично определяется условная вероятность $p_{\bar{A}}(B)$ события B при условии $\bar{A}$ (т. е. при условии, что A не произошло); в нашем случае $p_{\bar{A}}(B) = \frac{m}{n-1}$.

Очевидно, что условная вероятность $p_A(B)$ какого-либо события B при определенном условии A может быть и меньше и больше безусловной вероятности $p(B)$ этого события (т. е. вероятности, которую имеет B, если про результат опыта, с которым связано A, ничего неизвестно). Так, в рассмотренном выше примере $p(B) = \frac{m}{n}$, поскольку заранее мы можем с равной вероятностью ожидать, что при втором извлечении мы вынем любой из n содержащихся в урне шаров, а из этих n шаров ровно m черных. Таким образом, здесь $p_A(B) = \frac{m-1}{n-1} < \frac{m}{n} = p(B)$, а $p_{\bar{A}}(B) = \frac{m}{n-1} > \frac{m}{n} = p(B)$. Если события A и B независимы, то, очевидно, $p_A(B) = p(B)$. Последнее условие даже можно считать точным математическим определением понятия *независимости* событий, позволяющим для любой пары событий A и B проверить, являются ли они независимыми или нет (см. по этому поводу пример, напечатанный мелким шрифтом в конце параграфа).

Условные вероятности можно вычислять аналогично тому, как мы вычисляли в § 1 безусловные вероятности. Пусть событию A благоприятствуют N равновероятных исходов опыта, позволяющего определить, выполняются ли или нет событие A и некоторое другое событие B, причем из этих N исходов M благоприятствуют также и B, а остальные $N - M$ не благоприятствуют B. В таком случае $p_A(B) = \frac{M}{N}$ $\left(\text{и } p_A(\bar{B}) = \frac{N-M}{N}\right)$. Так, например, в разобранном выше примере опыт, состоящий в последовательном извлечении двух шаров из урны с n шарами, имеет $n(n-1)$ равновероятных исходов (в первый раз

мы можем вынуть любой из n имеющихся шаров, во второй раз — один из $n-1$ оставшихся), из которых событию A благоприятствуют $N = m(n-1)$ исходов (в первый раз извлекается один из m черных шаров, затем — любой из $n-1$ оставшихся); из этих $m(n-1)$ исходов событию B благоприятствуют $M = m(m-1)$ исходов (в первый раз извлекается любой из m черных шаров, затем — любой из $m-1$ оставшихся черных шаров) и, следовательно, здесь

$$p_A(B) = \frac{M}{N} = \frac{m(m-1)}{m(n-1)} = \frac{m-1}{n-1}.$$

Обозначим теперь общее число равновероятных исходов опыта, с которым связано выполнение событий A и B, через K. Так как из этих K исходов выполнению и события A и события B благоприятствуют M исходов, то вероятность события AB, состоящего в том, что имеют место и A и B, равна $\frac{M}{K}$. Но $\frac{M}{K} = \frac{N}{K}\frac{M}{N}$, а $\frac{M}{N} = p_A(B)$ и $\frac{N}{K} = p(A)$ (из K равновероятных исходов опыта событию A благоприятствуют N). Следовательно, мы имеем

$$p(AB) = p(A)\, p_A(B).$$

Это и есть общее правило для определения вероятности произведения AB двух событий; его также часто называют правилом умножения вероятностей (то правило, которое мы называли правилом умножения в § 2, является его частным случаем). Таким образом, для того, чтобы найти $p(AB)$, надо знать условную вероятность $p_A(B)$, характеризующую зависимость, существующую между событиями A и B; одними вероятностями $p(A)$ и $p(B)$ вероятность события AB не определяется. В том случае, когда вероятность события B не меняется в результате наступления или ненаступления события A, т. е. когда события A и B независимы, $p_A(B) = p(B)$ и $p(AB) = p(A)\, p(B)$ — результат, который мы уже имели выше.

Из определения условной вероятности сразу выводятся следующие свойства этой величины:

а) $0 \leqslant p_A(B) \leqslant 1$; $p_A(B) = 1$, если $A \subset B$ (в частности, если B есть достоверное событие); $p_A(B) = 0$, если A и B несовместимы (в частности, если B есть невозможное событие);

б) если $B \subset B_1$, то $p_A(B) \leqslant p_A(B_1)$;

в) если B и C несовместимы, то $p_A(B + C) = p_A(B) + p_A(C)$; если $B_1, B_2, \ldots, B_k$ попарно несовместимы, то

$$p_A(B_1 + B_2 + \ldots + B_k) = p_A(B_1) + p_A(B_2) + \ldots + p_A(B_k);$$

г) $p_A(\bar{B}) = 1 - p_A(B)$.

Доказательства этих свойств совершенно аналогичны приведенным в § 2 доказательствам тех же свойств для обычных (безусловных) вероятностей.

Заметим еще, что из формулы $p(AB) = p(A)\, p_A(B)$ следует, что

$$p(A)\, p_A(B) = p(B)\, p_B(A) \text{ или } \frac{p_B(A)}{p(A)} = \frac{p_A(B)}{p(B)}$$

(ибо события AB и BA, разумеется, не отличаются). Отсюда вытекает, в частности, что, зная вероятности $p(A)$ и $p(B)$ двух событий A и B и условную вероятность $p_A(B)$ события B при условии A, мы можем определить также и условную вероятность $p_B(A)$:

$$p_B(A) = p_A(B) \cdot \frac{p(A)}{p(B)}.$$

Так, в разобранном выше примере с извлечением шаров из урны $p(A) = p(B) = \frac{m}{n}$ (вероятности того, что при первом и что при втором извлечении будет вынут черный шар, обе равны $\frac{m}{n}$); поэтому $p_B(A) = p_A(B) = \frac{m-1}{n-1}$ (здесь $p_B(A)$ — вероятность того, что неизвестный нам первый вынутый шар был черным, если известно, что при втором извлечении оказался вынутым черный шар).

Наконец отметим, что поскольку одно из событий A и $\bar{A}$ обязательно имеет место, то сумма событий AB («B и A») и $\bar{A}B$ («B и $\bar{A}$») совпадает с событием B. А так как $p(AB) = p(A)\, p_A(B)$, $p(\bar{A}B) = p(\bar{A})\, p_{\bar{A}}(B)$ и $p(AB + \bar{A}B) = p(AB) + p(\bar{A}B)$ (события AB и $\bar{A}B$ несовместимы, ибо несовместимы A и $\bar{A}$), то

$$p(B) = p(A)\, p_A(B) + p(\bar{A})\, p_{\bar{A}}(B).$$

Так, в случае того же примера, что и выше,

$$p(A) = \frac{m}{n},\ p(\bar{A}) = \frac{n-m}{n},\ p_A(B) = \frac{m-1}{n-1},\ p_{\bar{A}}(B) = \frac{m}{n-1}$$

и $$p(A)\,p_A(B) + p(\bar{A})\,p_{\bar{A}}(B) = \frac{m}{n}\frac{m-1}{n-1} + \frac{n-m}{n}\frac{m}{n-1} =$$

$$= \frac{m}{n} = p(B).$$

Совершенно аналогично, если какой-либо опыт α может иметь k (и только k) попарно несовместимых исходов A_1, $A_2, \ldots, A_k$, то любое событие B можно представить в виде суммы событий $A_1B + A_2B + \ldots + A_kB$ и

$$p(B) = p(A_1)p_{A_1}(B) + p(A_2)\,p_{A_2}(B) + \ldots + p(A_k)\cdot p_{A_k}(B).$$

Эта формула называется формулой полной вероятности.

Задача 11. *В трех урнах находятся соответственно:*
1) 2 *белых и* 4 *черных шара*;
2) 4 *белых и* 2 *черных шара*;
3) 3 *белых и* 3 *черных шара.*

Из одной урны (неизвестно из какой) вынут наудачу шар. Какова вероятность того, что шар извлечен из первой урны, если он оказался: а) *белым*; б) *черным*?

Пусть событие A состоит в том, что вынутый шар оказался белым, а событие $\bar{A}$ — в том, что он оказался черным; далее, пусть B есть событие, которое состоит в том, что шар вынут из первой урны. Наш опыт извлечения одного шара может иметь $3 \cdot 6 = 18$ исходов (по числу шаров во всех трех урнах), которые мы считаем равновероятными (другими словами, мы считаем равновероятным, что шар вынут из любой из наших урн). Из этих 18 исходов событию A благоприятствуют 9, а из них событию B благоприятствуют 2. Событию $\bar{A}$ благоприятствуют тоже 9 из наших 18 исходов, но из них B благоприятствуют уже 4. Таким образом, имеем $p_A(B) = \frac{2}{9}$ и $p_{\bar{A}}(B) = \frac{4}{9}$.

Задача 12. *Слово* «папаха» *составлено из букв разрезной азбуки. Затем карточки с буквами тщательно перемешиваются и из них извлекаются по очереди и расклады-*

ваются в ряд какие-то четыре. Какова вероятность получить таким путем слово «папа»?

Пусть событие A состоит в том, что первой извлекается буква «*п*», событие B — в том, что второй извлекается буква «*а*», C — в том, что третьей извлекается снова «*п*» и D — в том, что четвертой буквой снова оказывается «*а*»; в таком случае то событие, вероятность которого нас интересует, можно записать как $ABCD$. Далее, применяя последовательно несколько раз формулу для вероятности произведения двух событий, имеем

$$p(A) = \frac{2}{6} = \frac{1}{3};$$

$$p(AB) = p(A)\,p_A(B) = \frac{1}{3}\cdot\frac{3}{5} = \frac{1}{5},$$

$$p(ABC) = p(AB)\,p_{AB}(C) = \frac{1}{5}\cdot\frac{1}{4} = \frac{1}{20}$$

и, наконец,

$$p(ABCD) = p(ABC)\,p_{ABC}(D) = \frac{1}{20}\cdot\frac{2}{3} = \frac{1}{30}.$$

Задача 13. *Имеется* 5 *урн, из которых две содержат по одному белому и по* 5 *черных шаров, одна урна —* 2 *белых и* 5 *черных шаров и, наконец, последние две урны — по* 3 *белых и по* 5 *черных шаров. Наудачу выбирается одна урна и из нее наудачу извлекается один шар. Какова вероятность того, что этот шар окажется белым*?

Обозначим через A_1, A_2 и A_3 события, состоящие в том, что шар извлечен из урны, содержащей один, или два, или три белых шара; в таком случае $p(A_1) = \frac{2}{5}$; $p(A_2) = \frac{1}{5}$ и $p(A_3) = \frac{2}{5}$. Далее, если B есть событие, состоящее в том, что извлекается белый шар, то по формуле полной вероятности имеем:

$$p(B) = p(A_1)\cdot p_{A_1}(B) + p(A_2)\cdot p_{A_2}(B) + p(A_3)\cdot p_{A_3}(B) = \frac{2}{5}\cdot\frac{1}{6} + \frac{1}{5}\cdot\frac{2}{7} + \frac{2}{5}\cdot\frac{3}{8} = \frac{23}{84}.$$

Приведем в заключение простой пример, иллюстрирующий применение данного на стр. 41 определения независимости случайных событий. Рассмотрим правильный тетраэдр из однородного материала, на трех гранях которого нанесены цифры *1*, *2* и *3*, а на четвертой — все эти три цифры одновременно (рис. 2). Через A

обозначим событие, состоящее в том, что подброшенный кверху тетраэдр упал на грань, на которой имеется цифра *1*; аналогично этому буквами B и C мы будем обозначать события, состоящие в падении тетраэдра на грань, имеющую на себе цифру *2*, соответственно, *3*. В таком случае ясно, что $p(A) = p(B) = p(C) = \frac{1}{2}$. Действительно, тетраэдр может упасть на каждую из своих граней с одинаковой вероятностью, а каждая из цифр имеется ровно на двух из четырех граней. Если теперь мы знаем, что событие A произошло, то это значит, что тетраэдр упал или на грань, на которой нанесена одна цифра *1*, или на грань, на которой имеются три цифры, *1*, *2* и *3*; при этом и событие B и событие C будут выполняться во втором случае и не будут выполняться в первом. Следовательно, здесь $p_A(B) = p_A(C) = \frac{1}{2}$, так что

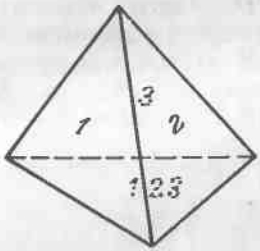


Рис. 2.

$$p_A(B) = p(B) \text{ и } p_A(C) = p(C),$$

т. е. и события A и B, и события A и C будут независимыми; соответственно этому и

$$p(AB) = p(A)\,p(B) = \frac{1}{4},\quad p(AC) = p(A)\,p(C) = \frac{1}{4}$$

(см. правило умножения вероятностей для независимых событий на стр. 29). Аналогично проверяется, что и события B и C являются независимыми: для них тоже $p_B(C) = p(C) = \frac{1}{2}$.

Из приведенного примера можно также сделать вывод, что из попарной независимости любых двух из трех событий A, B и C не вытекает еще независимость всех трех этих событий, т. е. выполнение равенства

$$p(ABC) = p(A)\,p(B)\,p(C)$$

(ср. стр. 30). В самом деле, ясно, что в нашем примере одновременное выполнение событий A и B уже влечет за собой выполнение события C, так что здесь

$$p_{AB}(C) = 1 \text{ и } p(ABC) = p(AB)\,p_{AB}(C) = \frac{1}{4}\cdot 1 = \frac{1}{4},$$

в то время как

$$p(A)\,p(B)\,p(C) = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{8}.$$

§ 4. Дисперсия случайной величины. Неравенство Чебышева и закон больших чисел

Важнейшей характеристикой случайной величины, бесспорно, является ее *среднее значение*. С помощью среднего значения мы можем сравнивать две случайные величины: так, например, из двух стрелков (см. задачу 4, стр. 25) лучшим естественно считать того, для которого среднее число выбиваемых им очков будет больше. Однако встречаются задачи, в которых знание одного лишь среднего значения случайной величины доставляет слишком мало данных об этой величине. Рассмотрим, например, пушку, ведущую прицельный огонь по мишени, удаленной от нее на расстояние a *км* (рис. 3). Если обозначить дальность полета снаряда через α (*км*), то среднее значение величины α, как правило, будет равно a; отклонение среднего значения от a свидетельствовало бы о наличии систематической погрешности стрельбы (систематического перелета или недолета снарядов), которую можно было бы устранить, изменив соответствующим образом наклон пушечного ствола. Однако отсутствие систематической ошибки нисколько не гарантирует высокую точность стрельбы: чтобы оценить точность, нам необходимо еще знать, насколько близко ложатся снаряды к цели (ибо равенство ср. зн. $\alpha = a$ лишь означает, что перелет снаряда встречается в среднем столь же часто, как и недолет).

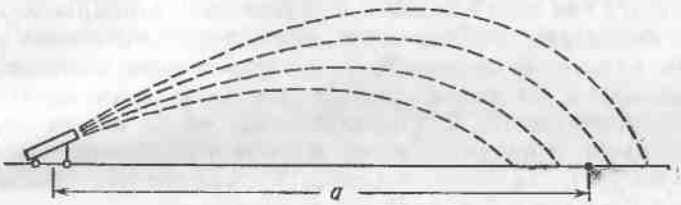


Рис. 3.

Как же определить точность стрельбы (и сравнить качество стрельбы по цели двух орудий)? Отклонение снаряда от цели задается числом $\alpha - a$; однако среднее значение величины $\alpha - a$ будет, очевидно, равно нулю:

$$\text{ср. зн. } (\alpha - a) = \text{ср. зн. } \alpha - a = a - a = 0,$$

что, впрочем, ясно, ибо в среднем положительные и отрицательные значения величины $\alpha - a$ взаимно сократятся. Разумеется, хорошей характеристикой «разброса» было бы среднее значение величины $|\alpha - a|$ (где вертикальные черточки, как всегда, обозначают а б с о л ю т н у ю в е л и ч и н у числа); однако математики не очень любят иметь дело с абсолютной величиной чисел, поскольку она плохо поддается дальнейшим алгебраическим преобразованиям. Поэтому принято характеризовать разброс случайной величины *средним значением квадрата ее отклонения от своего среднего значения*: ведь квадраты и положительных и отрицательных чисел всегда положительны, и никакого «сокращения» отклонений здесь не произойдет. Полученное таким образом число называется д и с п е р с и е й случайной величины α:

$$\text{дисп. } \alpha = \text{ср. зн. } (\alpha - a)^2 \ (= \text{ср. зн. } (\alpha - \text{ср. зн. } \alpha)^2).$$

Дисперсия случайной величины α и является наиболее распространенной мерой «разброса» (или отклонения от среднего значения) рассматриваемой случайной величины [1]). Ясно, что в случае орудий, ведущих прицельную стрельбу по мишени, лучшим следует считать орудие, для которого дисперсия величины α — длины полета снаряда — будет меньше (здесь мы считаем, что орудия уже отрегулированы так, что средняя длина полета снаряда совпадает с расстоянием a от орудия до цели).

Нетрудно понять, что для случайной величины α, характеризуемой таблицей вероятностей:

значения	a_1	a_2	...	a_k
вероятности	p_1	p_2	...	p_k

[1]) Ясно, что если — как в нашем примере — случайная величина α измеряется в *км*, то и ее среднее значение имеет размерность *км*, а дисперсия — размерность *км*^{2}. Поэтому наряду с дисперсией часто рассматривают число, равное *корню квадратному из дисперсии* случайной величины. Это число называется с р е д н и м к в а д р а т и ч н ы м у к л о н е н и е м случайной величины:

$$\text{ср. кв. укл. } \alpha = \sqrt{\text{дисп. } \alpha};$$

оно измеряется в тех же единицах, что и сама случайная величина α, и также служит мерой «разброса» ее значений.

среднее значение a равно

$$a = \text{ср. зн. } \alpha = p_1 a_1 + p_2 a_2 + \ldots + p_k a_k$$

(ср. выше, стр. 24), дисперсия найдется по формуле

$$\text{дисп. } \alpha = \text{ср. зн. } (\alpha - a)^2 =$$
$$= p_1 (a_1 - a)^2 + p_2 (a_2 - a)^2 + \ldots + p_k (a_k - a)^2.$$

Последнюю формулу можно записать и в несколько ином виде. Заметим, что

$$(\alpha - a)^2 = \alpha^2 - 2a\alpha + a^2.$$

Поэтому, поскольку среднее значение суммы (случайных) величин равно сумме их средних значений (см. стр. 38),

$$\text{дисп. } \alpha = \text{ср. зн. } (\alpha - a)^2 = \text{ср. зн. } (\alpha^2 - 2a\alpha + a^2) =$$
$$= \text{ср. зн. } \alpha^2 + \text{ср. зн. } (-2a\alpha) + \text{ср. зн. } a^2.$$

Но a^2 — это не случайная величина, а число, имеющее вполне определенное значение [1]); поэтому

$$\text{ср. зн. } a^2 = a^2.$$

С другой стороны, величина $-2a\alpha$ получается из случайной величины α умножением всех ее значений на число $-2a$; поэтому и ее среднее значение получается с помощью умножения среднего значения величины α на $-2a$:

$$\text{ср. зн. } (-2a\alpha) = -2a \cdot \text{ср.зн. } \alpha = -2a \cdot a = -2a^2.$$

Таким образом, окончательно получаем:

$$\text{дисп. } \alpha = \text{ср. зн. } \alpha^2 + \text{ср. зн. } (-2a\alpha) + \text{ср. зн. } a^2 =$$
$$= \text{ср. зн. } \alpha^2 - 2a^2 + a^2 = \text{ср. зн. } \alpha^2 - a^2 =$$
$$= \text{ср. зн. } (\alpha^2) - (\text{ср. зн. } \alpha)^2,$$

---

[1]) Можно, конечно, понимать a^2 как «случайную величину» со следующей «таблицей вероятностей»:

значения	a^2
вероятности	1

отсюда также вытекает, что

$$\text{ср. зн. } a^2 = 1 \cdot a^2 = a^2.$$

т. е. *дисперсия случайной величины равна среднему значению ее квадрата, уменьшенному на квадрат среднего значения*. Отсюда, в силу того, что дисперсия случайной величины всегда неотрицательна (ибо это есть среднее значение величины $(\alpha - a)^2$, все значения которой положительны), вытекает, что среднее значение квадрата случайной величины всегда не меньше квадрата ее среднего значения (ср. выше, стр. 40).

З а д а ч а 14. *Двум однотипным станкам отвечают следующие таблицы вероятностей (частот) того или иного числа бракованных изделий (на тысячу изделий)*:

1-й станок: число браков. изделий (на тысячу)	0	1	2	3	4
вероятности	0,1	0,2	0,4	0,2	0,1

2-й станок: число браков. изделий (на тысячу)	0	1	2	3	4
вероятности	0,15	0,2	0,25	0,3	0,1

Сравнить средние числа выпускаемых 1-м и 2-м станками бракованных изделий и дисперсии тех же величин.

Нетрудно видеть, что среднее число выпускаемых 1-м станком бракованных изделий (случайной величины α) и среднее число выпускаемых 2-м станком бракованных изделий (величины β) будет одинаковым:

$$\text{ср. зн. } \alpha = 0{,}1 \cdot 0 + 0{,}2 \cdot 1 + 0{,}4 \cdot 2 + 0{,}2 \cdot 3 + 0{,}1 \cdot 4 = 2$$

и

$$\text{ср. зн. } \beta = 0{,}15 \cdot 0 + 0{,}2 \cdot 1 + 0{,}25 \cdot 2 + 0{,}3 \cdot 3 + 0{,}1 \cdot 4 = 2;$$

с этой точки зрения оба станка можно считать равноценными. Однако дисперсия величины α меньше дисперсии β:

$$\text{дисп. } \alpha = 0{,}1 \cdot (0-2)^2 + 0{,}2 \cdot (1-2)^2 + 0{,}4 \cdot (2-2)^2 + 0{,}2 \cdot (3-2)^2 + 0{,}1 \cdot (4-2)^2 = 1{,}2,$$

а

$$\text{дисп. } \beta = 0{,}15 \cdot (0-2)^2 + 0{,}2 \cdot (1-2)^2 + 0{,}25 \cdot (2-2)^2 + 0{,}3 \cdot (3-2)^2 + 0{,}1 \cdot (4-2)^2 = 1{,}5.$$

Это значит, что продукция первого станка является более «стабильной»: здесь числа бракованных изделий в

разных тысячах выпущенных изделий более плотно группируются вокруг среднего значения 2, чем в случае второго станка.

Заметим теперь, что *дисперсия суммы двух независимых случайных величин* всегда *равна сумме их дисперсий.* В самом деле, пусть α и β — две независимые случайные величины, т. е. такие, что вероятности отдельных исходов одной из них нисколько не зависят от того, какие значения приняла в этот момент другая величина. В таком случае, как мы знаем (см. стр. 34—39),

если ср. зн. $\alpha = a$ и ср. зн. $\beta = b$, то ср. зн. $(\alpha + \beta) = a + b$ и ср. зн. $(\alpha\beta) = ab$.

Наряду с α и β рассмотрим еще случайные величины α^2 и β^2, значения которых равны квадратам значений величин α и β; для них также имеем:

$$\text{ср. зн. } (\alpha^2 + \beta^2) = \text{ср. зн. } \alpha^2 + \text{ср. зн. } \beta^2.$$

Далее,

$$\text{дисп. } \alpha = \text{ср. зн. } \alpha^2 - a^2; \quad \text{дисп. } \beta = \text{ср. зн. } \beta^2 - b^2$$

и

$$\begin{aligned}\text{дисп. } (\alpha + \beta) &= \text{ср. зн. } (\alpha + \beta)^2 - [\text{ср. зн. } (\alpha + \beta)]^2 = \\ &= \text{ср. зн. } (\alpha + \beta)^2 - (a + b)^2 = \\ &= \text{ср. зн. } (\alpha^2 + 2\alpha\beta + \beta^2) - (a^2 + 2ab + b^2).\end{aligned}$$

Но поскольку среднее значение суммы случайных величин равно сумме их средних значений, то

$$\text{ср. зн. } (\alpha^2 + 2\alpha\beta + \beta^2) = \text{ср. зн. } \alpha^2 + \text{ср. зн. } (2\alpha\beta) + \text{ср. зн. } \beta^2.$$

А так как случайная величина $2\alpha\beta$ в два раза больше случайной величины $\alpha\beta$, то

$$\text{ср. зн. } (2\alpha\beta) = 2 \text{ ср. зн. } (\alpha\beta) = 2ab.$$

Таким образом, окончательно получаем:

$$\begin{aligned}&\text{дисп. } (\alpha + \beta) = \\ &= (\text{ср. зн. } \alpha^2 + 2ab + \text{ср. зн. } \beta^2) - (a^2 + 2ab + b^2) = \\ &= (\text{ср. зн. } \alpha^2 + \text{ср. зн. } \beta^2) - (a^2 + b^2) = \\ &= (\text{ср. зн. } \alpha^2 - a^2) + (\text{ср. зн. } \beta^2 - b^2) = \text{дисп. } \alpha + \text{дисп. } \beta.\end{aligned}$$

Ясно, что и для произвольного числа попарно независимых случайных величин *дисперсия их суммы равна сумме их дисперсий*. Однако для не независимых случайных величин это будет уже не так. Пусть, например, α_1 и α_2 — это одна и та же случайная величина α со средним значением a; тогда $\alpha_1 + \alpha_2 = 2\alpha$. В этом случае, очевидно,

$$\text{ср. зн. } (2\alpha) = 2 \text{ ср. зн. } \alpha$$
$$(\text{т. е. ср. зн. } (\alpha_1 + \alpha_2) = \text{ср. зн. } \alpha_1 + \text{ср. зн. } \alpha_2).$$

Однако

$$\text{дисп. } (2\alpha) = 4 \text{ дисп. } \alpha$$
$$(\text{т. е. дисп. } (\alpha_1 + \alpha_2) = 2 \text{ дисп. } \alpha_1 + 2 \text{ дисп. } \alpha_2),$$

поскольку

$$\text{дисп. } (2\alpha) = \text{ср. зн. } [2\alpha - \text{ср. зн. } (2\alpha)]^2 = \text{ср. зн. } (2\alpha - 2a)^2 =$$
$$= \text{ср. зн. } [4(\alpha - a)^2] = 4 \text{ ср. зн. } (\alpha - a)^2 = 4 \text{ дисп. } \alpha.$$

Задача 15. *Предприятие выпускает какие-то изделия, причем для каждого отдельного изделия существует определенная вероятность p оказаться бракованным (скажем, $p = 0{,}002 = 0{,}2\%$). Считая, что все изделия из некоторой тысячи изделий независимо друг от друга могут оказаться бракованными с вероятностью p, найти среднее значение числа бракованных изделий на 1000 выпущенных изделий и дисперсию этой величины.*

Обозначим через α_i (где $i = 1, 2, 3, \ldots,$ или 1000) случайную величину, равную 1, если i-е изделие окажется бракованным, и 0 в противном случае; в таком случае все 1000 величин α_i имеют одну и ту же таблицу вероятностей:

значения	1	0
вероятности	p	$1-p$

Поэтому

$$\text{ср. зн. } \alpha_i = p \cdot 1 + (1 - p) \cdot 0 = p \ (= 0{,}002)$$

и

$$\text{дисп. } \alpha_i = \text{ср. зн. } \alpha_i^2 - (\text{ср. зн. } \alpha_i)^2 = [p \cdot 1 + (1 - p) \cdot 0] -$$
$$- p^2 = p - p^2 = p\,(1 - p) \ (= 0{,}002 \cdot 0{,}998 = 0{,}001996).$$

Интересующая же нас величина α равна сумме всех величин

$$\alpha = \alpha_1 + \alpha_2 + \alpha_3 + \ldots + \alpha_{1000},$$

причем, по предположению, все величины α_i взаимно независимы. Поэтому

$$\text{ср. зн. } \alpha = \text{ср. зн. } \alpha_1 + \text{ср. зн. } \alpha_2 + \ldots + \text{ср. зн. } \alpha_{1000} = 1000\,p \ (= 2)$$

и

$$\text{дисп. } \alpha = \text{дисп. } \alpha_1 + \text{дисп. } \alpha_2 + \ldots + \text{дисп. } \alpha_{1000} = 1000\,p\,(1-p) \ (= 1{,}996).$$

Приведенное решение задачи 15 использует то, что *среднее значение и дисперсия суммы n взаимно независимых случайных величин $\alpha_1, \alpha_2, \ldots, \alpha_n$ с одинаковым средним значением a и одинаковой дисперсией d равны n-кратным среднему значению и дисперсии одной величины α_1*:

$$\text{ср. зн. } (\alpha_1 + \alpha_2 + \ldots + \alpha_n) = n \text{ ср. зн. } \alpha_1 = na$$

и

$$\text{дисп. } (\alpha_1 + \alpha_2 + \ldots + \alpha_n) = n \text{ дисп. } \alpha_1 = nd.$$

В частности, *если α — число осуществлений некоторого события A в последовательности n взаимно независимых испытаний, причем вероятность осуществления A при каждом испытании равна p, то*

$$\text{ср. зн. } \alpha = np \text{ и дисп. } \alpha = np\,(1-p).$$

Из сказанного вытекает одно следствие, которое весьма часто оказывается полезным. Рассмотрим среднее арифметическое

$$\alpha_{\text{ср}} = \frac{\alpha_1 + \alpha_2 + \ldots + \alpha_n}{n}$$

n взаимно независимых случайных величин с одинаковым средним значением a и одинаковой дисперсией d. Так как все значения величины $\alpha_{\text{ср}}$ в n раз меньше соответствующих значений величины $\alpha_1 + \alpha_2 + \ldots + \alpha_n$, то среднее значение $\alpha_{\text{ср}}$ также в n раз меньше среднего значения

суммы $\alpha_1 + \alpha_2 + \ldots + \alpha_n$, т. е.

$$\text{ср. зн. } \alpha_{\text{ср}} = \frac{1}{n}(na) = a.$$

Дисперсия же величины $\alpha_{\text{ср}}$ в n^2 раз меньше дисперсии величины $\alpha_1 + \alpha_2 + \ldots + \alpha_n$ (ср. со сказанным на стр. 52 о дисперсиях величин α и 2α); поэтому

$$\text{дисп. } \alpha_{\text{ср}} = \frac{nd}{n^2} = \frac{d}{n}.$$

Таким образом, *среднее значение среднего арифметического n взаимно независимых случайных величин с одинаковыми средним значением и дисперсией равно среднему значению каждой из этих величин; дисперсия же среднего арифметического в n раз меньше дисперсии каждой из рассматриваемых случайных величин.*

Приведенный вывод можно проиллюстрировать на следующем примере. Пусть нам надо с возможно большей точностью определить значение какой-то физической величины α (для конкретности можно представлять себе, что речь идет, скажем, об определении некоторого расстояния на плоскости). Результат α одного измерения интересующей нас величины можно считать *случайной величиной*, ибо всегда существует определенная вероятность ошибки, связанная с неточностью измерительных приборов и недостаточной тщательностью измерения; при этом отсутствие с и с т е м а т и ч е с к о й ошибки измерения означает, что

$$\text{ср. зн. } \alpha = a$$

(ср. выше, стр. 47). Произведем теперь, скажем, 20 независимых измерений и образуем среднее арифметическое $\alpha_{\text{ср}}$ результатов $\alpha_1, \alpha_2, \ldots, \alpha_{20}$ этих измерений. При этом

$$\text{ср. зн. } \alpha_{\text{ср}} = \text{ср. зн. } \alpha = a,$$

т. е. значения величины $\alpha_{\text{ср}}$, так же как и значения величины α, группируются вокруг истинного значения a измеряемой величины. Однако, так как

$$\text{дисп. } \alpha_{\text{ср}} = \frac{1}{20} \text{ дисп. } \alpha,$$

то разброс значений $\alpha_{ср}$ является заметно меньшим, чем разброс значений α; поэтому приняв за a значение величины $\alpha_{ср}$, мы имеем все основания ожидать, что большая ошибка будет менее вероятной, чем в случае, когда за a принимается результат α одного измерения. Так, например, если мы измеряем на плоскости расстояние порядка 100 *м*, то ошибка в 1—2 *м* часто является вполне возможной; однако среднее арифметическое двадцати независимых измерений здесь почти наверное будет отличаться от истинного значения заметно меньше чем на 1 *м*.

Последнее замечание вплотную подводит нас к одному замечательному неравенству, вывод которого является основной целью этого параграфа. Так как дисп. $\alpha_{ср} <$ $<$ дисп. α, то мы предположили, что вероятность заметного отклонения величины $\alpha_{ср}$ от среднего значения a этой величины меньше вероятности большого отклонения α от числа $a =$ ср. зн. α. Это заключение можно строго обосновать, базируясь на следующем фундаментальном результате: *если α — это случайная величина со средним значением a и дисперсией d, то всегда*

$$P(|\alpha - a| > \varepsilon) < \frac{d}{\varepsilon^2}. \qquad (*)$$

Здесь ε — произвольное положительное число; запись же $P(|\alpha - a| > \varepsilon)$ означает вероятность того, что значение случайной величины α отклонится от среднего значения a той же величины больше чем на ε. Неравенство (*) называется неравенством Чебышева; оно показывает, что чем меньше дисперсия d случайной величины α, тем меньше вероятность значительного отклонения α от числа $a =$ ср. зн. α.

Неравенство Чебышева (*) представляет собой частный случай другого неравенства (также обычно называемого неравенством Чебышева), относящегося к произвольным случайным величинам β, принимающим только неотрицательные значения. А именно, *если β принимает только неотрицательные значения и среднее значение β равно b, то, каково бы ни было положительное число c,*

$$P(\beta > c) < \frac{b}{c}, \qquad (**)$$

где $P(\beta > c)$ — вероятность того, что случайная величина β примет значение, большее c. Ясно, что неравенство (*) вытекает из (**) — для того, чтобы в этом убедиться, надо только выбрать в качестве β неотрицательную случайную величину $(\alpha - a)^2$ (среднее значение которой по определению равно дисперсии d величины α) и заметить, что условие $|\alpha - a| > \varepsilon$ равносильно условию $(\alpha - a)^2 > \varepsilon^2$. Поэтому нам достаточно доказать (**).

Предположим, что таблица вероятностей величины β имеет вид

значения	b_1	b_2	b_3	...	b_n
вероятности	p_1	p_2	p_3	...	p_n

;

в таком случае

$$b = \text{ср. зн. } \beta = p_1 b_1 + p_2 b_2 + p_3 b_3 + \ldots + p_n b_n.$$

Будем считать, что возможные значения величины β перечислены в нашей таблице в порядке возрастания, так что $b_1 < b_2 < b_3 < \ldots < b_n$. Пусть b_k — первое из этих значений, которое превосходит c (т. е. значения $b_1, b_2, \ldots, b_{k-1}$ все меньше или равны c, а $b_k, b_{k+1}, \ldots, b_n$ — больше c); так как все значения β неотрицательны, то сумма в правой части предыдущего равенства не может увеличиться от того, что мы отбросим в ней слагаемые $p_1 b_1 + p_2 b_2 + \ldots + p_{k-1} b_{k-1}$. Следовательно,

$$b \geqslant p_k b_k + p_{k+1} b_{k+1} + \ldots + p_n b_n.$$

Заменим теперь в правой части полученного неравенства все значения $b_k, b_{k+1}, \ldots, b_n$ меньшим, чем они, числом c; при этом наша сумма только еще больше уменьшится и, значит,

$$b > p_k c + p_{k+1} c + \ldots + p_n c = (p_k + p_{k+1} + \ldots + p_n) c.$$

Таким образом мы приходим к неравенству

$$p_k + p_{k+1} + \ldots + p_n < \frac{b}{c},$$

которое в точности совпадает с нужным нам неравенством (**), поскольку сумма $p_k + p_{k+1} + \ldots + p_n$ вероятностей тех значений β, которые превосходят c, как раз и равна $P(\beta > c)$.

Вернемся теперь к случайной величине $\alpha_{ср}$, представляющей собой среднее арифметическое n независимых случайных величин $\alpha_1, \alpha_2, \ldots, \alpha_n$ с одним и тем же средним значением a и одной и той же дисперсией d:

$$\alpha_{ср} = \frac{\alpha_1 + \alpha_2 + \ldots + \alpha_n}{n}.$$

Выше мы видели, что

$$\text{ср. зн. } \alpha_{ср} = a \text{ и дисп. } \alpha_{ср} = \frac{d}{n}.$$

Применяя теперь к величине $\alpha_{ср}$ неравенство Чебышева (*), получаем

$$P(|\alpha_{ср} - a| \geqslant \varepsilon) < \frac{d}{n\varepsilon^2}. \qquad (***)$$

Так, например, пусть мы имеем 20 независимых измерений расстояния в 100 *м* (так что и среднее значение a результата каждого из этих измерений равно 100 *м*); предположим, что дисперсия каждого измерения близка к 2 ($м^2$). Иначе говоря, мы предполагаем, что квадрат ошибки каждого измерения в среднем равен 2, т. е. что абсолютная величина ошибки каждого измерения обычно имеет порядок 1—2 *м*. В таком случае формула (***) при $\varepsilon = 1$ (*м*), дает

$$P(|\alpha_{ср} - 100| > 1) < \frac{2}{20 \cdot 1^2} = 0{,}1.$$

Таким образом, вероятность того, что среднее арифметическое наших 20 измерений отклонится от истинного значения расстояния больше чем на 1 *м*, будет здесь заведомо меньше 0,1 [1]).

Отметим еще специально, что *если α — число осуществлений при n независимых испытаниях некоторого события A, вероятность осуществления которого при одном*

---

[1]) Следует еще иметь в виду, что неравенство Чебышева (*), так же как и следующее из него неравенство (***), являются весьма грубыми: реальная величина стоящих в левых частях этих неравенств вероятностей чаще всего оказывается намного меньшей стоящих справа значений. Так, например, применив более сложные методы, можно показать, что в рассмотренном нами примере величина $P(|\alpha_{ср} - 100| > 1)$ на самом деле будет меньше, чем 0,002.

испытании равна p, *то* (поскольку на стр. 53 было показано, что ср. зн. $\alpha = np$ и дисп. $\alpha = np(1-p)$) *при любом* $\varepsilon > 0$

$$P(|\alpha - np| > n\varepsilon) < \frac{np(1-p)}{(n\varepsilon)^2}$$

или, что то же самое,

$$P\left(\left|\frac{\alpha}{n} - p\right| > \varepsilon\right) < \frac{p(1-p)}{n\varepsilon^2}. \qquad (****)$$

Отсюда следует, что при любом (сколь угодно малом!) числе $\varepsilon > 0$ можно выбрать число n независимых испытаний столь большим, что вероятность $P\left(\left|\frac{\alpha}{n} - p\right| > \varepsilon\right)$ того, что частота $\frac{\alpha}{n}$ осуществлений события A в серии из n последовательных испытаний будет отличаться от вероятности p осуществления события A в одном испытании больше чем на ε, станет с к о л ь у г о д н о м а л о й. В самом деле, ведь при любых p и ε отношение $\frac{p(1-p)}{n\varepsilon^2}$, фигурирующее в правой части неравенства (****), стремится к нулю при $n \to \infty$; значит, при достаточно большом n оно будет сколь угодно мало. Но в жизни мы обычно пренебрегаем событиями достаточно малой вероятности, считая их «практически невозможными» (причем от того, насколько важно нам, чтобы мы не ошиблись в своем выводе, зависит то, начиная со сколь малой вероятности мы склонны считать, что соответствующее событие заведомо не произойдет). Поэтому последнее заключение означает, что *при любом положительном* ε *мы можем найти столь большое* N, *что неравенство* $n > N$ «*практически гарантирует*» *то, что отклонение частоты* $\frac{\alpha}{n}$ *от вероятности* p *будет меньше* ε. Это заключение, которое обосновывает высказанное в начале настоящей главы отождествление вероятностей случайных событий с их частотами, носит название з а к о н а б о л ь ш и х ч и с е л (поскольку оно связано с выбором большого числа N испытаний).

Аналогичный вывод можно сделать и из более общего чем (****), неравенства (***). А именно, из (***) вытекает, что для любого сколь угодно малого положитель-

ного числа ε мы всегда можем выбрать столь большое число n случайных величин $\alpha_1, \alpha_2, \ldots, \alpha_n$ (иначе говоря, выбрать столь большое число наблюдений или испытаний), чтобы оно гарантировало нам достаточную малость вероятности $P(|\alpha_{ср} - a| > \varepsilon)$. В самом деле, ведь при любом ε (и любом фиксированном значении d) правая часть $\frac{d}{n\varepsilon^2}$ неравенства (***) также стремится к нулю при неограниченном росте n. Таким образом, при любом $\varepsilon > 0$ мы можем при помощи выбора достаточно большого числа n гарантировать «практическую достоверность» неравенства $|\alpha_{ср} - a| < \varepsilon$. Общее утверждение о том, что *при достаточно большом числе однотипных независимых испытаний* (т. е. независимых испытаний, приводящих к результатам, имеющим одинаковое среднее значение и одинаковую дисперсию) *среднее арифметическое их результатов* $\alpha_1, \alpha_2, \ldots, \alpha_n$ *можно сделать сколь угодно близким к среднему значению* a *величин* $\alpha_1, \alpha_2, \ldots, \alpha_n$, также носит название з а к о н а б о л ь ш и х ч и с е л.

На самом деле, мы можем даже не требовать, чтобы участвующие в определении величины $\alpha_{ср}$ взаимно независимые случайные величины $\alpha_1, \alpha_2, \alpha_3, \ldots$ все имели о д и н а к о в ы е средние значения и одинаковые дисперсии. Действительно, если средние значения этих величин равны $a_1, a_2, a_3, \ldots$, а все дисперсии $d_1, d_2, d_3, \ldots$ ограничены (т. е. существует такое число D, что $d_i < D$ при всех i), то из неравенства Чебышева (*) следует, что

$$P(|\alpha_{ср} - a_{ср}| > \varepsilon) < \frac{D}{n\varepsilon^2}, \text{ где } a_{ср} = \frac{a_1 + a_2 + \ldots + a_n}{n}.$$

Отсюда, в свою очередь, вытекает, что *для любого числа* $\varepsilon > 0$ *мы можем, выбрав число* n *достаточно большим, «практически гарантировать» выполнение неравенства* $|\alpha_{ср} - a_{ср}| < \varepsilon$. Это утверждение представляет собой еще одну форму з а к о н а б о л ь ш и х ч и с е л.

§ 5. Алгебра событий и общее определение вероятности

В предыдущих параграфах у нас играли значительную роль две операции, сопоставляющие двум событиям A и B некоторое третье событие; эти операции мы назвали *суммой* и *произведением* событий A и B и обозначили через $A + B$ и AB (см. стр. 26 и 28). Некоторым основанием для этих названий служило то, что правила

«сложения» и «умножения» событий во многом напоминают правила сложения и умножения чисел. Так, из самого определения суммы и произведения событий следует, что $A + B = B + A$ и $AB = BA$; в одном месте мы воспользовались также равенством $(A + B)C = AC + BC$ (см. стр. 33). В настоящем параграфе мы более тщательно проанализируем сходство и отличие «алгебры событий» от «алгебры чисел».

В арифметике и алгебре рассматриваются числа разной природы — целые, рациональные, действительные (как рациональные, так и иррациональные), комплексные. Во всех случаях каждым двум числам a и b можно поставить в соответствие два других числа — их сумму $a + b$ и произведение ab. При этом правила, относящиеся к сложению, во многом напоминают правила, относящиеся к умножению; так, например,

$$a + b = b + a \quad \text{и} \quad ab = ba,$$
$$(a + b) + c = a + (b + c) \quad \text{и} \quad (ab)c = a(bc).$$

Эта аналогия между действиями сложения и умножения находит также отражение в существовании двух замечательных чисел 0 и 1, таких, что прибавление одного из них и умножение на второе не меняет никакого числа:

$$a + 0 = a \quad \text{и} \quad a \cdot 1 = a.$$

Однако аналогия между действиями сложения и умножения не простирается, к сожалению, особенно далеко. Причиной этого является несимметричный распределительный (дистрибутивный) закон

$$(a + b)c = ac + bc,$$

в записи которого сложение и умножение фигурируют совсем по-разному. Действительно, если заменить в последней формуле всюду знак сложения на знак умножения и наоборот, то мы придем к нелепому равенству

$$a \cdot b + c = (a + c) \cdot (b + c).$$

Поэтому многие свойства сложения и умножения весьма далеки друг от друга. Так, например, число 0 играет совершенно особую роль по отношению к умножению: эта особая роль определяется замечательным равенством

$$a \cdot 0 = 0$$

(из которого, в частности, вытекает, что деление отличного от нуля числа a на 0 невозможно); в противоположность этому аналогичное равенство, относящееся к сложению:

$$a + 1 = 1,$$

разумеется, не имеет места.

Существуют, однако, и отличные от чисел объекты, для которых можно определить операции сложения и умножения, обладающие многими обычными свойствами сложения и умножения чисел. При этом в некоторых случаях мы получаем алгебраические системы, в которых имеет место большая, чем в случае чисел, близость между определенными в этих системах действиями сложения и

умножения. Для примера рассмотрим совокупность всевозможных множеств («фигур») плоскости. *Сумму* $A + B$ двух множеств A и B естественно определить как их о б ъ е д и н е н и е (рис. 4, *а*). При этом, очевидно, будем иметь

$$A + B = B + A$$

и

$$(A + B) + C = A + (B + C).$$

(в последнем равенстве слева и справа стоит объединение трех множеств A, B и C, которое можно было бы обозначить и просто как $A + B + C$ без скобок). Роль нуля здесь будет играть так называемое «пустое» множество O, вовсе не содержащее точек; для такого множества имеем

$$A + O = A.$$

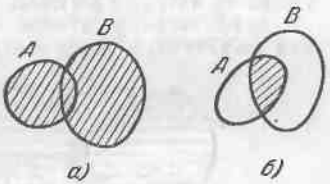


Рис. 4.

Определим теперь *произведение* AB двух множеств A и B как их общую часть или п е р е с е ч е н и е (рис. 4, *б*). В таком случае, очевидно,

$$AB = BA$$

и

$$(AB)C = A\,(BC)$$

(в последнем равенстве слева и справа стоит общая часть трех множеств A, B и C, которую естественно обозначить просто через ABC). Роль единицы здесь играет вся плоскость I. Действительно, для любого множества A имеем

$$AI = A.$$

Для определенной таким образом «алгебры множеств» имеет место также распределительный или дистрибутивный закон

$$(A + B)\cdot C = A\cdot C + B\cdot C,$$

для доказательства которого достаточно рассмотреть рис. 5, *а*, где заштрихованы двумя разными способами множества $A + B$ и C, так что их произведение (пересечение) $(A + B)\cdot C$ оказывается покрытым двойной штриховкой; цифрой I обозначено произведение $A\cdot C$ и цифрой II — произведение $B\cdot C$. Однако здесь имеет место также и «второй дистрибутивный закон»

$$A\cdot B + C = (A + C)\cdot(B + C),$$

получающийся из первого заменой сложения на умножение и наоборот. Для доказательства этого закона достаточно рассмотреть рис. 5, *б*, где двумя разными способами заштрихованы множества $A + C$ и $B + C$, так что их произведение $(A + C)\cdot(B + C)$ оказывается покрытым двойной штриховкой; цифрой I обозначено множество $A\cdot B$ и цифрой II — множество C.

Аналогия между этими двумя распределительными законами определяет полное сходство между правилами, относящимися к сложению множеств, и правилами, относящимися к их умножению. Так, например, здесь, очевидно,

$$A \cdot O = O \quad \text{и} \quad A + I = I;$$

можно сравнить также равенства

$$A \cdot A = A \quad \text{и} \quad A + A = A,$$

ни одно из которых не имеет места в алгебре чисел.

В арифметике и алгебре значительную роль играет сравнение чисел по величине. Если считать основным знаком сравнения знак

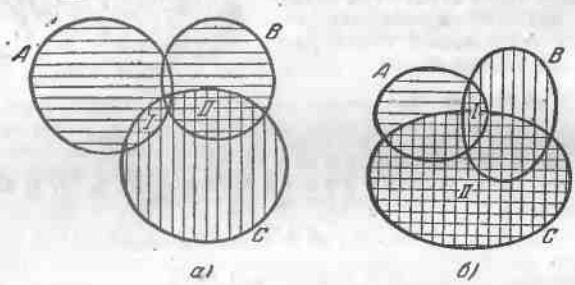


Рис. 5.

$\leqslant$ (соотношение $a \leqslant b$ означает, что число a *не больше* числа b), то основные правила, относящиеся к сравнению чисел, примут следующий вид:

$a \leqslant a$ (всякое число a не больше самого себя);

если $a \leqslant b$ и $b \leqslant a$, то $a = b$ (если a не больше b и b не больше a, то числа a и b равны);

если $a \leqslant b$ и $b \leqslant c$, то $a \leqslant c$ (если число a не больше b и b не больше c, то a не больше c).

Можно также ввести в рассмотрение *сравнение множеств*, условившись писать $A \subset B$ (этот знак $\subset$ заменяет «составной знак» $\leqslant$), если A есть часть множества B (могущая и совпасть со всем множеством B). Здесь тоже очевидно [1]),

[1]) Отметим одно существенное отличие между сравнением чисел и сравнением множеств. Для любых двух (действительных) чисел a и b имеет место одно из двух соотношений $a \leqslant b$ или $b \leqslant a$ (могут даже выполняться оба эти соотношения, если числа a и b равны). В противоположность этому для двух множеств A и B чаще всего не будет выполняться ни одно из двух соотношений $A \subset B$ и $B \subset A$. (Подобное положение будет иметь место и для комплексных чисел, если условиться, как это иногда делают, писать $a \leqslant b$ в том случае, когда комплексные числа a и b имеют одинаковый аргумент и модуль числа a не больше модуля числа b.)

$$A \subset A;$$

если $A \subset B$ и $B \subset A$, то $A = B$;

если $A \subset B$ и $B \subset C$, то $A \subset C$.

Из других правил, относящихся к сравнению множеств, заслуживают упоминания правила:

$$A \subset A + B \text{ и } AB \subset A,$$

а также

$$A \subset I \text{ и } O \subset A$$

(последнее соотношение означает, что пустое множество O не содержит точек, отличных от точек множества A,— это верно при любом A, ибо O вовсе не содержит никаких точек).

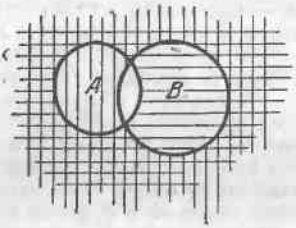


Рис. 6.

Существенное отличие алгебры множеств от алгебры чисел заключается в наличии в алгебре множеств еще одной операции, ставящей в соответствие каждому множеству A новое множество $\overline{A}$ (*дополнение* A). Эта операция определяется следующим образом: $\overline{A}$ состоит из всех точек плоскости не принадлежащих множеству A. Основные правила, относящиеся к этой новой операции таковы:

$$A + \overline{A} = I \text{ и } A\overline{A} = O;$$

$$\overline{O} = I \text{ и } \overline{I} = O;$$

$$\overline{\overline{A}} = A;$$

если $A \subset B$, то $\overline{B} \subset \overline{A}$;

и, наконец,

$$\overline{A + B} = \overline{A} \cdot \overline{B} \text{ и } \overline{A \cdot B} = \overline{A} + \overline{B}$$

(см. рис. 6, на котором по-разному заштрихованы множества $\overline{A}$ и $\overline{B}$, причем дважды заштрихованным оказалось множество $\overline{A + B}$, а хоть один раз — множество $\overline{A} \cdot \overline{B}$).

Существует также много других совокупностей некоторых объектов, для которых естественно определяются понятия суммы, произведения, а также «упорядочение» $A \subset B$ и «дополнение» $\overline{A}$, удовлетворяющие всем перечисленным выше алгебраическим свойствам. Одним из примеров таких совокупностей является рассмотренная в §§ 1—3 совокупность случайных событий: алгебра событий, как легко видеть, обладает всеми теми же свойствами, что и алгебра множеств. Другие примеры можно получить, рассматривая вместо множеств точек плоскости множества элементов какой-либо другой природы, например, множества целых чисел. Если при этом под суммой и произведением множеств A и B по-прежнему понимать их объединение и пересечение (например, если A_2 и A_3 — множеств а

чисел, делящихся, соответственно, на 2 и на 3, то в множество $A_2 + A_3$ входят все четные числа и те из нечетных чисел, которые делятся на 3, а множество A_2A_3 состоит из всех целых чисел, кратных 6) и считать, что $A \subset B$, если A составляет часть B (например, $A_4 \subset A_2$, где A_4 — множество чисел, делящихся на 4), и что $\overline{A}$ — это множество всех целых чисел, *не принадлежащих* A (если A — множество всех простых чисел, то $\overline{A}$ содержит все составные числа и число 1), а под I и O понимать соответственно множество *всех* целых чисел и множество, не содержащее *ни одного* числа, то сохранят силу *все* выписанные выше соотношения.

В качестве еще одного примера можно рассмотреть множество всех делителей некоторого числа N, не делящегося ни на какой полный квадрат, больший 1 (в частности, при $N = 30$— множество чисел 1, 2, 3, 5, 6, 10, 15 и 30); если под $A + B$ и AB понимать соответственно *общее наименьшее кратное* и *общий наибольший делитель* чисел A и B, под $A \subset B$ — соотношение «A есть *делитель* B» и обозначить через O и I — числа 1 и N (т. е. 1 и 30) и через $\overline{A}$ — число N/A (в нашем случае $30/A$), то по-прежнему

$$A + B = B + A \quad \text{и} \quad AB = BA,$$

$$(A + B)\cdot C = A\cdot C + B\cdot C \quad \text{и} \quad A\cdot B + C = (A + C)\cdot(B + C),$$

$$A \subset A + B \quad \text{и} \quad AB \subset A,$$

$$\overline{A + B} = \overline{A}\cdot\overline{B} \quad \text{и} \quad \overline{A\cdot B} = \overline{A} + \overline{B}$$

и т. д.

Наконец, весьма важный пример того же рода составляет множество всех логических предложений (т. е. утверждений, относительно каждого из которых имеет смысл говорить о том, что оно истинно или что оно ложно); это множество составляет предмет изучения *математической логики*. Здесь под суммой $A + B$ и произведением $A\cdot B$ предложений A и B следует понимать предложения «или A, или B» и соответственно «и A, и B», под $A \subset B$ — то, что из истинности предложения A следует и истинность предложения B (короче: «из A следует B»), под $\overline{A}$ — отрицание предложения A (предложение «A неверно»), под I и O — предложение, которое заведомо истинно, и предложение, которое заведомо ложно. В таком случае снова выполняются все вышеописанные соотношения, которые выражают определенные законы логики. Так, например,

$$A + \overline{A} = I$$

есть *закон исключенного третьего*: во всех случаях предложение A либо истинно, либо ложно; соотношение

$$A\cdot\overline{A} = O$$

есть *закон противоречия*: ни в каком случае предложение A не может одновременно оказаться и истинным и ложным.

Многочисленность и важность алгебраических систем, обладающих всеми перечисленными выше свойствами, заставила математиков специально заняться их изучением. В настоящее время такие системы называются *алгебрами Буля*, по имени

известного английского математика и логика XIX в. Джорджа Буля, впервые применившего такую алгебру в своих исследованиях по логике [1]).

Элементы алгебры Буля, вообще говоря, не являются числами. Однако часто удается поставить в соответствие каждому элементу A число $|A|$ или $p(A)$, удовлетворяющее следующим условиям:

$$0 \leqslant p(A) \leqslant 1;\ p(O) = 0,\ p(I) = 1;$$

$$\text{если } A \subset B, \text{ то } p(A) \leqslant p(B);$$

$$\text{если } A \cdot B = O, \text{ то } p(A + B) = p(A) + p(B).$$

Это число называют абсолютной величиной элемента A или его нормой, а саму алгебру Буля в этом случае называют нормированной. В качестве примеров можно привести семейство плоских фигур, принадлежащих квадрату со стороной единица (сам квадрат играет роль элемента I этой алгебры Буля), где за абсолютную величину или норму фигуры A принята ее площадь, или множество всех делителей не делящегося ни на какой квадрат целого числа N (например, числа 30), где под нормой числа A понимается $\log_N A$ (в нашем случае $\log_{30} A$); совокупность всех предложений математической логики также можно рассматривать как нормированную алгебру Буля, если условиться считать абсолютную величину (норму) предложения равной 1, если это предложение истинно, и равной 0, если оно ложно. Примером нормированной алгебры Буля является и та алгебра событий, которая изучалась в §§ 1—3; здесь роль абсолютной величины или нормы события A играет вероятность $p(A)$ этого события.

Связь теории вероятностей с алгебрами Буля может быть положена в основу общего определения самого предмета этой науки. А именно, можно сказать, что *теория вероятностей изучает совокупности объектов, образующие нормированную алгебру Буля*; эти объекты называются событиями, а норма $p(A)$ события A называется вероятностью. Так, например, в «задаче об урне» (или в любой сводящейся к ней задаче) фактически рассматривается нормированная алгебра Буля всевозможных множеств,

---

[1]) Алгебру Буля можно охарактеризовать как совокупность элементов, в которой определены две операции $\overline{A}$ и $A + B$ (сопоставляющие каждому элементу A, соответственно паре элементов A и B, какой-то элемент того же множества), обладающие свойствами:

$$A + B = B + A,$$

$$(A + B) + C = A + (B + C),$$

$$\overline{\overline{A} + \overline{B}} + \overline{\overline{A} + B} = A.$$

Все остальные свойства алгебр Буля могут уже быть выведены из этих трех основных свойств, если определить «произведение» AB как $\overline{\overline{A} + \overline{B}}$, соотношение $A \subset B$ — равенством $A + B = B$, элементы I и O — равенствами $A + \overline{A} = I$ и $A\overline{A} = O$ (где A — любое).

которые можно составить из заданных n элементов («точек»). При этом сумма и произведение двух множеств здесь (как и во всех примерах ниже) определяются как их объединение и пересечение; норма же задается условием, что для всех множеств из одного элемента (т. е. отдельных точек) она равна одному и тому же числу $\frac{1}{n}$. Столь же законными, однако, с нашей новой точки зрения будут задачи, возникающие из рассмотрения той же алгебры Буля, но при более общем условии, что нормы отдельных точек равны произвольным положительным числам $p_1, p_2, \ldots, p_n$, удовлетворяющим единственному условию $p_1 + p_2 + \ldots + p_n = 1$ (к алгебре Буля такого типа с $n = 6$ сводится, в частности, задача о «неправильной» игральной кости — имеющей искаженную форму или изготовленной из неоднородного материала). Ниже нам встретится также случай, в котором элементами алгебры Буля являются всевозможные части заданного отрезка AB, а норма определяется как отношение длины рассматриваемой части ко всей длине отрезка AB (см. задачу 22 на стр. 112); аналогично этому иногда приходится рассматривать совокупность всех множеств, принадлежащих некоторой плоской фигуре или пространственному телу, и задавать норму, как отношение площади или объема соответствующего множества к площади всей фигуры или объему всего тела (ср., например, «задачи на геометрические вероятности» на стр. 45—50 книги [37]). Можно также обобщить на все эти случаи «задачу о неправильной кости», т. е. и при рассмотрении алгебры Буля всех множеств, принадлежащих заданному отрезку, фигуре или телу, вводить норму совершенно произвольным образом, требуя лишь, чтобы она удовлетворяла условиям, наложенным выше на функцию $p(A)$; при этом мы придем к новому широкому классу интересных теоретико-вероятностных задач.

Если принять набранное на стр. 65 курсивом утверждение в качестве определения предмета теории вероятностей, то отсюда вытекает, что в любой относящейся к этой теории задаче исходная алгебра Буля обязательно должна быть задана заранее (т. е. так или иначе указана в самом условии задачи). Основной задачей теории вероятностей при этом следует считать нахождение вероятностей составных событий, образованных из заданных нам основных или первоначальных событий $A, B, C, D, \ldots$ при помощи операций булевской алгебры (например, событий $AB + BC + CA$ или $(A + B \cdot C)(A + D)$) по вероятностям этих основных событий (подобно тому, как основная задача геометрии состоит в вычислении каких-либо расстояний или углов по другим, исходным, расстояниям и углам, предполагающимся известными — например, длины гипотенузы прямоугольного треугольника по известным длинам катетов). При таком подходе к теории вероятностей (указанном впервые в 1917 г. С. Н. Бернштейном) очень важный вопрос о способах нахождения основных вероятностей $p(A)$, $p(B)$ и т. д. остается, разумеется открытым; однако для того, чтобы развиваемая теория имела практическое значение, эти вероятности обязательно должны задаваться так, чтобы они совпадали с эмпирическими частотами соответствующих событий в длинной серии опытов. Один возможный путь определения «основных вероятностей»,

удовлетворяющий этому условию, дает приведенное в § 1 «классическое определение вероятности», опирающееся на понятие о «полной системе равновероятных исходов опыта»; в других случаях, когда такой полной системы не существует, для определения значений $p(A)$ приходится привлекать другие соображения (например, непосредственно находить приближенное значение $p(A)$ с помощью многократного выполнения опыта, с которым связано осуществление события A). Суть дела, однако, заключается в том, что методы определения исходных вероятностей нисколько не отражаются на всех дальнейших операциях с ними, составляющих основное содержание теории.

Заметим еще, что то обстоятельство, что во всех приведенных выше примерах алгебра Буля задавалась как совокупность множеств, составленных из точек одного «наибольшего множества», не является случайным — такое задание этой алгебры возможно во в с е х теоретико-вероятностных задачах. Исходя отсюда, можно даже с само[illegible] начала считать основным объектом изучения теории вероятностей не нормированную алгебру Буля всевозможных событий, а некоторое «полное множество элементарных событий», различные части («подмножества») которого и отождествляются затем с «событиями». Для того, чтобы сделать эти рассуждения вполне законченными, надо только сопоставить еще подмножествам A нашего «множества всех элементарных событий» определенную норму $p(A)$ и перечислить основные требования (аксиомы), которым должны удовлетворять сами рассматриваемые подмножества и их нормы, чтобы мы действительно имели нормированную алгебру Буля. Такой метод аксиоматического построения теории вероятностей (предложенный в 1929 г. А. Н. Колмогоровым) обладает определенными преимуществами перед методом, изложенным выше в настоящем параграфе, при исследовании более сложных и тонких вопросов теории и поэтому он является в настоящее время наиболее распространенным; более подробное его изложение увело бы нас, однако, слишком далеко в сторону от нашей основной темы.

Глава II

ЭНТРОПИЯ И ИНФОРМАЦИЯ

§ 1. Энтропия как мера степени неопределенности

Главным свойством случайных событий, изучение которых составляет основной предмет этой книжки, является отсутствие полной уверенности в их наступлении, создающее известную неопределенность при выполнении связанных с этими событиями опытов. Однако совершенно ясно, что степень этой неопределенности в различных случаях будет совершенно разной. Если наш опыт состоит в определении цвета первой встретившейся нам вороны, то мы можем почти с полной уверенностью рассчитывать, что этот цвет будет черным — хотя зоологи и утверждают, что встречаются иногда белые вороны, вряд ли кто-нибудь усомнится в исходе такого опыта. Несколько менее определенен опыт, состоящий в выяснении того, окажется ли первый встреченный нами человек левшой или нет — здесь тоже предсказать результат опыта можно почти не колеблясь, но опасения относительно правильности этого предсказания будут более обоснованны, чем в первом случае. Значительно труднее предсказать заранее, будет ли первый встретившийся нам на улице города человек мужчиной или женщиной. Но и этот опыт имеет относительно небольшую степень неопределенности по сравнению, например, с попыткой заранее указать победителя в турнире с двадцатью совершенно незнакомыми нам участниками или определить номер лотерейного билета, на который выпадет наибольший выигрыш в предстоящем тираже лотереи: если, скажем, предсказав, что первый встреченный нами на улице человек будет мужчиной, мы еще можем надеяться угадать, то вряд ли кто-нибудь рискнет сделать прогноз в предпоследнем или, тем более, в последнем случае.

Для практики важно уметь численно оценивать *степень неопределенности* самых разнообразных опытов, чтобы иметь возможность сравнить их с этой

стороны. Мы начнем здесь с рассмотрения опытов, имеющих k равновероятных исходов. Очевидно, что степень неопределенности каждого такого опыта определяется числом k: если при $k = 1$ исход опыта вообще не является случайным, то при большом k, т. е. при наличии большого числа разных исходов, предсказание результата опыта становится весьма затруднительным. Таким образом, совершенно ясно, что искомая численная характеристика степени неопределенности должна зависеть от k, т. е. являться функцией $f(k)$ числа k. При этом для $k = 1$ эта функция должна обращаться в нуль (ибо в этом случае неопределенность полностью отсутствует), а при возрастании числа k она должна возрастать.

Для более полного определения функции $f(k)$ надо предъявить к ней дополнительные требования. Рассмотрим два независимых опыта α и β (т. е. такие два опыта, что любые сведения об исходе первого из них никак не меняют вероятностей исходов второго). Пусть опыт α имеет k равновероятных исходов, а опыт β имеет l равновероятных исходов; рассмотрим также сложный опыт $\alpha\beta$, состоящий в одновременном выполнении опытов α и β. Очевидно, что неопределенность опыта $\alpha\beta$ больше неопределенности опыта α, так как к неопределенности α здесь добавляется еще неопределенность исхода опыта β. Естественно считать, что *степень неопределенности опыта* $\alpha\beta$ *равна сумме неопределенностей, характеризующих опыты* α *и* β. А так как опыт $\alpha\beta$ имеет, очевидно, kl равновероятных исходов (они получаются, если комбинировать каждый из k возможных исходов опыта α с l исходами β), то мы приходим к следующему условию, которому должна удовлетворять наша функция $f(k)$:

$$f(kl) = f(k) + f(l).$$

Последнее условие наталкивает на мысль *принять за меру неопределенности опыта, имеющего* k *равновероятных исходов, число* $\log k$ (ибо $\log(kl) = \log k + \log l$). Такое определение меры неопределенности согласуется также с условиями, что при $k = 1$ она равна нулю и что при возрастании k она возрастает [1]).

[1]) Нетрудно показать, что логарифмическая функция является единственной функцией аргумента k, удовлетворяющей условиям $f(kl) = f(k) + f(l)$, $f(1) = 0$ и $f(k) > f(l)$ при $k > l$ (ср. ниже § 4, стр. 134 — 135).

Заметим, что выбор основания системы логарифмов здесь несуществен, так как в силу известной формулы

$$\log_b k = \log_b a \cdot \log_a k$$

переход от одной системы логарифмов к другой сводится лишь к умножению функции $f(k) = \log k$ на постоянный множитель (м о д у л ь п е р е х о д а $\log_b a$), т. е. равносилен простому изменению *единицы измерения* степени неопределенности. В конкретных применениях «меры степени неопределенности» обычно используются логарифмы при основании д в а (другими словами — считается, что $f(k) = \log_2 k$). Это означает, что за единицу измерения степени неопределенности здесь принимается неопределенность, содержащаяся в опыте, имеющем *два* равновероятных исхода (например, в опыте, состоящем в подбрасывании монеты и выяснении того, какая сторона ее оказалась сверху, или в выяснении ответа «да» или «нет» на вопрос, по поводу которого мы с равными основаниями можем ожидать, что ответ будет утвердительным или отрицательным). Такая единица измерения неопределенности называется д в о и ч н о й е д и н и ц е й (сокращенно дв. ед.) или б и т о м [1]); в немецкой литературе используется также выразительное ее название «Ja-Nein Einheit» («да-нет единица»). Подобная «да-нет единица» является в каком-то смысле самой естественной; дополнительные соображения, указывающие, почему именно ей отдается предпочтение в технике, станут ясны из содержания гл. IV этой книги. Мы тоже в дальнейшем будем все время пользоваться двоичными единицами (битами); таким образом запись $\log k$ (где мы, как правило, не будем указывать основания системы логарифмов) будет обычно означать $\log_2 k$. Заметим только, что в содержании книги практически ничего не изменилось бы, если бы мы использовали более привычные д е с я т и ч н ы е л о г а р и ф м ы; это лишь означало бы, что за единицу степени неопределенности принимается неопределенность опыта, имеющего 10 равновероятных исходов (таким является, например, опыт, состоящий в извлечении шара из урны с десятью

[1]) Английское слово *bit* было образовано с помощью сжатия слов *binary digit*, означающих «двоичная цифра» или «двоичная единица».

перенумерованными шарами, или опыт по отгадыванию одной цифры, если любая из десяти цифр имеет одинаковую вероятность быть загаданной). Эта последняя единица степени неопределенности (которую называют *десятичной единицей* или *дитом*) примерно в $3\frac{1}{3}$ раза крупнее двоичной единицы (ибо $\log_2 10 \approx 3{,}32 \approx \approx 3\frac{1}{3}$).

Таблица вероятностей для опыта, имеющего k равновероятных исходов, имеет вид:

исходы опыта	A_1	A_2	A_3	...	A_k
вероятности	$\frac{1}{k}$	$\frac{1}{k}$	$\frac{1}{k}$	...	$\frac{1}{k}$

Так как общая неопределенность опыта по нашему условию равна $\log k$, то можно считать, что каждый отдельный исход, имеющий вероятность $\frac{1}{k}$, вносит неопределенность, равную $\frac{1}{k}\log k = -\frac{1}{k}\log\frac{1}{k}$. Но тогда естественно считать, что в результат опыта, таблица вероятностей для которого имеет вид

исходы опыта	A_1	A_2	A_3
вероятности	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$

исходы A_1, A_2 и A_3 вносят неопределенность, равную соответственно $-\frac{1}{2}\log\frac{1}{2}$, $-\frac{1}{3}\log\frac{1}{3}$ и $-\frac{1}{6}\log\frac{1}{6}$, так что общая неопределенность этого опыта равна

$$-\frac{1}{2}\log\frac{1}{2} - \frac{1}{3}\log\frac{1}{3} - \frac{1}{6}\log\frac{1}{6}.$$

Аналогично этому можно положить, что в самом общем случае, *для опыта α с таблицей вероятностей*

исходы опыта	A_1	A_2	A_3	...	A_k
вероятности	$p(A_1)$	$p(A_2)$	$p(A_3)$	...	$p(A_k)$

мера неопределенности равна

$$-p(A_1)\log p(A_1) - p(A_2)\log p(A_2) - \\ - p(A_3)\log p(A_3) - \ldots - p(A_k)\log p(A_k)$$

(см. также § 4 этой главы, напечатанный мелким шрифтом). Это последнее число мы, руководствуясь некоторыми глубокими физическими аналогиями, несущественными, впрочем, для всего дальнейшего, будем называть *энтропией* опыта α и обозначать через $H(\alpha)$ [1]).

Перейдем к изучению свойств энтропии $H(\alpha)$. Отметим, прежде всего, что она не может принимать отрицательных значений: так как всегда $0 \leqslant p(A) \leqslant 1$, то $\log p(A)$ не может быть положительным, а $-p(A)\log p(A)$ — отрицательным. Заметим далее, что если p очень мало, то и произведение $-p\log p$ будет весьма малым, хотя $-\log p$ здесь и будет большим положительным числом. В самом деле, пусть, например, $p = \frac{1}{2^n}$; в таком случае $\log p = -n$ и $-p\log p = \frac{n}{2^n}$, а дробь $\frac{n}{2^n}$ при большом n (что соответствует малому $p = \frac{1}{2^n}$) будет очень маленькой (ибо с ростом n число 2^n растет несравненно быстрее, чем само число n — так, например, число 2^{64} состоит из 20 цифр [2])!). Отсюда вытекает, что при $p \to 0$ произведение $-p\log p$ неограниченно убывает, так что

$$\lim_{p\to 0}(-p\log p) = 0$$

(ср. ниже рис. 7 и 9, на которых изображен график функции $y = -p\log p$: из графика видно, что при $p = 0$

---

[1]) Относительно отношения введенного здесь понятия энтропии к термодинамическому понятию энтропии, играющему важную роль в физике, см., например, книги И. А. Полетаева [10] и, особенно, Л. Бриллюэна [2].

[2]) С этим связана известная, видимо, многим из читателей этой книги легенда об изобретателе шахмат, который в качестве награды попросил, чтобы ему выдали столько хлебных зерен, сколько получится, если положить на 1-ю клетку шахматной доски одно зерно, на 2-ю — два и далее на каждую клетку доски вдвое больше зерен, чем на предшествующую. Эта награда первоначально показалась обещавшему ее шаху очень скромной; однако на самом деле соответствующее количество зерен (равное $2^{64} - 1$) намного превосходит все имеющиеся на земле запасы зерна.

значение этой функции равно нулю). Поэтому если вероятность $p(A_i)$ исхода A_i равна нулю (т. е. исход A_i невозможен), то соответствующий член $-p(A_i)\log p(A_i)$ в выражении для энтропии можно просто отбросить (строго говоря, этот член не имеет смысла, так как $\log p(A_i)$ в этом случае не существует; именно поэтому нам и пришлось искать п р е д е л выражения $-p\log p$ при $p \to 0$). В обратном случае, когда $p(A_i)$ очень велико (т. е. близко к 1), член $-p(A_i)\log p(A_i)$ также будет очень мал, так как $\log p$ при $p \to 1$ стремится к нулю; если вероятность $p(A_i)$ точно равна единице (т. е. появление исхода A_i нашего опыта является достоверным событием), то $\log p(A_i) = 0$ и, значит, также $-p(A_i)\log p(A_i) = 0$ (см. снова рис. 7 и 9).

Так как $-p\log p$ равно нулю лишь при $p = 0$ или $p = 1$, то ясно, что *энтропия $H(\alpha)$ опыта α равна нулю лишь в том случае, когда одна из вероятностей $p(A_1)$, $p(A_2)$, . . ., $p(A_k)$ равна единице, а все остальные равны нулю* (напоминаем, что $p(A_1) + p(A_2) + \ldots + p(A_k) = 1$; см. выше стр. 27). Это обстоятельство хорошо согласуется со смыслом величины $H(\alpha)$ как меры степени неопределенности: действительно, только в этом случае опыт вообще не содержит никакой неопределенности.

Далее, естественно считать, что среди всех опытов, имеющих k исходов, н а и б о л е е н е о п р е д е л е н н ы м является опыт с таблицей вероятностей:

исходы опыта	A_1	A_2	A_3	. . .	A_k
вероятности	$\frac{1}{k}$	$\frac{1}{k}$	$\frac{1}{k}$	. . .	$\frac{1}{k}$

который мы обозначим через α_0: в этом случае предсказать исход опыта труднее всего. Этому отвечает то обстоятельство, что опыт α_0 имеет н а и б о л ь ш у ю энтропию: *если α — произвольный опыт, имеющий k исходов A_1, A_2, . . ., A_k, то*

$$H(\alpha) = -p(A_1)\log p(A_1) - p(A_2)\log p(A_2) - \ldots - p(A_k)\log p(A_k) \leqslant$$

$$\leqslant \log k = \underbrace{-\frac{1}{k}\log\frac{1}{k} - \frac{1}{k}\log\frac{1}{k} - \ldots - \frac{1}{k}\log\frac{1}{k}}_{k \text{ раз}} = H(\alpha_0),$$

причем равенство достигается только в том случае, когда $p(A_1) = p(A_2) = \ldots = p(A_k) = \frac{1}{k}$. Полное доказательство этого результата мы пока отложим (см. Приложение I в конце книги); здесь же мы ограничимся тем, что поясним соответствующую теорему на примере, когда

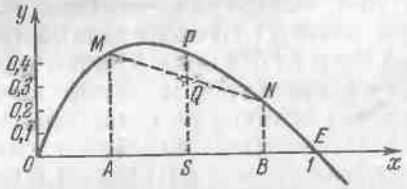


Рис. 7.

$k = 2$. В этом случае она сводится к доказательству следующего неравенства:

$$-p(A_1)\log p(A_1) - p(A_2)\log p(A_2) \leqslant \log 2 = 1.$$

Как мы уже отмечали, значение функции $F(x) = -x\log x$ при $x \to 0$ стремится к нулю; с другой стороны, при $x = 1$ ее значение также равно нулю, а при $0 \leqslant x \leqslant 1$ эта функция положительна (ибо в этом случае $\log x$ отрицателен); при $x > 1$ функция $-x\log x$ отрицательна. График рассматриваемой функции изображен на рис. 7, где $OE = 1$, $OA = p(A_1)$, $OB = p(A_2)$ и отрезки AM и BN изображают величины $-p(A_1)\log p(A_1)$ и $-p(A_2)\log p(A_2)$. Так как

$$OA + OB = p(A_1) + p(A_2) = 1 = OE,$$

то расстояние OS от начала до середины S отрезка AB равно $\frac{1}{2}$; поэтому на рис. 7 отрезок SP равен $-\frac{1}{2}\log\frac{1}{2} = \frac{1}{2}$. Но полусумма отрезков AM и BN равна средней линии SQ трапеции $ABNM$, которая не превосходит SP; следовательно,

$$\frac{1}{2}\left(-p(A_1)\log p(A_1) - p(A_2)\log p(A_2)\right) \leqslant \frac{1}{2},$$

т. е.

$$-p(A_1)\log p(A_1) - p(A_2)\log p(A_2) \leqslant 1,$$

где равенство имеет место лишь в том случае, когда отрезки OA и OB оба совпадают с OS. Итак, мы доказали, что функция

$$h(p) = -p \log p - (1-p) \log (1-p),$$

определяющая энтропию опыта с двумя исходами (вероятности которых равны p и $1-p$), принимает наибольшее значение (равное $\log 2 = 1$) при $p = \frac{1}{2}$. На рис. 8 изображен график этой функции, показывающий, как меняется энтропия $h(p)$ при изменении p от 0 до 1.

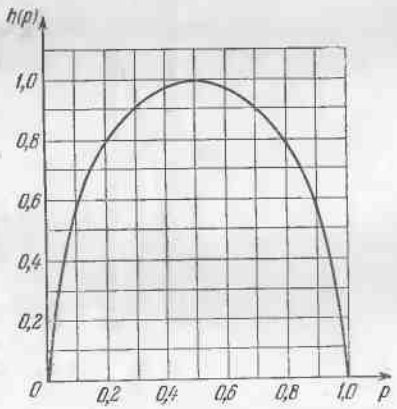


Рис. 8.

В случае опыта с k возможными исходами энтропия задается формулой

$$H(p_1, p_2, \ldots, p_k) = -p_1 \log p_1 - p_2 \log p_2 - \ldots - p_k \log p_k,$$

где $p_1, p_2, \ldots, p_k$ — вероятности отдельных исходов, так что всегда $p_1 + p_2 + \ldots + p_k = 1$. В этом более общем случае (ибо при $k = 2$ функция $H(p_1, p_2, \ldots, p_k)$ обращается в $H(p_1, 1-p_1) = h(p_1)$) также можно доказать, что функция $H(p_1, p_2, \ldots, p_k)$ принимает наибольшее значение (равное $\log k$) при $p_1 = p_2 = \ldots = p_k = \frac{1}{k}$; соответствующее доказательство приведено в Приложении I (см. стр. 453). Для того чтобы представить себе характер зависимости функции $H(p_1, p_2, \ldots, p_k)$ от отдельных вероятностей $p_1, p_2, \ldots, p_k$, рассмотрим более внимательно график функции $-p \log p$, $0 < p < 1$ (см. рис. 9, где в несколько большем масштабе воспроизведена часть рис. 7) [1]). Из этого графика видно, что при $p < 0{,}1$ величина $-p \log p$ растет чрезвычайно быстро; поэтому в этой области сравнительно небольшому уменьшению

[1]) Таблица значений функции $-p \log p$ (логарифмы — двоичные!) составляет Приложение III к книге.

вероятности p_i ($i = 1, \ldots, k-1$ или k) отвечает очень значительное уменьшение соответствующего слагаемого $-p_i \log p_i$ в выражении функции $H(p_1, p_2, \ldots, p_k)$. Это приводит к тому, что обычно слагаемые $-p_i \log p_i$, отвечающие очень малым значениям вероятности p_i, вносят много меньший вклад в выражение $H(p_1, p_2, \ldots, p_k)$,

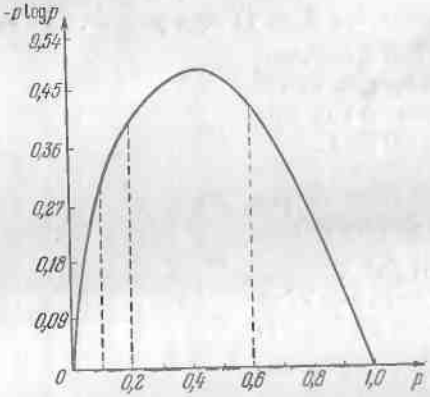


Рис. 9.

чем прочие члены, так что при вычислении энтропии сравнительно маловероятные исходы часто можно без большой ошибки просто опустить (ср. текст, напечатанный на стр. 86—87 мелким шрифтом). Наоборот, в области между $p = 0{,}2$ и $p = 0{,}6$, где функция $-p \log p$ принимает наибольшие значения, она меняется сравнительно плавно; поэтому в этой области даже довольно значительное изменение вероятностей p_i сравнительно мало отражается на величине энтропии. Отметим еще, что из непрерывности графика функции $-p \log p$ следует, что энтропия $H(\alpha)$ непрерывно зависит от вероятностей отдельных исходов опыта α, т. е. что при очень малом изменении этих вероятностей и энтропия изменится очень мало.

Задача 16. *Имеются две урны, содержащие по* 20 *шаров —* 10 *белых,* 5 *черных и* 5 *красных в первой и* 8 *белых,* 8 *черных и* 4 *красных во второй. Из каждой урны вытаскивают по одному шару. Исход какого из этих двух опытов следует считать более неопределенным?*

Таблицы вероятностей для соответствующих опытов (обозначим их через α_1 и α_2) имеют вид:

опыт α_1 (извлечение шара из 1-й урны):

цвет вынутого шара	белый	черный	красный
вероятность	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

опыт α_2 (извлечение шара из 2-й урны):

цвет вынутого шара	белый	черный	красный
вероятность	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{5}$

Энтропия первого опыта равна

$$H(\alpha_1) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} = \\ = \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 2 = 1{,}5 \text{ бита},$$

а энтропия второго несколько больше:

$$H(\alpha_2) = -\frac{2}{5}\log\frac{2}{5} - \frac{2}{5}\log\frac{2}{5} - \frac{1}{5}\log\frac{1}{5} \approx \\ \approx \frac{4}{5}\cdot 1{,}32 + \frac{1}{5}\cdot 2{,}32 \approx 1{,}52 \text{ бита}.$$

Поэтому, если оценивать (как мы это условились делать) степень неопределенности исхода опыта его энтропией, то надо считать, что исход второго опыта является более неопределенным, чем исход первого.

Задача 17. *Пусть из многолетних наблюдений за погодой известно, что для определенного пункта вероятность того, что* 15 *июня будет идти дождь, равна* 0,4, *а вероятность того, что в указанный день дождя не будет, равна* 0,6. *Пусть далее для этого же пункта вероятность того, что* 15 *ноября будет идти дождь равна* 0,65, *вероятность того, что* 15 *ноября будет идти снег, равна* 0,15 *и вероятность того, что* 15 *ноября вовсе не будет осадков, равна* 0,2. *Если из всех характеристик погоды интересоваться лишь вопросом о наличии и о характере осадков, то*

в какой из двух перечисленных дней погоду в рассматриваемом пункте следует считать более неопределенной?

Согласно тому, как понимается здесь слово «погода», опыты α_1 и α_2, состоящие в выяснении того, какая погода имела место 15 июня и 15 ноября, характеризуются следующими таблицами вероятностей:

опыт α_1:

исходы опыта	дождь	отсутствие осадков
вероятность	0,4	0,6

опыт α_2:

исходы опыта	дождь	снег	отсутствие осадков
вероятность	0,65	0,15	0,2

Поэтому энтропии наших двух опытов равны

$$H(\alpha_1) = -0{,}4 \log 0{,}4 - 0{,}6 \log 0{,}6 \approx 0{,}97 \text{ бита},$$

и

$$H(\alpha_2) = -0{,}65 \log 0{,}65 - 0{,}15 \log 0{,}15 - 0{,}2 \log 0{,}2 \approx$$
$$\approx 1{,}28 \text{ бита} > H(\alpha_1).$$

Следовательно, погоду 15 ноября в рассматриваемом пункте следует считать более неопределенной, чем 15 июня.

Полученный результат, разумеется, существенно зависит от того, как понимать слово «погода»; без точного разъяснения того, что под этим понимается, наша задача вообще не имеет смысла. В частности, если интересоваться только тем, будут ли в рассматриваемый день осадки или нет, то исходы «дождь» и «снег» опыта α_2 следует объединить. При этом вместо α_2 мы будем иметь опыт α_2', энтропия которого равна

$$H(\alpha_2') = -0{,}8 \log 0{,}8 - 0{,}2 \log 0{,}2 \approx 0{,}72 < H(\alpha_1).$$

Поэтому при таком понимании погоды надо считать, что 15 ноября погода является менее неопределенной, чем 15 июня. Если же интересоваться не только осадками, но и, например, температурой воздуха, то решение задачи становится более сложным и требует привлечения дополнительных данных о распределении значений температуры в рассматриваемом пункте 15 июня и 15 ноября.

Соображения, развитые в решении задачи 17, представляют интерес для оценки качества предсказания погоды по тому или иному методу (аналогично обстоит дело и в случае любого другого прогноза). В самом деле, при оценке качества прогноза нельзя учитывать лишь его точность (т. е. процент случаев, в которых прогноз оправдывается); иначе нам пришлось бы высоко оценивать любой прогноз, имеющий большие шансы оказаться правильным — в том числе, например, и предсказание отсутствия снега в Москве 1 июня, не представляющее, разумеется, никакой ценности. При сравнении качества различных прогнозов следует учитывать не только их точность, но и трудность удачного прогноза, которую можно характеризовать степенью неопределенности соответствующего опыта. К этому вопросу мы еще вернемся ниже (см. задачу 21 в § 3 этой главы, стр. 108).

Исторически первые шаги к введению понятия энтропии были сделаны еще в 1928 г. американским инженером-связистом Хартли[1]), предложившим характеризовать степень неопределенности опыта с k различными исходами числом $\log k$. Разумеется, Хартли хорошо понимал, что предложенная им мера степени неопределенности, очень удобная в некоторых практических задачах, во многих случаях оказывается мало показательной, поскольку она полностью игнорирует различие между характером имеющихся исходов (почти невероятному исходу здесь придается такое же значение, как и исходу весьма правдоподобному). Однако он считал, что различия между отдельными исходами определяются в первую очередь «психологическими факторами» и должны учитываться поэтому лишь психологами, но никак не инженерами или математиками.

Ошибочность точки зрения Хартли была показана Клодом Шенноном, предложившим принять в качестве меры неопределенности опыта α с возможными исходами $A_1, A_2, \ldots, A_k$ величину

$$H(\alpha) = -p(A_1)\log p(A_1) - p(A_2)\log p(A_2) - \ldots$$
$$\ldots - p(A_k)\log p(A_k),$$

[1]) Русский перевод работы Хартли напечатан в сборнике «Теория информации и ее приложения», М., Физматгиз, 1969, стр. 5—35.

где $p(A_1)$, $p(A_2)$, . . ., $p(A_n)$ — вероятности отдельных исходов; он же предложил называть эту величину «энтропией». Иначе говоря, согласно Шеннону, исходу A_i опыта α следует приписать неопределенность, равную $-\log p(A_i)$ (подобно тому, как в случае k равновероятных исходов, имеющих вероятность $p = \frac{1}{k}$, за меру неопределенности, согласно Хартли, следует принять число $\log k = -\log p$). Далее в качестве меры неопределенности всего опыта α принимается *среднее значение* неопределенности отдельных исходов (т. е. среднее значение случайной величины, принимающей значения $-\log p(A_1)$, $-\log p(A_2)$, . . ., $-\log p(A_n)$ с вероятностями $p(A_1)$, $p(A_2)$, . . ., $p(A_n)$; согласно приведенному на стр. 24 определению это среднее значение и равно $H(\alpha)$). Таким образом, загадочные «психологические факторы» Хартли здесь учитываются с помощью использования понятия вероятности, имеющего чисто математический (точнее статистический) характер.

Использование величины $H(\alpha)$ в качестве меры неопределенности опыта α оказывается очень удобным для весьма многих целей; раскрытию этого обстоятельства и посвящена, в основном, последующая часть книги. Следует, однако, иметь в виду, что мера Шеннона, как и мера Хартли, не может претендовать на полный учет всех факторов, определяющих «неопределенность опыта» в любом смысле, какой может встретиться в жизни. Так, например, мера $H(\alpha)$ зависит лишь от вероятностей $p(A_1)$, $p(A_2)$, ..., $p(A_n)$ различных исходов опыта, но вовсе не зависит от того, каковы сами эти исходы — являются ли они в некотором смысле «близкими» один к другому или очень «далекими». Поэтому наша «степень неопределенности» будет одинаковой для двух случайных величин, характеризующихся следующими таблицами вероятностей:

значения	0,9	1	1,1
вероятности	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

и

значения	−200	1	1000
вероятности	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

или для двух методов лечения больного, один из которых приводит к полному выздоровлению в 90 случаях из 100 и к заметному улучшению состояния больного — в осталь-

ных 10 случаях, а второй также вполне успешен в 90 случаях из 100, но зато в остальных 10 случаях завершается смертельным исходом. Существенное различие между двумя опытами в этих случаях должно оцениваться совсем другими характеристиками, отличными от энтропии Шеннона.

Отмеченная особенность энтропии $H(\alpha)$, как и ряд других особенностей этой величины, естественно объясняется тем, что понятие энтропии первоначально было введено специально для решения некоторых вопросов теории передачи сообщений по линиям связи и поэтому оно особенно удобно именно для такого использования. То обстоятельство, что для определения времени, требующегося для передачи некоторого сообщения, или стоимости такой передачи, конкретное содержание самого сообщения совершенно несущественно, и проявляется в независимости энтропии $H(\alpha)$ от значений $A_1, A_2, \ldots, A_k$ самих исходов опыта. С другой стороны, вероятности отдельных сообщений вовсе не безразличны для теории связи; об этом подробнее мы еще будем говорить в гл. IV. И особенно важно то обстоятельство, что при работе линии связи основную роль играют статистические закономерности, так как по такой линии всегда передается большое количество разнообразных сообщений. Поэтому мера неопределенности, используемая в решении задач, касающихся работы линий связи, должна быть приспособлена, в первую очередь, для оценки степени неопределенности сложных «составных опытов», состоящих из целой совокупности следующих друг за другом испытаний.

Любопытно, что с точки зрения исследователя, изучающего степень неопределенности таких составных опытов, различие между взглядами Хартли и Шеннона оказывается совсем не таким значительным, как это может показаться сначала. В самом деле, ведь даже с точки зрения Хартли нельзя совершенно игнорировать вероятности появления исходов — иначе можно было бы произвольно увеличить число k исходов нашего опыта, добавив к реально возможным исходам любое число фиктивных исходов, имеющих вероятность нуль. Поэтому при вычислении меры неопределенности опыта по Хартли мы непременно должны отбросить все «невозможные» исходы, имеющие

нулевую вероятность. Но при этом вряд ли стоит учитывать и «практически невозможные» исходы, осуществление которых имеет столь малую вероятность, что на практике ее можно считать нулевой. Заменим теперь опыт α, имеющий k различных исходов, другим опытом α_N, состоящим в N-кратном повторении (при одинаковых условиях) опыта α. Число различных исходов этого последнего опыта будет равно k^N; эти k^N исходов мы получим комбинируя k возможных исходов первого выполнения опыта α с k возможными исходами второго выполнения, k исходами третьего выполнения и т. д. вплоть до k исходов N-го выполнения α. Поэтому степень неопределенности опыта α_N по Хартли равна $\log k^N = N \log k$, что снова приводит к выражению $\log k$ для степени неопределенности опыта α (ибо естественно считать, что степень неопределенности опыта, состоящего в N-кратном повторении α, должна быть ровно в N раз больше степени неопределенности α; ср. аналогичное рассуждение на стр. 69).

До сих пор, однако, мы ничего не говорили о вероятностях наших k^N исходов опыта α_N. Ясно, что если k исходов α являются равновероятными, то равновероятными будут и все k^N исходов опыта α_N, так как здесь ни один из этих k^N исходов ничем не выделяется среди других. Если же k исходов опыта α имеют разные вероятности $p(A_1), p(A_2), \ldots, p(A_k)$, то разные вероятности будут иметь и $k^N = 2^{N \log k}$ исходов составного опыта α_N. Оказывается, что при больших значениях N подавляющее большинство из этих $2^{N \log k}$ исходов будет иметь настолько ничтожно малую вероятность, что даже суммарная вероятность всех таких маловероятных исходов будет очень мала. Что же касается остальных (более вероятных) исходов опыта α_N, то вероятности всех этих исходов при большом N почти не отличаются друг от друга. Точнее говоря, можно показать, *что при достаточно большом N всегда можно отбросить некоторую* (как правило, очень большую!) *часть исходов опыта α_N так, чтобы общая вероятность всех отброшенных исходов была меньше любого заранее выбранного очень малого числа* (например, меньше чем 0,01, или чем 0,001, или чем 0,000001; при этом только, чем меньшим мы выберем это число, тем большим придется взять N) *и чтобы оставшиеся исходы опыта α_N имели бы уже все практически одинаковую веро-*

ятность. Весьма важно при этом, что *число оставшихся после такого отбрасывания исходов опыта* α_N *оказывается имеющим порядок* $2^{NH(\alpha)}$, *где* $H(\alpha) = -p(A_1)\log p(A_1) - \ldots - p(A_k)\log p(A_k)$ — *энтропия опыта* α[1]). Поэтому ясно, что за меру степени неопределенности опыта α_N даже с точки зрения Хартли естественно принять число $\log 2^{N \cdot H(\alpha)} = N \cdot H(\alpha)$ (ибо исходами, суммарная вероятность которых ничтожно мала, естественно пренебречь); при этом для степени неопределенности исходного опыта α получается значение $N \cdot H(\alpha)/N = H(\alpha)$. Таким образом, мы видим, что точка зрения Шеннона отличается от точки зрения Хартли в первую очередь привлечением длинных цепочек, составленных из повторных осуществлений одного и того же опыта α; рассмотрение подобных цепочек является характерным для теоретико-вероятностного (или статистического) подхода.

Утверждение, выделенное курсивом, поясняет статистический смысл понятия энтропии; оно лежит в основе большинства технических приложений этого понятия. Однако доказательство этого утверждения не очень просто; мы отложим его (а также и несколько более аккуратную формулировку самого утверждения) до заключительной главы книги, непосредственно посвященной применениям понятия энтропии к теории передачи сообщений.

Реальная ценность понятия энтропии определяется в первую очередь тем, что выражаемая им «степень неопределенности» опытов оказывается во многих случаях именно той характеристикой, которая играет роль в разнообразных процессах, встречающихся в природе и технике и так или иначе связанных с передачей и хранением каких-либо сообщений. О некоторых технических применениях понятия энтропии мы будем сравнительно подробно говорить дальше; здесь же мы остановимся лишь на одном примере совсем другого рода.

Одной из основных задач, с которыми имеет дело экспериментальная психология, является изучение

---

[1]) Отсюда, в частности, вытекает, что если только не все исходы опыта α равновероятны и, следовательно, $H(\alpha) < \log k$, то число отброшенных исходов составляет подавляющую часть исходов опыта α_N (ибо отношение $2^{N \cdot H(\alpha)} : k^N = 2^{N \cdot H(\alpha)} : 2^{N \cdot \log k} = 2^{-N \cdot [\log k - H(\alpha)]}$ при большом N будет очень мало).

психических реакций, т. е. ответов организма на какое-либо раздражение или воздействие. При этом различаются простая реакция — какой-то определенный ответ на некоторый заданный сигнал,— и сложная реакция, важнейшей из которых является реакция выбора, состоящая в том, что на разные сигналы даются разные ответы. Известно, что время простой реакции у человека не зависит от подаваемого сигнала (для тренированных взрослых людей его минимальное значение близко к 0,1 сек). Значительно более сложным является вопрос о времени сложной реакции, существенно зависящем от условий эксперимента и, прежде всего, от «степени сложности» реакции. Еще в 80-е годы прошлого столетия психологами было выяснено, что средняя скорость, с какой человек может реагировать на последовательность беспорядочно чередующихся сигналов k различных типов (при условии, что на каждый тип сигнала он должен реагировать по-разному) монотонно уменьшается с возрастанием k. Подтверждающие этот факт опыты по определению среднего времени реакции выбора производились очень много раз и всегда приводили к примерно одинаковым результатам; наиболее обычная постановка их заключалась в том, что на стенке перед испытуемым через определенные промежутки времени вспыхивала одна из k лампочек или появлялась одна из k цифр, и в зависимости от номера сигнала он должен был нажать одну из k кнопок, на которых заранее лежали его пальцы, или же произнести одно из k заранее назначенных слов. Специальное устройство при этом отмечало время, проходящее между появлением сигнала и реакцией на него испытуемого; зависимость получаемого среднего значения T такого времени от числа k и изучалась.

Естественно, что среднее время, требующееся для реакции на сигнал, можно также рассматривать как определенную меру «степени неопределенности» ожидаемого сигнала: чем большая неопределенность в исходе имеет место, тем больше требуется времени на уяснение того, какой именно сигнал был подан. Имеющиеся опыты показывают, что *среднее время реакции растет с увеличением числа k различных типов сигналов примерно как* $\log k$, *т. е. как шенноновская энтропия* $H(\alpha)$ *опыта* α, *состоявшего в подаче сигнала* (во всех экспериментах, о которых

здесь идет речь, вероятности сигналов различных типов всегда были одинаковыми). Для примера на рис. 10 (заимствованном из работы американского психолога Р. Хаймана [38]) кружками отмечены данные восьми опытов, состоящих в определении среднего времени, требующегося испытуемому, чтобы указать, какая из k лампочек (где k менялось от 1 до 8) зажглась. Это среднее

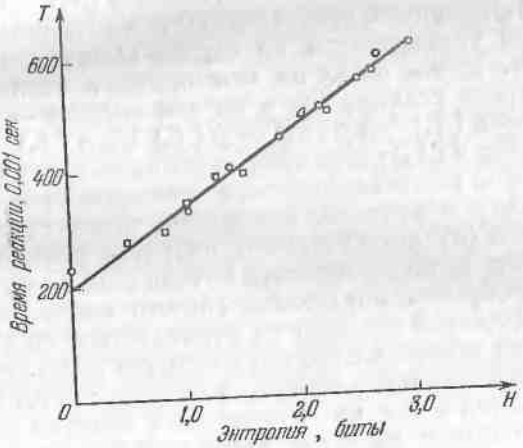


Рис. 10.

время определялось из большого числа серий зажиганий, в каждой из которых частоты зажиганий всех лампочек были одинаковыми, причем предварительно испытуемый специально тренировался в подобных опытах. По оси ординат на рис. 10 отложено среднее время реакции, а по оси абсцисс — величина $\log k$; при этом, как мы видим, все 8 кружков довольно точно укладываются на одну прямую.

Исходя из этих данных, можно было бы предположить, что *среднее время реакции во всех случаях определяется энтропией опыта* α, *состоящего в подаче сигнала*. Из этого предположения следует, что уменьшение степени неопределенности опыта путем замены равновероятных сигналов

неравновероятными должно на столько же уменьшить среднее время реакции, на сколько оно уменьшается при уменьшении числа используемых типов сигналов, приводящем к такому же изменению энтропии $H(\alpha)$. Это утверждение допускает прямую экспериментальную проверку, полностью его подтверждающую. Так, на том же рис. 10 квадратиками отмечены результаты восьми опытов (проведенных с тем же испытуемым, что и раньше), в которых k лампочек (где k равнялось 2, 4, 6 или 8) зажигались с разными относительными частотами $p(A_1)$, $p(A_2)$, ..., $p(A_k)$, причем предварительно испытуемый некоторое время тренировался на сериях зажиганий с такими частотами. Здесь снова по оси ординат откладывалось среднее время реакции T, а по оси абсцисс — энтропия $H(\alpha) = -p(A_1)\log p(A_1) - p(A_2)\log p(A_2) - \ldots - p(A_k)\log p(A_k)$; при этом оказывается, что квадратики с большой степенью точности укладываются на ту же прямую, что и кружки. Мы видим, таким образом, что энтропия $H(\alpha)$ действительно является именно той мерой степени неопределенности исхода опыта, которая решающим образом определяет среднее время, требуемое для определенной реакции на появившийся сигнал.

Причина изменения среднего времени реакции при изменении относительной частоты различных сигналов, очевидно, кроется в том, что испытуемый быстрее реагирует должным образом на более часто повторяющийся (т. е. более привычный для него) сигнал, но зато медленнее реагирует на редкий сигнал, являющийся для него неожиданным. Разумеется, эти факторы носят психологический характер. Тем не менее мы видим, что и они могут быть количественно охарактеризованы величиной энтропии $H(\alpha)$ опыта α, вопреки опасениям Хартли, предполагавшего, что никакие «психологические факторы» (которые, впрочем, в его понимании имели гораздо более косвенное отношение к психологии, чем в настоящем примере) не могут быть количественно учтены.

В заключение этого параграфа приведем некоторые данные, характеризующие незначительность роли, которую играют в определении энтропии опыта со многими исходами многочисленные маловероятные исходы.

Рассмотрим опыт, состоящий в выборе из печатного текста наудачу одного слова из четырех букв; при этом наш текст мы будем считать написанным по-английски, что позволяет исполь-

зовать данные, содержащиеся в известном «Словаре Торндайка» (E. L. Thorndike «A Teacher's Word Book», New York, 1932), в котором указаны частоты 20 000 наиболее распространенных английских слов, полученные путем статистической обработки очень большого и разнообразного английского текста. Всего в этом словаре содержится 1550 четырехбуквенных слов; в соответствии с этим мы можем считать, что наш опыт α имеет 1550 различных исходов. Вычислим теперь энтропию

$$H(\alpha) = -p(A_1)\log p(A_1) - p(A_2)\log p(A_2) - \ldots \\ \ldots - p(A_{1550})\log p(A_{1550})$$

этого опыта, приняв вероятность $p(A_i)$ каждого исхода равной частоте n_i/N соответствующего слова; здесь n_i есть число повторений этого слова, указанное в словаре Торндайка, а $N = n_1 + n_2 + \ldots + n_{1550}$. Оказывается, что эта энтропия близка к 8,14 бит [1]). Отбросим теперь все слова, для которых $n_i < 150$; при этом остается лишь 865 четырехбуквенных слов, т. е. немного больше 50% от их исходного числа (точнее говоря — 55,8%). В то же время отвечающая этим 865 исходам опыта α часть суммы $H(\alpha)$ равна примерно 8 бит, т. е. составляет более 98% от всей величины $H(\alpha)$. Отбросим теперь все слова, для которых $n_i < 750$; при этом у нас останется 395 слов, т. е. всего около четверти (25,5%) первоначального количества; однако этим 395 исходам будет отвечать часть суммы $H(\alpha)$, большая чем 7,47 бит, т. е. составляющая свыше 92% всей величины $H(\alpha)$. Если мы отбросим затем все слова с $n_i < 1550$, то у нас останется только 214 слов (13,8% от исходного количества); однако этим 214 исходам опыта будет отвечать часть суммы $H(\alpha)$, близкая к 6,88 бит, т. е. составляющая около 85% ее первоначального значения. Наконец, если отбросить все слова с $n_i < 3150$, то останется всего 119 четырехбуквенных слов (7,7% от первоначального количества); однако этим 7,7% исходов будет отвечать около 78% суммы $H(\alpha)$ (эта часть суммы $H(\alpha)$ превышает 6,44 бит).

§ 2. Энтропия сложных событий. Условная энтропия

Пусть α и β — два независимых опыта с таблицами вероятностей:

опыт α

исходы опыта	A_1	A_2	...	A_k
вероятности	$p(A_1)$	$p(A_2)$	...	$p(A_k)$

опыт β

исходы опыта	B_1	B_2	...	B_l
вероятности	$p(B_1)$	$p(B_2)$	...	$p(B_l)$

[1]) Это значение, как и все последующие численные данные, заимствовано из сборника [46].

опыт β

исходы опыта	B_1	B_2	...	B_l
вероятности	$p(B_1)$	$p(B_2)$	...	$p(B_l)$

Рассмотрим сложный опыт $\alpha\beta$, состоящий в том, что одновременно осуществляются опыты α и β. Этот опыт может иметь kl исходов:

$$A_1B_1,\ A_1B_2,\ \ldots,\ A_1B_l;\ A_2B_1,\ A_2B_2,\ \ldots,\ A_2B_l;\ \ldots$$
$$\ldots;\ A_kB_1,\ A_kB_2,\ \ldots,\ A_kB_l,$$

где, например, A_1B_1 означает, что опыт α имел исход A_1, а опыт β — исход B_1. Очевидно, что неопределенность опыта $\alpha\beta$ больше неопределенности каждого из опытов α и β, так как здесь осуществляются сразу оба эти опыта, каждый из которых может иметь разные исходы в зависимости от случая.

Докажем равенство

$$H(\alpha\beta) = H(\alpha) + H(\beta)$$

(правило сложения энтропий), которое хорошо согласуется со смыслом энтропии как меры степени неопределенности. Согласно определению $H(\alpha\beta)$ имеем:

$$\begin{aligned} H(\alpha\beta) = &- p(A_1B_1)\log p(A_1B_1) - p(A_1B_2)\log p(A_1B_2) - \ldots \\ &\ldots - p(A_1B_l)\log p(A_1B_l) - \\ &- p(A_2B_1)\log p(A_2B_1) - p(A_2B_2)\log p(A_2B_2) - \ldots \\ &\ldots - p(A_2B_l)\log p(A_2B_l) - \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &- p(A_kB_1)\log p(A_kB_1) - p(A_kB_2)\log p(A_kB_2) - \ldots \\ &\ldots - p(A_kB_l)\log p(A_kB_l). \end{aligned}$$

Но так как опыты α и β независимы, то $p(A_1B_1) = p(A_1)\,p(B_1)$, $p(A_1B_2) = p(A_1)\,p(B_2)$ и т. д. (см. § 2 гл. I. Поэтому первая строка стоящего справа

выражения может быть записана так:

$$
\begin{aligned}
&- p(A_1)\, p(B_1)\,(\log p(A_1) + \log p(B_1)) - \\
&\quad - p(A_1)\, p(B_2)\,(\log p(A_1) + \log p(B_2)) - \ldots \\
&\quad \ldots - p(A_1)\, p(B_l)\,(\log p(A_1) + \log p(B_l)) = \\
&= - p(A_1)\,(p(B_1) + p(B_2) + \ldots + p(B_l)) \log p(A_1) + \\
&+ p(A_1)\,(- p(B_1) \log p(B_1) - p(B_2) \log p(B_2) - \ldots \\
&\quad \ldots - p(B_l) \log p(B_l)) = \\
&\qquad = - p(A_1) \log p(A_1) + p(A_1)\, H(\beta)
\end{aligned}
$$

(так как $p(B_1) + p(B_2) + \ldots + p(B_l) = 1$). Совершенно аналогично 2-я, ..., k-я строки в выражении для $H(\alpha\beta)$ равны

$$
\begin{aligned}
&- p(A_2) \log p(A_2) + p(A_2)\, H(\beta), \\
&\qquad \ldots\ldots\ldots\ldots\ldots \\
&- p(A_k) \log p(A_k) + p(A_k)\, H(\beta)
\end{aligned}
$$

и, значит,

$$
\begin{aligned}
H(\alpha\beta) &= - p(A_1) \log p(A_1) - p(A_2) \log p(A_2) - \ldots \\
&\ldots - p(A_k) \log p(A_k) + \\
&+ (p(A_1) + p(A_2) + \ldots p(A_k))\, H(\beta) = H(\alpha) + H(\beta)
\end{aligned}
$$

(так как и $p(A_1) + p(A_2) + \ldots + p(A_k) = 1$).

Предположим теперь, что опыты α и β *не независимы* (например, что α и β — последовательные извлечения двух шаров из одной урны; ср. выше, стр. 40). В этом более общем случае мы не можем ожидать, что энтропия сложного опыта $\alpha\beta$ будет равна сумме энтропий α и β. В самом деле, здесь может представиться такой случай, когда результат второго опыта полностью определяется результатом первого (например, если опыты α и β состоят в последовательном извлечении шаров из урны, содержащей всего два разноцветных шара). В этом случае после осуществления α опыт β уже не будет содержать *никакой* неопределенности; поэтому здесь естественно предполагать, что энтропия (мера степени неопределенности) сложного опыта $\alpha\beta$ будет равна энтропии одного опыта α, а не сумме энтропий опытов α и β (в дальнейшем мы убедимся, что это на самом деле так). Постараемся выяснить, чему равна энтропия сложного опыта $\alpha\beta$ в общем случае.

Повторим вывод формулы для энтропии $H(\alpha\beta)$ сложного опыта $\alpha\beta$, отказавшись от предположения о независимости опытов α и β. Очевидно, мы, как прежде, будем

иметь

$$H(\alpha\beta) = -p(A_1B_1)\log p(A_1B_1) - p(A_1B_2)\log p(A_1B_2) - \ldots$$
$$\ldots - p(A_1B_l)\log p(A_1B_l) -$$
$$- p(A_2B_1)\log p(A_2B_1) - p(A_2B_2)\log p(A_2B_2) - \ldots$$
$$\ldots - p(A_2B_l)\log p(A_2B_l) -$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$- p(A_kB_1)\log p(A_kB_1) - p(A_kB_2)\log p(A_kB_2) - \ldots$$
$$\ldots - p(A_kB_l)\log p(A_kB_l),$$

где снова через $A_1, A_2, \ldots, A_k$ и $B_1, B_2, \ldots, B_l$ обозначены соответственно исходы опытов α и β. Однако здесь уже нельзя заменить вероятности $p(A_1B_1)$, $p(A_1B_2)$ и т. д. просто произведениями соответствующих вероятностей: теперь $p(A_1B_1)$ равно не $p(A_1)\,p(B_1)$, а $p(A_1)\,p_{A_1}(B_1)$, где $p_{A_1}(B_1)$ — условная вероятность события B_1 при условии A_1 (см. § 3 гл. I). Это обстоятельство вносит существенные изменения в дальнейшие рассуждения.

Как прежде, рассмотрим сначала лишь члены, стоящие в первой строке выписанного выше выражения для $H(\alpha\beta)$. Очевидно, что их можно переписать в виде

$$-p(A_1)\,p_{A_1}(B_1)(\log p(A_1) + \log p_{A_1}(B_1)) -$$
$$- p(A_1)\,p_{A_1}(B_2)(\log p(A_1) + \log p_{A_1}(B_2)) - \ldots$$
$$\ldots - p(A_1)\,p_{A_1}(B_l)(\log p(A_1) + \log p_{A_1}(B_l)) =$$
$$= -p(A_1)(p_{A_1}(B_1) + p_{A_1}(B_2) + \ldots + p_{A_1}(B_l))\log p(A_1) +$$
$$+ p(A_1)(-p_{A_1}(B_1)\log p_{A_1}(B_1) - p_{A_1}(B_2)\log p_{A_1}(B_2) - \ldots$$
$$\ldots - p_{A_1}(B_l)\log p_{A_1}(B_l)).$$

Но

$$p_{A_1}(B_1) + p_{A_1}(B_2) + \ldots + p_{A_1}(B_l) =$$
$$= p_{A_1}(B_1 + B_2 + \ldots + B_l) = 1,$$

ибо событие $B_1 + B_2 + \ldots + B_l$ — достоверное (какой-либо из исходов $B_1, B_2, \ldots, B_l$ опыта β наверное имеет место). С другой стороны, сумма

$$-p_{A_1}(B_1)\log p_{A_1}(B_1) - p_{A_1}(B_2)\log p_{A_1}(B_2) - \ldots$$
$$\ldots - p_{A_1}(B_l)\log p_{A_1}(B_l)$$

представляет собой энтропию опыта β при условии, что имело место событие A_1 (энтропия опыта β зависит от ис-

хода опыта α, так как от исхода α зависят вероятности отдельных исходов β). Это выражение естественно назвать условной энтропией опыта β при условии A_1 и обозначить через $H_{A_1}(\beta)$.

Таким образом, первая строка выражения для $H(\alpha\beta)$ может быть переписанав следующем виде:

$$-p(A_1)\log p(A_1) + p(A_1)H_{A_1}(\beta).$$

Точно так же 2-я, . . ., k-я строки этого выражения соответственно равны

$$-p(A_2)\log p(A_2) + p(A_2)H_{A_2}(\beta),$$

. . . . . .

$$p-(A_k)\log p(A_k) + p(A_k)H_{A_k}(\beta),$$

где $H_{A_2}(\beta), \ldots, H_{A_k}(\beta)$ — условные энтропии опыта β при условиях $A_2, \ldots, A_k$. Отсюда вытекает следующая формула:

$$\begin{aligned} H(\alpha\beta) = &-p(A_1)\log p(A_1) - \\ &- p(A_2)\log p(A_2) - \ldots - p(A_k)\log p(A_k) + \\ &+ p(A_1)H_{A_1}(\beta) + p(A_2)H_{A_2}(\beta) + \ldots + p(A_k)H_{A_k}(\beta) = \\ &= H(\alpha) + \{p(A_1)H_{A_1}(\beta) + p(A_2)H_{A_2}(\beta) + \ldots \\ &\qquad \ldots + p(A_k)H_{A_k}(\beta)\}. \end{aligned}$$

Первый член последнего выражения представляет собой энтропию опыта α. Что же касается второго, то это есть среднее значение случайной величины, принимающей с вероятностями $p(A_1)$, $p(A_2)$, . . ., $p(A_k)$ значения $H_{A_1}(\beta)$, $H_{A_2}(\beta)$, . . ., $H_{A_k}(\beta)$, т. е. значения, равные условной энтропии опыта β при условии, что опыт α имеет исходы $A_1, A_2, \ldots, A_k$. Это среднее значение естественно назвать средней условной энтропией опыта β при условии выполнения опыта α, или, короче, условной энтропией β при условии выполнения α; мы будем обозначать его через $H_\alpha(\beta)$:

$$H_\alpha(\beta) = p(A_1)H_{A_1}(\beta) + p(A_2)H_{A_2}(\beta) + \ldots + p(A_k)H_{A_k}(\beta).$$

Таким образом, окончательно имеем

$$H(\alpha\beta) = H(\alpha) + H_\alpha(\beta).$$

Это и есть общее правило для определения энтропии сложного опыта $\alpha\beta$. Его тоже можно назвать п р а в и л о м с л о ж е н и я э н т р о п и й, аналогично выведенному выше правилу, относящемуся к тому частному случаю, когда опыты α и β независимы.

Следует отметить, что именно средняя условная энтропия $H_\alpha(\beta)$ играет существенную роль в рассматриваемых в этой книге вопросах. Дело в том, что коль скоро мы знаем заранее, какой именно исход A_i опыта α имел место, то при последующем определении условной энтропии $H_{A_i}(\beta)$ опыта β мы можем полностью игнорировать все строки таблицы условных вероятностей

$$\begin{aligned}
&p_{A_1}(B_1),\ p_{A_1}(B_2),\ \ldots,\ p_{A_1}(B_l),\\
&p_{A_2}(B_1),\ p_{A_2}(B_2),\ \ldots,\ p_{A_2}(B_l),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\\
&p_{A_k}(B_1),\ p_{A_k}(B_2),\ \ldots,\ p_{A_k}(B_l),
\end{aligned}$$

кроме единственной строки, отвечающей исходу A_i. Поэтому условная энтропия $H_{A_i}(\beta)$ совсем не зависит от того, как изменяются вероятности отдельных исходов β при $k-1$ исходах опыта α (из общего числа k исходов) и, следовательно, она лишь в весьма малой степени характеризует связь между опытами α и β, полное выражение которой дается всей таблицей условных вероятностей [1]). Напротив того, средняя условная энтропия $H_\alpha(\beta)$, вычисление которой не предполагает известным исход α, глубоко отражает взаимную зависимость опытов α и β. Подробнее об этом мы будем говорить в § 3 настоящей главы.

Укажем некоторые важнейшие свойства величины $H_\alpha(\beta)$. Очевидно, что это есть неотрицательное число. Ясно, что если все вероятности $p(A_1),\ p(A_2),\ \ldots,\ p(A_k)$ отличны от нуля, т. е. если опыт α имеет действительно k исходов, то $H_\alpha(\beta)=0$ *в том и только в том случае*, если

---

[1]) Заметим, что знание этой таблицы (и таблиц вероятностей опытов α и β) позволяет вычислить также и условные вероятности исходов $A_1, A_2, \ldots, A_k$ опыта α при условии, что опыт β имел какой-либо определенный исход B_1, или B_2, ..., или B_l; об этом см. выше, стр. 43.

$H_{A_1}(\beta) = H_{A_2}(\beta) = \ldots = H_{A_k}(\beta) = 0$, т. е. *если при любом исходе опыта α результат опыта β становится полностью определенным* (тривиальным образом это условие выполняется в том случае, если опыт β с самого начала не является неопределенным). При этом мы имеем

$$H(\alpha\beta) = H(\alpha)$$

(см. выше, стр. 89). Если же опыты α и β являются независимыми, то $H_{A_1}(\beta) = H_{A_2}(\beta) = \ldots = H_{A_k}(\beta) = H(\beta)$ и

$$H_\alpha(\beta) = H(\beta).$$

В этом случае формула $H(\alpha\beta) = H(\alpha) + H_\alpha(\beta)$ переходит в более простую: $H(\alpha\beta) = H(\alpha) + H(\beta)$ (см. выше, стр. 88).

Очень существенно, что во всех случаях *условная энтропия $H_\alpha(\beta)$ заключается между нулем и энтропией $H(\beta)$ опыта* β (безусловной):

$$0 \leqslant H_\alpha(\beta) \leqslant H(\beta).$$

Таким образом, случаи, когда исход опыта β полностью определяется исходом α и когда опыты α и β независимы, являются в определенном смысле крайними.

Это утверждение тоже хорошо согласуется со смыслом энтропии как меры неопределенности: совершенно ясно, что предварительное выполнение опыта α может лишь уменьшить степень неопределенности β или, в крайнем случае (например, в случае независимости опытов α и β), не изменить эту степень неопределенности, но никак не может ее увеличить [1]). Полное доказательство сделанного утверждения (включающее также доказательство того, что $H_\alpha(\beta) = H(\beta)$ т о л ь к о тогда, когда опыты α и β независимы) мы отложим до Приложения I в конце книги; здесь же мы лишь проиллюстрируем его на примере случая когда опыт α имеет два р а в н о в е р о я т н ы х исхода

---

[1]) Во избежание возможных заблуждений отметим, что условная энтропия $H_{A_1}(\beta)$ может быть и меньше и больше величины $H(\beta)$ (см., например, ниже задачи 18 и 19). Это связано с тем, что изменение таблицы вероятностей опыта β, обусловленное тем обстоятельством, что другой опыт α имел определенный исход A_1, может быть довольно произвольным (ср. выше, стр. 41).

A_1 и A_2. В этом случае

$$H_\alpha(\beta) = p(A_1)\,H_{A_1}(\beta) + p(A_2)\,H_{A_2}(\beta) = \frac{1}{2}H_{A_1}(\beta) + \frac{1}{2}H_{A_2}(\beta).$$

Итак, наша задача сводится к доказательству неравенства

$$\frac{1}{2}H_{A_1}(\beta) + \frac{1}{2}H_{A_2}(\beta) \leqslant H(\beta),$$

т. е. неравенства

$$\frac{1}{2}[-p_{A_1}(B_1)\log p_{A_1}(B_1) - p_{A_1}(B_2)\log p_{A_1}(B_2) - \ldots - p_{A_1}(B_l)\log p_{A_1}(B_l)] + \frac{1}{2}[-p_{A_2}(B_1)\log p_{A_2}(B_1) - p_{A_2}(B_2)\log p_{A_2}(B_2) - \ldots - p_{A_2}(B_l)\log p_{A_2}(B_l)] \leqslant -p(B_1)\log p(B_1) - p(B_2)\log p(B_2) - \ldots - p(B_l)\log p(B_l),$$

где, как всегда $B_1, B_2, \ldots, B_l$ означают исходы опыта β. Рассмотрим снова график функции $F(x) = -x\log x$, и

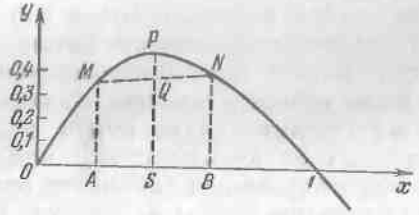


Рис. 11.

пусть на рис. 11 $OA = p_{A_1}(B_1)$, $OB = p_{A_2}(B_1)$; тогда отрезки AM и BN имеют длины $-p_{A_1}(B_1)\log p_{A_1}(B_1)$ и $-p_{A_2}(B_1)\log p_{A_2}(B_1)$. Сумма $-\frac{1}{2}p_{A_1}(B_1)\log p_{A_1}(B_1) - \frac{1}{2}p_{A_2}(B_1)\log - p_{A_2}(B_1)$ равна средней линии SQ трапеции $ABNM$. С другой стороны, отрезок SP, превосходящий SQ, равен $-p(B_1)\log p(B_1)$, так как

$$OS = \frac{1}{2}OA + \frac{1}{2}OB = p(A_1)\,p_{A_1}(B_1) + p(A_2)\,p_{A_2}(B_1) = p(B_1)$$

(см. формулу полной вероятности на стр. 44). Следовательно,

$$-\frac{1}{2}p_{A_1}(B_1)\log p_{A_1}(B_1)-\frac{1}{2}p_{A_2}(B_1)\log p_{A_2}(B_1)\leqslant$$
$$\leqslant -p(B_1)\log p(B_1).$$

Аналогично этому доказываются неравенства

$$-\frac{1}{2}p_{A_1}(B_2)\log p_{A_1}(B_2)-\frac{1}{2}p_{A_2}(B_2)\log p_{A_2}(B_2)\leqslant$$
$$\leqslant -p(B_2)\log p(B_2),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$-\frac{1}{2}p_{A_1}(B_l)\log p_{A_1}(B_l)-\frac{1}{2}p_{A_2}(B_l)\log p_{A_2}(B_l)\leqslant$$
$$\leqslant -p(B_l)\log p(B_l).$$

Сложив все эти неравенства, мы придем к требуемому результату.

Заметим еще, что так как сложные события $\alpha\beta$ и $\beta\alpha$ не отличаются одно от другого, то $H(\alpha\beta)=H(\beta\alpha)$, т. е.

$$H(\alpha)+H_\alpha(\beta)=H(\beta)+H_\beta(\alpha).$$

Отсюда следует, в частности, что, зная энтропии $H(\alpha)$ и $H(\beta)$ опытов α и β и условную энтропию $H_\alpha(\beta)$ опыта β при условии выполнения α, мы можем определить также и условную энтропию $H_\beta(\alpha)$ опыта α при условии выполнения β:

$$H_\beta(\alpha)=H_\alpha(\beta)+\{H(\alpha)-H(\beta)\}.$$

Поскольку $0\leqslant H_\beta(\alpha)\leqslant H(\alpha)$, то из формулы $H_\alpha(\beta)=H_\beta(\alpha)+H(\beta)-H(\alpha)$ следует, что

$$H(\beta)-H(\alpha)\leqslant H_\alpha(\beta)\leqslant H(\beta);$$

при $H(\beta)>H(\alpha)$ эта оценка величины условной энтропии $H_\alpha(\beta)$ оказывается более точной, чем приведенная на стр. 93. Равенство

$$H_\alpha(\beta)=H(\beta)-H(\alpha)$$

имеет место при $H_\beta(\alpha)=0$, т. е. если исход опыта β полностью определяет исход опыта α; при этом всегда будет $H(\beta)\geqslant H(\alpha)$ (что, разумеется, также хорошо согласуется со смыслом слова «неопределенность опыта»).

Задача 18. *Известно, что некоторой болезнью в среднем болеют 2 человека из 100. Для выявления больных используется определенная реакция, которая всегда оказывается положительной в том случае, когда человек болен; если же человек здоров, то она столь же часто бывает положительной, как и отрицательной. Пусть опыт* β *состоит в определении того, болен или здоров человек, а опыт* α *— в определении результата указанной реакции. Спрашивается, какова будет энтропия* $H(\beta)$ *опыта* β *и условная энтропия* $H_\alpha(\beta)$ *опыта* β *при условии осуществления* α?

Очевидно, здесь два исхода опыта β — исход B_1 (человек здоров) и исход B_2 (человек болен) — имеют вероятности: $p(B_1) = 0{,}98$ и $p(B_2) = 0{,}02$. Поэтому

$$H(\beta) = -0{,}98 \cdot \log 0{,}98 - 0{,}02 \cdot \log 0{,}02 \approx 0{,}14 \text{ бита}.$$

Опыт α также имеет два исхода: A_1 (положительная реакция) и A_2 (отрицательная реакция). Вероятности этих исходов равны

$$p(A_1) = 0{,}51 \text{ и } p(A_2) = 0{,}49$$

(ибо исход A_1 имеет место в половине тех случаев, когда опыт β имеет исход B_1, и во всех случаях, когда β имеет исход B_2, а исход A_2 — лишь в половине случаев, когда β имеет исход B_1). При этом, если опыт α имел исход A_1 (а таких случаев большинство!), то условные вероятности исходов β будут равны

$$p_{A_1}(B_1) = \frac{49}{51} \text{ и } p_{A_1}(B_2) = \frac{2}{51}$$

(ибо из 51 случая, когда реакция оказывалась положительной, в 49 случаях человек оказывался здоровым и в двух случаях — больным); поэтому условная энтропия $H_{A_1}(\beta)$ будет заметно больше безусловной энтропии $H(\beta)$:

$$H_{A_1}(\beta) = -\frac{49}{51}\log\frac{49}{51} - \frac{2}{51}\log\frac{2}{51} \approx 0{,}24 \text{ бита}.$$

Зато если опыт α имеет исход A_2, то мы с уверенностью можем утверждать, что опыт β имел исход B_1 (человек здоров); следовательно,

$$H_{A_2}(\beta) = 0.$$

Таким образом, средняя условная энтропия опыта β при

условии осуществления α будет все же меньше безусловной энтропии $H(\beta)$:

$$H_\alpha(\beta) = 0{,}51 \cdot H_{A_1}(\beta) + 0{,}49 \cdot H_{A_2}(\beta) \approx 0{,}51 \cdot 0{,}24 \approx 0{,}12 \text{ бита}.$$

Иначе говоря, выполнение опыта α уменьшает степень неопределенности опыта β примерно на 0,02 бита.

Задача 19. *Пусть опыты α и β состоят в последовательном извлечении двух шаров из урны, содержащей m черных и $n - m$ белых шаров (α — извлечение первого шара, β — извлечение второго шара). Чему равны энтропии $H(\alpha)$ и $H(\beta)$ опытов α и β и условные энтропии $H_\alpha(\beta)$ и $H_\beta(\alpha)$ тех же опытов? Решите ту же задачу при условии, что опыт α состоит в извлечении k шаров из урны, а опыт β — в последующем извлечении еще одного шара.*

Начнем со случая, когда опыт α состоит в извлечении одного шара. Пусть события A_1 и A_2 состоят в появлении черного и белого шара при первом извлечении, а события B_1 и B_2 — в появлении черного и белого шара при втором извлечении. Пока нам ничего не известно ни о первом, ни о втором опыте, мы можем ожидать осуществления этих событий со следующими вероятностями:

опыт α	исходы опыта	A_1	A_2
	вероятности	$\frac{m}{n}$	$\frac{n-m}{n}$
опыт β	исходы опыта	B_1	B_2
	вероятности	$\frac{m}{n}$	$\frac{n-m}{n}$

Таким образом, оба эти опыта имеют одинаковую энтропию:

$$H(\alpha) = H(\beta) = -\frac{m}{n}\log\frac{m}{n} - \frac{n-m}{n}\log\frac{n-m}{n}.$$

Если нам известен исход опыта α, то вероятности отдельных исходов опыта β будут иметь другие значения. А именно (ср. выше, стр. 40 и след.):

$$p_{A_1}(B_1) = \frac{m-1}{n-1}, \qquad p_{A_1}(B_2) = \frac{n-m}{n-1};$$

$$p_{A_2}(B_1) = \frac{m}{n-1}, \qquad p_{A_2}(B_2) = \frac{n-m-1}{n-1}.$$

Отсюда следует, что

$$H_{A_1}(\beta) = -\frac{m-1}{n-1}\log\frac{m-1}{n-1} - \frac{n-m}{n-1}\log\frac{n-m}{n-1},$$

$$H_{A_2}(\beta) = -\frac{m}{n-1}\log\frac{m}{n-1} - \frac{n-m-1}{n-1}\log\frac{n-m-1}{n-1}.$$

При этом, если $m < n - m$, то

$$H_{A_1}(\beta) < H(\beta), \quad H_{A_2}(\beta) > H(\beta)$$

(ибо неопределенность опыта, состоящего в извлечении одного шара из урны с m черными и $m_1 = n - m$ белыми шарами тем больше, чем ближе к единице отношение $\frac{m}{m_1}$). Наконец, имеем

$$H_\alpha(\beta) = p(A_1)H_{A_1}(\beta) + p(A_2)H_{A_2}(\beta) =$$
$$= \frac{m}{n}\left[-\frac{m-1}{n-1}\log\frac{m-1}{n-1} - \frac{n-m}{n-1}\log\frac{n-m}{n-1}\right] +$$
$$+ \frac{n-m}{n}\left[-\frac{m}{n-1}\log\frac{m}{n-1} - \frac{n-m-1}{n-1}\log\frac{n-m-1}{n-1}\right]$$

(во всех случаях $H_\alpha(\beta) < H(\beta)$) и

$$H_\beta(\alpha) = H_\alpha(\beta) + \{H(\alpha) - H(\beta)\} = H_\alpha(\beta).$$

Перейдем теперь к поставленной во второй части условия более общей задаче. Опыт α, состоящий в извлечении из урны k шаров, мы теперь будем обозначать через α_k. Пусть число k не превосходит чисел m и $n - m$. В таком случае опыт α_k может иметь $k + 1$ различных исходов, соответствующих тому, что среди извлеченных шаров окажется $0, 1, 2, \ldots, k$ черных; обозначим эти исходы через $A_0, A_1, A_2, \ldots, A_k$. Вероятность $p(A_i)$ исхода A_i $(i = 0, 1, \ldots, k)$ будет равна отношению $\frac{C_m^i C_{n-m}^{k-i}}{C_n^k}$: общее число равновероятных исходов опыта α_k равно C_n^k (числу всевозможных групп из k шаров, которые можно составить из имеющихся n шаров), а благоприятствовать исходу A_i из них будут $C_m^i \cdot C_{n-m}^{k-i}$ исходов (так как i черных шаров из имеющихся m можно выбрать C_m^i способами, а остальные $k - i$ белых шаров — C_{n-m}^{k-i} способами).

Отсюда следует, что энтропия опыта α_k равна

$$H(\alpha_k) = -\frac{C_{n-m}^k}{C_n^k}\log\frac{C_{n-m}^k}{C_n^k} - \frac{C_m^1 C_{n-m}^{k-1}}{C_n^k}\log\frac{C_m^1 C_{n-m}^{k-1}}{C_n^k} -$$
$$- \frac{C_m^2 C_{n-m}^{k-2}}{C_n^k}\log\frac{C_m^2 C_{n-m}^{k-2}}{C_n^k} - \dots$$
$$\dots - \frac{C_m^{k-1} C_{n-m}^1}{C_n^k}\log\frac{C_m^{k-1} C_{n-m}^1}{C_n^k} - \frac{C_m^k}{C_n^k}\log\frac{C_m^k}{C_n^k}.$$

Опыт β имеет два исхода B_1 (извлечение черного шара) и B_2 (извлечение белого шара). Вероятности этих исходов соответственно равны $\frac{m}{n}$ и $\frac{n-m}{n}$. Энтропия опыта β по-прежнему равна

$$H(\beta) = -\frac{m}{n}\log\frac{m}{n} - \frac{n-m}{n}\log\frac{n-m}{n}.$$

Пусть теперь мы знаем, что имел место исход A_i опыта α_k. Это значит, что в урне после осуществления этого опыта осталось $m - i$ черных и $n - m - k + i$ белых шаров. Соответственно этому

$$p_{A_i}(B_1) = \frac{m-i}{n-k}, \quad p_{A_i}(B_2) = \frac{n-m-k+i}{n-k}$$

и

$$H_{A_i}(\beta) =$$
$$= -\frac{m-i}{n-k}\log\frac{m-i}{n-k} - \frac{n-m-k+i}{n-k}\log\frac{n-m-k+i}{n-k}.$$

Для вычисления $H_{\alpha_k}(\beta)$ остается только воспользоваться формулой

$$H_{\alpha_k}(\beta) = \frac{C_{n-m}^k}{C_n^k} H_{A_0}(\beta) + \frac{C_m^1 C_{n-m}^{k-1}}{C_n^k} H_{A_1}(\beta) + \dots$$
$$\dots + \frac{C_m^k}{C_n^k} H_{A_k}(\beta).$$

Наконец, условная энтропия $H_\beta(\alpha_k)$ определяется по формуле

$$H_\beta(\alpha_k) = H_{\alpha_k}(\beta) + H(\alpha_k) - H(\beta).$$

Аналогично рассматриваются случаи, когда k больше одного из чисел m и $n-m$, или даже больше обоих этих чисел. Мы здесь не будем разбирать все представляющиеся возможности, а ограничимся лишь несколькими замечаниями.

а) Пусть $k=n-1$. Опыт α_{n-1} имеет всего два исхода A_1 и A_2, отвечающих тем случаям, когда последний оставшийся в урне шар является черным и когда он является белым. Вероятности этих двух исходов равны $\frac{m}{n}$ и $\frac{n-m}{n}$, ибо выбор $n-1$ извлеченных шаров равносилен выбору одного остающегося шара и, следовательно, наш опыт α_{n-1} по существу не отличается от опыта α_1, состоящего в извлечении из урны с n шарами одного единственного шара. Таким образом, энтропия опыта α_{n-1} равна

$$H(\alpha_{n-1}) = -\frac{m}{n}\log\frac{m}{n} - \frac{n-m}{n}\log\frac{n-m}{n},$$

т. е. совпадает с энтропией опыта β. Что же касается условной энтропии $H_{\alpha_{n-1}}(\beta)$, то она, разумеется, равна нулю, поскольку исход опыта α_{n-1} полностью предрешает исход опыта β. По аналогичной причине будет равна нулю и условная энтропия $H_\beta(\alpha_{n-1})$.

б) Пусть $k=n-2$. Опыт α_{n-2} имеет три исхода A_0, A_1 и A_2, отвечающих тем случаям, когда в урне остается два черных шара, или черный и белый шар, или два белых шара (мы считаем здесь, что ни одно из чисел m и $n-m$ не меньше 2). Вероятности этих исходов равны

$$p(A_0) = \frac{C_m^2}{C_n^2} = \frac{m(m-1)}{n(n-1)}, \quad p(A_1) = \frac{C_m^1 \cdot C_{n-m}^1}{C_n^2} = \frac{2m(n-m)}{n(n-1)},$$

$$p(A_2) = \frac{C_{n-m}^2}{C_n^2} = \frac{(n-m)(n-m-1)}{n(n-1)}.$$

Соответственно этому энтропия опыта α_{n-2} равна

$$H(\alpha_{n-2}) = -\frac{m(m-1)}{n(n-1)}\log\frac{m(m-1)}{n(n-1)} - \frac{2m(n-m)}{n(n-1)} \times$$

$$\times \log\frac{2m(n-m)}{n(n-1)} - \frac{(n-m)(n-m-1)}{n(n-1)}\log\frac{(n-m)(n-m-1)}{n(n-1)}.$$

Условные энтропии опыта β при условии реализации определенного исхода опыта α_{n-2} будут равны [1])

$$H_{A_0}(\beta) = 0, \quad H_{A_1}(\beta) = 1, \quad H_{A_2}(\beta) = 0,$$

а условная энтропия опыта β при условии осуществления α_{n-2} есть

$$H_{\alpha_{n-2}}(\beta) = \frac{2m(n-m)}{n(n-1)}.$$

Наконец, условная энтропия опыта α_{n-2} при условии осуществления опыта β будет равна

$$H_\beta(\alpha_{n-2}) = H_{\alpha_{n-2}}(\beta) + H(\alpha_{n-2}) - H(\beta).$$

в) Если $m = 1$, то опыт α_k имеет только два исхода A_1 и A_0, отвечающих тем случаям, когда единственный черный шар находится среди k извлеченных шаров или среди $n - k$ оставшихся в урне; вероятности этих исходов равны

$$p(A_1) = \frac{k}{n}, \quad p(A_0) = \frac{n-k}{n}.$$

Условная энтропия опыта β при условии, что опыт α_k имел исход A_1, равна нулю:

$$H_{A_1}(\beta) = 0,$$

ибо исход A_1 опыта α_k, очевидно, однозначно определяет исход опыта β. Условная энтропия опыта β при условии, что опыт α_k имел исход A_0, равна

$$H_{A_0}(\beta) = -\frac{1}{n-k}\log\frac{1}{n-k} - \frac{n-k-1}{n-k}\log\frac{n-k-1}{n-k};$$

она превосходит (безусловную) энтропию того же опыта

$$H(\beta) = -\frac{1}{n}\log\frac{1}{n} - \frac{n-1}{n}\log\frac{n-1}{n}$$

(ибо если среди заключенных в урне шаров лишь один по цвету отличается от остальных, то степень неопределенности опыта, состоящего в извлечении одного шара,

---

[1]) Здесь $H_{A_1}(\beta) \geqslant H(\beta)$, так как опыт β, имеющий два исхода, не может иметь энтропию, превышающую 1 бит.

будет тем меньше, чем больше общее число шаров). Средняя же условная энтропия опыта β

$$H_{\alpha_k}(\beta) = \frac{n-k}{n} H_{A_0}(\beta)$$

меньше (безусловной) энтропии $H(\beta)$.

Если производить много раз пару опытов α и β так, чтобы каждый раз опыт β следовал за опытом α, то условная энтропия $H_\alpha(\beta)$ будет характеризовать ту среднюю

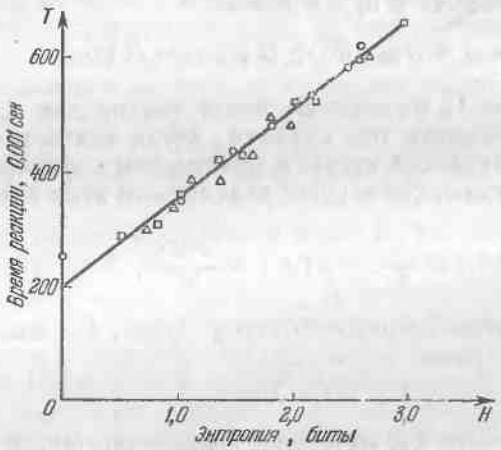


Рис. 12.

степень неопределенности исхода опыта β, которая остается после того, как становится известным исход предшествующего ему опыта α. В частности, в опытах по определению среднего времени реакции (см. выше, стр. 83 и след.) всегда производится целая серия подач сигнала, причем перед каждой из них испытуемый знает, какие сигналы ему подавались ранее. Поэтому степень неопределенности подаваемого сигнала здесь равна условной энтропии соответствующего опыта при условии, что исходы всех предыдущих опытов (т. е. предыдущих подач сигнала) являются известными. В описанных на стр. 83—86 опытах последовательные подачи сигналов всегда выбирались независимыми друг от друга; поэтому в этих

опытах условная энтропия опыта α совпадала с его безусловной энтропией $H(\alpha)$. Если, однако, время реакции действительно определяется степенью неопределенности подаваемого сигнала, измеряемой его энтропией, то из сказанного выше должно следовать, что изменение степени неопределенности при помощи введения зависимости между последовательными подачами сигналов должно оказать то же самое влияние на изменение среднего времени реакции, как такое же изменение степени неопределенности при помощи изменения общего числа используемых равновероятных сигналов или при помощи изменения относительных частот этих сигналов. Результаты проверки этого заключения приведены на рис. 12, заимствованном из той же статьи [38], на которую мы ссылались на стр. 85. На этом чертеже нанесены 8 кружков и 8 квадратиков, которые мы уже видели на рис. 10, и, кроме того, еще 8 треугольников, отвечающих результатам 8 опытов (проведенных над тем же испытуемым, что и раньше), в которых требовалось по-разному реагировать на зажигание каждой из k лампочек (о п ы т β; в разных опытах k принимало значения 2, 3, 4, 5 и 8), зажигавшихся в среднем с одинаковой частотой $p = \frac{1}{k}$, но так, что частота зажигания каждой лампочки существенно зависела от того, какая лампочка зажглась непосредственно перед ней (о п ы т α). На рис. 12 по оси ординат по-прежнему откладывалось среднее время реакции T (получаемое из длинной серии испытаний, проводимых после долгой предварительной тренировки испытуемого при фиксированных условиях зажигания отдельных лампочек), а по оси абсцисс — средняя у с л о в н а я энтропия

$$H_\alpha(\beta) = p(A_1) H_{A_1}(\beta) + p(A_2) H_{A_2}(\beta) + \ldots + p(A_k) H_{A_k}(\beta) = \\ = \frac{1}{k} [H_{A_1}(\beta) + H_{A_2}(\beta) + \ldots + H_{A_k}(\beta)]$$

($A_1, A_2, \ldots, A_k$ — исходы опыта α). То обстоятельство, что на рис. 12 треугольники с большой степенью точности попали на ту же прямую, вокруг которой группируются кружки и квадратики, показывает, что условная энтропия $H_\alpha(\beta)$ действительно является именно той мерой степени неопределенности, которая определяет время реакции человека на исход опыта.

§ 3. Понятие об информации

Вернемся снова к величине $H(\beta)$, характеризующей степень неопределенности опыта β. Равенство этой величины нулю означает, что исход опыта β заранее известен; большее или меньшее значение числа $H(\beta)$ отвечает большей или меньшей проблематичности результата опыта. Какое-либо измерение или наблюдение α, предшествующее опыту β, может ограничить количество возможных исходов опыта β и тем самым уменьшить степень его неопределенности; так, степень неопределенности опыта, состоящего в нахождении самого тяжелого из трех грузов, уменьшается после сравнения на весах двух из них. Для того чтобы результат измерения (наблюдения) α мог сказаться на последующем опыте β, разумеется, необходимо, чтобы этот результат не был известен заранее; поэтому α можно рассматривать как вспомогательный опыт, также имеющий несколько допустимых исходов. Тот факт, что осуществление α уменьшает степень неопределенности β, находит свое отражение в том, что условная энтропия $H_\alpha(\beta)$ опыта β при условии выполнения α оказывается меньше (точнее — не больше) первоначальной энтропии $H(\beta)$ того же опыта. При этом, если опыт β не зависит от α, то осуществление α не уменьшает энтропии β, т. е. $H_\alpha(\beta) = H(\beta)$; если же результат α полностью предопределяет исход β, то энтропия β уменьшается до нуля: $H_\alpha(\beta) = 0$. Таким образом, разность

$$I(\alpha, \beta) = H(\beta) - H_\alpha(\beta)$$

указывает, насколько осуществление опыта α уменьшает неопределенность β, т. е. как много нового узнаем мы об исходе опыта β, произведя измерение (наблюдение) α; эту разность называют **количеством информации относительно опыта β, содержащимся в опыте α**, или, короче, **информацией о β, содержащейся в α**.

Таким образом, мы получаем возможность **численного измерения информации**, что весьма полезно во многих случаях. Так, например, в условиях задачи 18 (стр. 95—97) можно сказать, что используемая реакция дает *информацию* о заболеваниях рассматриваемой болезнью, близкую к 0,14—0,12 = 0,02 (где за еди-

нице принята информация, доставляемая нам одним ответом «да» или «нет» на вопрос, в отношении которого мы заранее склонны были считать утвердительный и отрицательный ответы одинаково вероятными); цифра 0,02 и оценивает пользу реакции. Другие примеры использования понятия количества информации будут приведены в гл. III и IV.

Соотношение между понятиями *энтропии* и *информации* в известном смысле напоминает соотношение между физическими понятиями потенциала и разности потенциалов. Энтропия есть абстрактная «мера неопределенности»; ценность этого понятия в значительной мере заключается в том, что оно позволяет оценить влияние на определенный опыт β какого-либо другого опыта α как «разность энтропий» $I(\alpha, \beta) = H(\beta) - H_\alpha(\beta)$. Так как понятие информации, связанное с определенными изменениями в условиях опыта β, является, так сказать, «более активным», чем понятие энтропии, то для лучшего уяснения смысла энтропии полезно свести это последнее понятие к первому. Энтропию $H(\beta)$ опыта β можно определить как информацию относительно β, содержащуюся в самом этом опыте (ибо осуществление самого опыта β, разумеется, полностью определяет его исход и, следовательно, $H_\beta(\beta) = 0$), или как наибольшую информацию относительно β, какую только можно иметь («полную информацию» относительно β). Иначе говоря, энтропия $H(\beta)$ опыта β равна той информации, которую мы получаем, осуществив этот опыт, т. е. средней информации, содержащейся в одном исходе опыта β [1]). Эти выражения, которые будут

[1]) Заметим, что выражение для энтропии

$$H(\beta) = -p(B_1)\log p(B_1) - p(B_2)\log p(B_2) - \ldots - p(B_l)\cdot\log p(B_l)$$

имеет вид среднего значения случайной величины, принимающей значения $-\log p(B)_1, -\log p(B_2), \ldots, -\log p(B_l)$ с вероятностями, равными соответственно $p(B_1), p(B_2), \ldots, p(B_l)$ (ср. стр. 24). В связи с этим можно считать, что при осуществлении определенного исхода B_i нашего опыта мы получаем информацию, равную $-\log p(B_i)$. В таком случае, если, например, опыт β имеет всего два возможных исхода B_1 и B_2 с вероятностями 0,99 и 0,01, то при осуществлении исхода B_1 мы получим очень небольшую информацию $-\log 0{,}99 \approx 0{,}017$ бит. Это представляется вполне естест-

широко использоваться в гл. III и IV, понятно, имеют тот же смысл, что и «мера неопределенности»: чем больше неопределенность какого-либо опыта, тем бо́льшую информацию дает определение его исхода.

Подчеркнем еще, что информация относительно β, содержащаяся в опыте α, по определению представляет собой среднее значение случайной величины $H(\beta) - H_{A_i}(\beta)$, связанной с отдельными исходами A_i опыта α; поэтому ее можно было бы назвать также «средней информацией относительно β, содержащейся в α». Часто может случиться, что, желая узнать исход какого-либо опыта β, мы можем с этой целью по-разному выбирать вспомогательные опыты (измерения, наблюдения) α; так, например, при нахождении самого тяжелого из определенной системы грузов мы можем в разном порядке сравнивать отдельные грузы. При этом рекомендуется начинать с того опыта α_0, который содержит наибольшую информацию относительно β, ибо при ином опыте α мы вероятно, добьемся менее значительного уменьшения степени неопределенности β (энтропии $H(\beta)$). Реально же вполне может случиться, что опыт α окажется более полезным, чем α_0; может даже оказаться, что исход A опыта α_0 будет настолько неудачен, что энтропия $H_A(\beta)$ окажется больше первоначальной энтропии $H(\beta)$. Такое положение дела является вполне естественным, так как случайный характер исходов опыта β, разумеется, не позволяет заранее указать кратчайший путь к выяснению результата этого опыта: самое большее, на что мы можем рассчиты-

---

венным: в самом деле мы и до опыта знали, что почти наверное осуществится исход B_1, так что результат опыта мало что изменил в имеющихся у нас сведениях. Наоборот, если осуществится исход B_2, то полученная информация будет равна $-\log 0{,}01 = 6{,}6$ бит, т. е. будет гораздо больше, чем в первом случае; это естественно, так как сведения, полученные в результате опыта, здесь много более интересны (осуществилось событие, которое трудно было ожидать). Однако такое большое количество информации при многократном повторении опыта мы будем получать очень редко; поэтому среднее количество информации, содержащееся в одном исходе, оказывается здесь меньшим, чем в том случае, когда вероятности обоих исходов равны. Заметим еще, что в практических задачах нас всегда интересует только это среднее количество информации; представление же о количестве информации, связанном с отдельными исходами опыта, почти никогда не употребляется.

вать — это указать путь, который в е р о я т н е е в с е г о окажется кратчайшим; именно эту возможность и предоставляет теория информации [1]). Отдельные же величины $H(\beta) - H_{A_i}(\beta)$ фактически даже не являются характеристиками опыта β, поскольку если нам известен результат A_i опыта α (и опыты α и β н е п е з а в и с и м ы), то мы тем самым теряем право говорить о первоначальном опыте β, а должны учитывать те изменения в условиях этого опыта, которые вытекают из того, что α имеет исход A_i; таким образом $H_{A_i}(\beta)$ это есть просто энтропия некоторого н о в о г о опыта, к которому сводится опыт β при условии, что реализуется событие A_i.

З а д а ч а 20. *Пусть опыт β состоит в извлечении одного шара из урны, содержащей 5 черных и 10 белых шаров, опыт α_k — в предварительном извлечении из той же урны (без возвращения обратно) k шаров. Чему равна энтропия опыта β и информация об этом опыте, содержащаяся в опытах α_1, α_2, α_{13} и α_{14}?*

Энтропия опыта β, очевидно, равна

$$H(\beta) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} \approx 0{,}92 \text{ бита}.$$

Далее, согласно формулам, полученным в ходе решения задачи 19, имеем (в битах):

$$I(\alpha_1, \beta) = H(\beta) - H_{\alpha_1}(\beta) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\left(\frac{2}{7}\log\frac{2}{7} + \frac{5}{7}\log\frac{5}{7}\right) + \frac{2}{3}\left(\frac{5}{14}\log\frac{5}{14} + \frac{9}{14}\log\frac{9}{14}\right) \approx 0{,}004;$$

[1]) Не следует думать, что методы теории информации ни в каких случаях не позволяют получить такую оценку, скажем, для числа вспомогательных опытов α, необходимых для определения результата определенного опыта β, которая имела бы а б с о л ю т н ы й характер, а не являлась бы лишь наиболее вероятной. Так, например, если информация $I(\alpha, \beta)$ равна энтропии $H(\beta)$ опыта β, то мы можем быть уверены, что п р и л ю б о м и с х о д е о п ы т а α результат β становится полностью определенным (и аналогично этому — если информация $I(\alpha, \beta)$ равна нулю, т о п р и л ю б о м и с х о д е A_i о п ы т а α энтропия $H_{A_i}(\beta)$ равна первоначальной энтропии $H(\beta)$). См. по этому поводу гл. III настоящей книги.

$$I(\alpha_2, \beta) = H(\beta) - H_{\alpha_2}(\beta) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} +$$
$$+ \frac{C_{10}^2}{C_{15}^2}\left(\frac{5}{13}\log\frac{5}{13} + \frac{8}{13}\log\frac{8}{13}\right) +$$
$$+ \frac{C_{10}^1 \cdot C_5^1}{C_{15}^2}\left(\frac{4}{13}\log\frac{4}{13} + \frac{9}{13}\log\frac{9}{13}\right) +$$
$$+ \frac{C_5^2}{C_{15}^2}\left(\frac{3}{13}\log\frac{3}{13} + \frac{10}{13}\log\frac{10}{13}\right) \approx 0{,}008;$$

$$I(\alpha_{13}, \beta) = H(\beta) - H_{\alpha_{13}}(\beta) =$$
$$= -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} - \frac{2\cdot 5\cdot 10}{15\cdot 14} \approx 0{,}44$$

и, наконец,

$$I(\alpha_{14}, \beta) = H(\beta) - H_{\alpha_{14}}(\beta) = H(\beta) \quad (\approx 0{,}92).$$

Задача 21. *Пусть для некоторого пункта вероятность того, что* 15 *июня будет идти дождь, равна* 0,4, *а вероятность того, что дождя не будет, равна* 0,6. *Пусть далее для этого же пункта вероятность дождя* 15 *октября равна* 0,8, *а вероятность отсутствия дождя в этот день — всего* 0,2. *Предположим, что определенный метод прогноза погоды* 15 *июня оказывается правильным в* $^3/_5$ *всех тех случаев, в которых предсказывается дождь, и в* $^4/_5$ *тех случаев, в которых предсказывается отсутствие осадков; в применении же к погоде* 15 *октября этот метод оказывается правильным в* $^9/_{10}$ *тех случаев, в которых предсказывается дождь, и в половине случаев, в которых предсказывается отсутствие дождя* (сравнительно большой процент ошибок в последнем случае естественно объясняется тем, что предсказывается маловероятное событие, предугадать которое довольно трудно). *Спрашивается, в какой из двух указанных дней прогноз дает нам больше информации о реальной погоде?*

Обозначим через β_1 и β_2 опыты, состоящие в определении погоды в рассматриваемом пункте 15 июня и 15 октября. Мы считаем, что эти опыты имеют всего по два исхода — B (дождь) и $\overline{B}$ (отсутствие осадков); соответствующие

таблицы вероятностей имеют вид:

опыт β_1	исходы	B	$\overline{B}$
	вероятн.	0,4	0,6

опыт β_2	исходы	B	$\overline{B}$
	вероятн.	0,8	0,2

Следовательно, энтропии опытов β_1 и β_2 равны

$$H(\beta_1) = -0{,}4 \log 0{,}4 - 0{,}6 \log 0{,}6 \approx 0{,}97 \text{ бита},$$

$$H(\beta_2) = -0{,}8 \log 0{,}8 - 0{,}2 \log 0{,}2 \approx 0{,}72 \text{ бита}.$$

Пусть теперь α_1 и α_2 — предсказания погоды на 15 июня и на 15 октября. Опыты α_1 и α_2 также имеют по два исхода: A (предсказание дождя) и $\overline{A}$ (предсказание сухой погоды); при этом пары опытов (α_1, β_1) и (α_2, β_2) характеризуются следующими таблицами условных вероятностей:

пара (α_1, β_1)	$p_A^{(1)}(B)$	$p_A^{(1)}(\overline{B})$	$p_{\overline{A}}^{(1)}(B)$	$p_{\overline{A}}^{(1)}(\overline{B})$
	0,6	0,4	0,2	0,8

пара (α_2, β_2)	$p_A^{(2)}(B)$	$p_A^{(2)}(\overline{B})$	$p_{\overline{A}}^{(2)}(B)$	$p_{\overline{A}}^{(2)}(\overline{B})$
	0,9	0,1	0,5	0,5

(напомним, что $p_A(B) + p_A(\overline{B}) = p_{\overline{A}}(B) + p_{\overline{A}}(\overline{B}) = 1$). Эти таблицы позволяют определить также и неизвестные нам вероятности $p_1(A)$ и $p_1(\overline{A})$, $p_2(A)$ и $p_2(\overline{A})$ исходов A и $\overline{A}$ опытов α_1 и α_2. В самом деле, по формуле полной вероятности (см. выше, стр. 44) имеем для опыта β_1

$$0{,}4 = p(B) = p_1(A)\, p_A^{(1)}(B) + p_1(\overline{A})\, p_{\overline{A}}^{(1)}(B) = \\ = 0{,}6 \cdot p_1(A) + 0{,}2 \cdot p_1(\overline{A})$$

и для опыта β_2

$$0{,}8 = p(B) = p_2(A)\, p_A^{(2)}(B) + p_2(\overline{A})\, p_{\overline{A}}^{(2)}(B) = \\ = 0{,}9 \cdot p_2(A) + 0{,}5 \cdot p_2(\overline{A}).$$

Так как $p_1(\overline{A}) = 1 - p_1(A)$, $p_2(\overline{A}) = 1 - p_2(A)$, то отсюда получаем

$$p_1(A) = p_1(\overline{A}) = 0{,}5, \quad p_2(A) = 0{,}75, \quad p_2(\overline{A}) = 0{,}25.$$

Подсчитаем теперь энтропии $H_A(\beta_1)$, $H_{\bar{A}}(\beta_1)$, $H_A(\beta_2)$ и $H_{\bar{A}}(\beta_2)$ (в битах):

$$H_A(\beta_1) = -0{,}6 \cdot \log 0{,}6 - 0{,}4 \cdot \log 0{,}4 \approx 0{,}97,$$
$$H_{\bar{A}}(\beta_1) = -0{,}2 \cdot \log 0{,}2 - 0{,}8 \cdot \log 0{,}8 \approx 0{,}72$$

и

$$H_A(\beta_2) = -0{,}9 \cdot \log 0{,}9 - 0{,}1 \cdot \log 0{,}1 \approx 0{,}47,$$
$$H_{\bar{A}}(\beta_2) = -0{,}5 \cdot \log 0{,}5 - 0{,}5 \cdot \log 0{,}5 = 1.$$

Следовательно,

$$H_{\alpha_1}(\beta_1) = p_1(A) H_A(\beta_1) + p_1(\bar{A}) H_{\bar{A}}(\beta_1) \approx 0{,}84,$$
$$H_{\alpha_2}(\beta_2) = p_2(A) H_A(\beta_2) + p_2(\bar{A}) H_{\bar{A}}(\beta_2) \approx 0{,}60.$$

Таким образом, мы видим, что информация, содержащаяся в прогнозе погоды на 15 июня (опыт α_1) о реальной погоде в этот день (об опыте β_1), равна

$$I(\alpha_1, \beta_1) = H(\beta_1) - H_{\alpha_1}(\beta_1) \approx 0{,}97 - 0{,}84 = 0{,}13 \text{ бит},$$

что несколько б о л ь ш е, чем информация о реальной погоде 15 октября (об опыте β_2), содержащаяся в прогнозе погоды на этот день (в опыте α_2):

$$I(\alpha_2, \beta_2) = H(\beta_2) - H_{\alpha_2}(\beta_2) \approx 0{,}72 - 0{,}60 = 0{,}12 \text{ бит}.$$

Этот результат позволяет считать прогноз погоды на 15 июня более ценным, чем прогноз на 15 октября, несмотря на то, что последний прогноз ч а щ е о к а з ы в а е т с я п р а в и л ь н ы м: действительно, в силу формулы полной вероятности, для прогноза погоды на 15 июня вероятность оказаться правильным равна

$$p_1(A) p_A^{(1)}(B) + p_1(\bar{A}) p_{\bar{A}}^{(1)}(\bar{B}) = 0{,}5 \cdot 0{,}6 + 0{,}5 \cdot 0{,}8 = 0{,}7,$$

в то время как для прогноза погоды на 15 октября эта вероятность равна

$$p_2(A) p_A^{(2)}(B) + p_2(\bar{A}) p_{\bar{A}}^{(2)}(\bar{B}) = 0{,}75 \cdot 0{,}9 + 0{,}25 \cdot 0{,}5 = 0{,}8.$$

Вообще, количество информации $I(\alpha, \beta)$, содержащееся в предсказании α исхода некоторого случайного события (или опыта) β, является объективной характеристикой ценности прогноза. Оно равно нулю, если $H_\alpha(\beta) = H(\beta)$, т. е. если события α и β независимы (так что «прогноз» α

никак не связан с событием β) или если $H(\beta) = 0$ (так что исход β известен заранее и не нуждается в предсказании); во всех остальных случаях количество информации положительно, но не больше степени неопределенности $H(\beta)$ опыта β (причем $I(\alpha, \beta) = H(\beta)$, лишь если $H_\alpha(\beta) = 0$, т. е. если «прогноз» α однозначно определяет исход β). Отметим, впрочем, что подобный способ оценки качества л ю б о г о прогноза уже по самой своей универсальности не может охватить все возможные аспекты вопроса. В частности, наша характеристика ценности прогноза полностью игнорирует с о д е р ж а н и е различных исходов прогнозируемого опыта β, а оперирует лишь с в е р о я т н о с т я м и этих исходов. Между тем вполне реальна такая ситуация, когда в силу различного характера разных исходов β один из них значительно важнее предсказать безошибочно, чем другие. Так при предсказании какого-либо стихийного бедствия B (землетрясения, наводнения; в более скромном варианте — заморозков) обычно крайне важно не ошибиться, предсказывая, что B н е п а с т у п и т, в то время как ошибка в предсказании п а с т у п л е н и я B чаще всего является значительно более безобидной (она влечет за собой лишь необоснованное принятие мер предосторожности). Подобные различия между исходами опыта β должны учитываться иными численными характеристиками, отличными от информации I.

В этой связи мы можем повторить относительно информации I то же самое, что было сказано выше (см. стр. 80—81) об энтропии H. Понятие информации возникло непосредственно из задач теории связи и специально было подобрано так, чтобы отвечать запросам этой теории. Поскольку передача по линии связи (например, телеграфной) сообщения определенной длины требует в случае совершенно несущественного или даже лживого сообщения и в случае сообщения о величайшем открытии примерно одинакового времени и одинаковых затрат, то с точки зрения теории связи приходится считать, что и количество информации в этих сообщениях является одинаковым. Разумеется, подобное определение количества информации, полностью отвлекающееся от смыслового содержания рассматриваемого сообщения, не может быть годным во всех случаях, в которых в обыденной жизни употребляется слово «информация». Ясно, однако, что ценность любого

научного понятия определяется никак не количеством случаев, не обслуживаемых этим понятием, а единственно лишь важностью и распространенностью конкретных задач, при решении которых данное понятие оказывается полезным. В отношении же понятия информации таких задач оказывается множество (см., в частности, ниже гл. III и IV).

Задача 22. *Пусть опыт* β *состоит в определении положения некоторой точки* M, *относительно которой заранее известно только, что она расположена на отрезке* AB *длины* L (рис. 13), *а опыт* α — *в измерении длины отрезка* AM *с помощью некоторого измерительного прибора, дающего нам значение длины с точностью до определенной «ошибки измерения»* Δ (например, с помощью линейки, на которой нанесена шкала с делениями длины Δ). *Чему равна информация* $I(\alpha, \beta)$, *содержащаяся в результате измерения, относительно истинного положения точки* M?

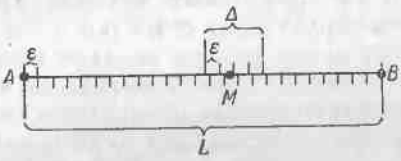


Рис. 13.

С первого взгляда может показаться, что эта задача не может быть решена с помощью приведенных выше формул: ведь в этих формулах речь все время шла об опытах, могущих иметь лишь конечное число исходов, а теперь у нас β может иметь бесконечно много исходов (точка M может совпасть с любой точкой отрезка AB). И действительно, мы не можем здесь приписать опыту β никакой конечной энтропии; тем не менее оказывается, что информация $I(\alpha, \beta)$ (являющаяся разностью двух энтропий $H(\beta)$ и $H_\alpha(\beta)$) в рассматриваемом случае имеет вполне определенное конечное значение. Чтобы пояснить это предположим сначала, что длины L и Δ соизмеримы между собой и разобьем весь отрезок AB на маленькие отрезки длины ε, выбранной так, чтобы и на всем отрезке AB и на отрезке длины Δ уложилось целое число таких малых отрезков (т. е. чтобы отношения L/ε и Δ/ε оба выражались целыми числами). Поставим задачу об определении положения точки M с точностью до величины ε. Так как заранее нам было извест-

но только, что точка M располагается где-то на отрезке AB, то мы можем считать, что опыт β_ε, состоящий в определении ее положения с точностью до ε, имеет L/ε равновероятных исходов, так что его энтропия равна $H(\beta_\varepsilon) = \log \frac{L}{\varepsilon}$. После того как мы произвели опыт α, т. е. измерили длину AM с помощью нашего измерительного прибора, мы выяснили, что на самом деле точка M помещается внутри меньшего интервала длины Δ (определяющего точность измерения); поэтому при известном исходе α опыт β_ε будет иметь уже всего Δ/ε равновероятных исходов, так что $H_\alpha(\beta_\varepsilon) = \log \frac{\Delta}{\varepsilon}$. Следовательно,

$$I(\alpha, \beta_\varepsilon) = H(\beta_\varepsilon) - H_\alpha(\beta_\varepsilon) = \log \frac{L}{\varepsilon} - \log \frac{\Delta}{\varepsilon} = \log \frac{L}{\Delta}.$$

При неограниченном уменьшении ε (т. е. при определении положения нашей точки со все большей и большей точностью) обе энтропии $H(\beta_\varepsilon)$ и $H_\alpha(\beta_\varepsilon)$ будут неограниченно возрастать; однако информация $I(\alpha, \beta_\varepsilon)$ при этом вовсе не меняется, оставаясь все время равной $\log \frac{L}{\Delta}$. Естественно поэтому, что информацию $I(\alpha, \beta)$ (которую мы можем определить, например, как предел $I(\alpha, \beta_\varepsilon)$ при $\varepsilon \to 0$), надо считать также равной $\log \frac{L}{\Delta}$ — это и есть информация относительно истинного положения M, содержащаяся в результате измерения с точностью Δ. При неограниченном увеличении точности прибора (т. е. неограниченном уменьшении Δ) эта информация неограниченно возрастает, однако возрастание это сравнительно медленное: при увеличении точности в n раз мы получаем дополнительно лишь $\log n$ единиц информации (например, при увеличении точности в 2 раза мы выгадываем 1 бит информации, а при увеличении точности в 1000 раз — меньше 10 бит информации).

В наших рассуждениях мы предполагали, что длины L и Δ являются соизмеримыми. Ясно, однако, что это предположение не является существенным; если выбрать ε достаточно малым, то предположение о том, что на отрезках AB и Δ укладывается целое число малых отрезков длины ε, всегда будет выполняться с большой степенью

точности, так что полученный нами результат не может измениться и в случае несоизмеримых L и Δ.

Более подробное обсуждение затронутого в этой задаче вопроса об информации, содержащейся в результате измерения, можно найти в книге Л. Бриллюэна [2].

Отметим еще, что при решении задачи 22 мы встретились с новой для нас ситуацией. Нам пришлось здесь иметь дело с опытом β, имеющим бесконечное число исходов, так что соответствующую энтропию $H(\beta)$ мы должны были считать бесконечной. Для того чтобы подсчитать информацию об этом опыте, содержащуюся в другом опыте α, мы рассмотрели вспомогательный опыт β_ε, получаемый при помощи объединения в один-единственный исход целой группы исходов β, отличающихся друг от друга не больше чем на малую величину ε. При этом оказалось, что как энтропия $H(\beta_\varepsilon)$ этого нового опыта, так и условная энтропия $H_\alpha(\beta_\varepsilon)$ имеют уже конечное значение; так как их разность к тому же оказалась не зависящей от выбора ε, то эту разность мы и приняли за информацию $I(\alpha, \beta)$.

Подобного рода обстановка постоянно встречается при рассмотрении опытов β, имеющих непрерывное множество исходов. Во всех таких случаях энтропия $H(\beta)$ оказывается бесконечной; однако вместо нее часто можно рассматривать конечную энтропию $H(\beta_\varepsilon) = H_\varepsilon(\beta)$, получаемую при объединении исходов β, отличающихся не более чем на некоторое малое ε, в один исход. В практических задачах обычно только энтропия $H_\varepsilon(\beta)$ (называемая ε-энтропией опыта β) и имеет смысл, так как мы вообще не можем различить между собой исходы β, отличающиеся меньше чем на некоторую малую величину (определяемую точностью имеющихся в нашем распоряжении измерительных приборов). К этому вопросу мы еще вернемся ниже (см. стр. 290—292).

Приравняв энтропию $H(\alpha)$ средней информации, содержащейся в исходе опыта α, мы можем, в частности, дать новое истолкование результатам психологических экспериментов, описанных на стр. 183—186 и 102—103. А именно, мы видим теперь, что согласно этим результатам среднее время, требующееся для четкого уяснения значения некоторого сигнала и правильной реакции на него, возра-

стает пропорционально средней информации, содержащейся в этом сигнале. Исходя отсюда, можно предположить, что в случае достаточно регулярно происходящих событий, характеризующихся определенной статистической устойчивостью (т. е. подчиняющихся законам теории вероятностей), сообщение о возникновении такого события передается через органы чувств и нервную систему в среднем за время, пропорциональное содержащейся в этом событии информации. Иначе говоря, можно предположить, что передача сообщений в живом организме во многих случаях происходит так, что *за одинаковое время в среднем передается одинаковое количество информации.* Отметим здесь же, что, как будет видно из содержания гл. IV, такой же закономерностью характеризуется передача сообщений по всем техническим линиям связи.

Из сделанного предположения вытекает простое следствие, которое может быть проверено экспериментально. В самом деле, пусть при проведении опыта по определению среднего времени реакции мы предлагаем испытуемому реагировать очень быстро — быстрее чем он может полностью уяснить себе, какой именно сигнал перед ним появился. Пусть, например, рассматриваемые сигналы состоят в зажигании одной из n лампочек и требуется при зажигании i-й лампочки нажать i-ю кнопку. При уменьшении времени реакции T испытуемый, естественно, будет все чаще и чаще ошибаться — нажимать вместо i-й кнопки какую-нибудь другую, например, j-ю. Это означает, что из-за необходимости очень быстро реагировать он становится не в состоянии полностью воспринять всю информацию, заключающуюся в появлении определенного сигнала. Если, однако, время T не слишком мало, то некоторую полезную информацию о сигнале испытуемый все же успеет уловить; это будет проявляться в том, что реакция его не будет совершенно беспорядочной, а в среднем он будет при зажигании i-й лампочки все же чаще нажимать i-ю кнопку, чем какую-либо другую. Опыт α, состоящий в нажимании испытуемым одной из n кнопок, здесь будет содержать определенную информацию об опыте β, состоящем в зажигании одной из n лампочек; это и будет та средняя информация, которую способен воспринять за время T испытуемый. Согласно нашему предположению эта информация должна так же зависеть от времени

реакции T, как зависит от T энтропия $H(\beta)$ в том случае, когда T определяется как наименьшее время, достаточное для б е з о ш и б о ч н о й реакции.

Проверка последнего заключения была проведена английским психологом У. Х и к о м [39]; полученные им результаты мы изобразили на рис. 14. Кружками здесь

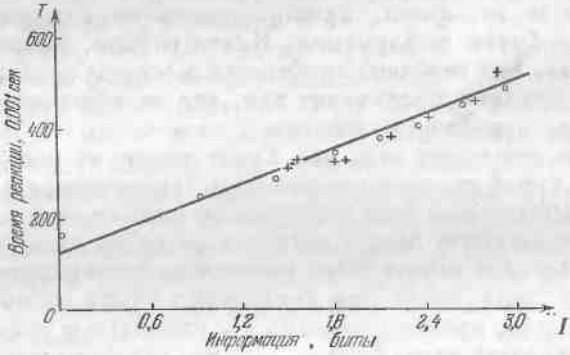


Рис. 14.

обозначено среднее время реакции, определенное из опыта, совпадающего с описанным на стр. 83—85; перед испытуемым (которым в данном случае являлся сам исследователь) зажигались с равными частотами n различных лампочек (где n в разных опытах менялось от 1 до 10) и измерялось среднее время T, требующееся для правильной реакции на поступивший сигнал. Как мы уже знаем, T при этом линейно возрастает с ростом энтропии $H(\beta) = I(\beta, \beta)$; это проявляется в том, что на нашем рисунке, где по оси ординат отложено время T, а по оси абсцисс $I(\beta, \beta)$, все кружки со значительной степенью точности попадают на одну прямую. Крестиками же здесь обозначены результаты опытов, в которых использовались все 10 лампочек, зажигавшихся с одинаковой частотой, но время реакции T заранее устанавливалось столь малым, что реакция испытуемого в ряде случаев поневоле оказывалась ошибочной. Для того чтобы оценить среднюю информацию, содержащуюся в опыте α (нажимании испытуемым одной из 10 кнопок) относительно опыта β (появления одного из 10 сигналов), производилась большая

серия из N испытаний с одним и тем же T и подсчитывалось общее число $n_{i,j}$ всех тех случаев, в которых в ответ на зажигание i-й лампочки была нажата j-я кнопка (i и j принимают всевозможные значения от 1 до 10; при этом сумма всех $n_{i,j}$ равна N, а общее число всех случаев, в которых испытуемый реагировал правильно, равно $n_{1,1} + n_{2,2} + \ldots + n_{10,10}$). Ясно, что вероятности 10 исходов опыта β здесь можно приближенно считать равными

$$q_1 = \frac{n_{1,1} + n_{1,2} + \cdots + n_{1,10}}{N},$$

$$q_2 = \frac{n_{2,1} + n_{2,2} + \cdots + n_{2,10}}{N}, \ldots, q_{10} = \frac{n_{10,1} + n_{10,2} + \cdots + n_{10,10}}{N},$$

а вероятности 10 исходов опыта α — равными

$$p_1 = \frac{n_{1,1} + n_{2,1} + \cdots + n_{10,1}}{N},$$

$$p_2 = \frac{n_{1,2} + n_{2,2} + \cdots + n_{10,2}}{N}, \ldots, p_{10} = \frac{n_{1,10} + n_{2,10} + \cdots + n_{10,10}}{N};$$

сложный опыт $\alpha\beta$ здесь будет иметь $10^2 = 100$ различных исходов, вероятности которых приближенно равны соответствующим частотам

$$p_{1,1} = \frac{n_{1,1}}{N}, \; p_{1,2} = \frac{n_{1,2}}{N}, \ldots, p_{10,10} = \frac{n_{10,10}}{N}.$$

Отсюда для энтропий опытов β, α и $\alpha\beta$ получаются выражения:

$$H(\beta) = -q_1 \log q_1 - q_2 \log q_2 - \ldots - q_{10} \log q_{10},$$

$$H(\alpha) = -p_1 \log p_1 - p_2 \log p_2 - \ldots - p_{10} \log p_{10},$$

$$H(\alpha\beta) = -p_{1,1} \log p_{1,1} - p_{1,2} \log p_{1,2} - \ldots - p_{10,10} \log p_{10,10},$$

позволяющие приближенно подсчитать эти энтропии по определяемым из эксперимента числам $n_{i,j}$. После этого из формулы

$$H(\alpha\beta) = H(\alpha) + H_\alpha(\beta)$$

(см. стр. 91) мы можем определить среднюю условную энтропию $H_\alpha(\beta)$:

$$H_\alpha(\beta) = H(\alpha\beta) - H(\alpha),$$

а по $H(\beta)$ и $H_\alpha(\beta)$ можно найти и информацию $I(\alpha, \beta)$ об опыте β, содержащуюся в опыте α:

$$I(\alpha, \beta) = H(\beta) - H_\alpha(\beta).$$

Это значение информации $I(\alpha, \beta)$ и откладывалось по оси абсцисс на рис. 14 при нанесении на него крестиков.

Отметим, что постановка опыта здесь в некотором смысле обратна той, которая рассматривалась на стр. 83—86 и 102—103: если раньше мы задавали заранее информацию $I(\beta, \beta) = H(\beta)$ и исследовали зависимость от нее времени реакции T, то теперь заранее задается время T (т. е. требуется, чтобы испытуемый реагировал через определенное время T после появления сигнала) и изучается зависимость от этого времени информации $I(\alpha, \beta)$. То обстоятельство, что крестики на рис. 14 группируются вокруг той же прямой, что и кружки, подтверждает предположение о линейной зависимости времени реакции именно от и н ф о р м а ц и и, содержащейся в сигнале.

Разумеется, было бы неоправданным распространять результаты этих нескольких опытов, проводившихся в весьма специфической обстановке, на все вообще процессы передачи информации в живом организме. На самом деле простая линейная зависимость между временем реакции и информацией, содержащейся в поданом сигнале, наблюдается не во всех опытах; кроме того, даже в тех случаях, когда такая зависимость имеет место, коэффициенты соответствующей линейной функции могут принимать весьма разные значения (см., например, сводный рис. 15, заимствованный из статьи В. И. Н и к о л а е в а [40]; ср. также книгу [41], включающую список литературы, содержащий более 500 названий). Факторы, от которых зависят эти коэффициенты, изучались многими авторами (см., например, обзорные статьи [42] — [44]); однако в этой области остается еще очень много открытых вопросов. Тем не менее имеющиеся данные (по поводу которых см. также книгу [45] и сборник [46]) определенно показывают, что введенное выше понятие информации часто может быть с успехом использовано для количественного описания процессов восприятия и усвоения живыми организмами сигналов разного рода, поступающих к ним из внешнего мира.

Докажем теперь, что *информация относительно опыта* β, *содержащаяся в опыте* α, *всегда равна информации относительно* α, *содержащейся в* β. Это непосредственно следует из результатов предыдущего параграфа: так как

$$H(\alpha) + H_\alpha(\beta) = H(\beta) + H_\beta(\alpha)$$

(см. выше, стр. 95), то

$$I(\alpha, \beta) = H(\beta) - H_\alpha(\beta) = H(\alpha) - H_\beta(\alpha) = I(\beta, \alpha).$$

Таким образом, информацию $I(\alpha, \beta)$, которую содержит

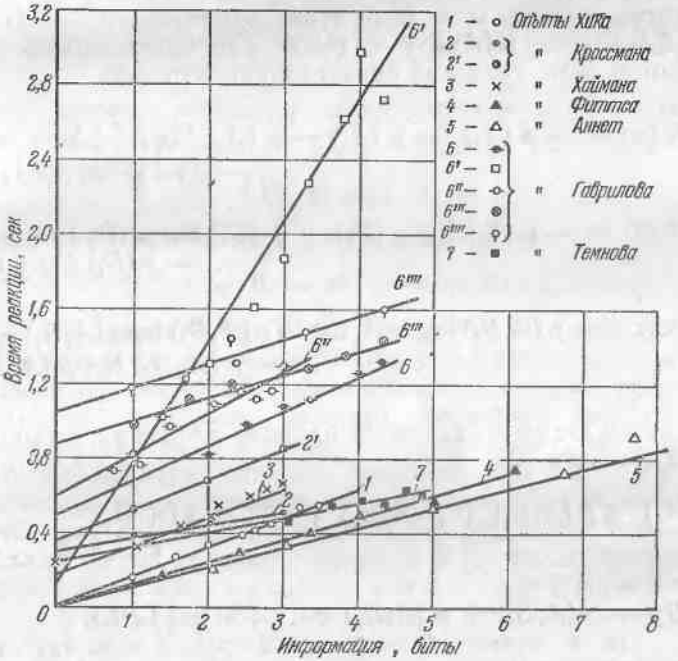


Рис. 15.

опыт α относительно опыта β, можно также назвать взаимной информацией двух опытов α и β друг относительно друга. Равенство информаций $I(\alpha, \beta)$ и $I(\beta, \alpha)$ подчеркивается следующей простой формулой, которая во многих случаях оказывается

весьма удобной

$$I(\alpha, \beta) = H(\alpha) + H(\beta) - H(\alpha\beta)$$

(ср., например, выше, стр. 117—118). Эта формула вытекает из того, что $H_\alpha(\beta) = H(\alpha\beta) - H(\alpha)$ (ибо $H(\alpha\beta) = H(\alpha) + H_\alpha(\beta)$); в ее правую часть опыты α и β входят совершенно симметрично.

Приведенную здесь симметричную формулу для количества информации можно преобразовать так, чтобы ее правая часть непосредственно выражалась через вероятности $p(A_1), \ldots, p(A_k)$ и $p(B_1), \ldots, p(B_l)$ различных исходов опытов α и β и через вероятности $p(A_1B_1)$, $p(A_1B_2), \ldots, p(A_kB_l)$ исходов сложного опыта $\alpha\beta$. В самом деле, согласно определению энтропии

$$H(\alpha) = -p(A_1)\log p(A_1) - p(A_2)\log p(A_2) - \ldots \\ \ldots - p(A_k)\log p(A_k),$$

$$H(\beta) = -p(B_1)\log p(B_1) - p(B_2)\log p(B_2) - \ldots \\ \ldots - p(B_l)\log p(B_l)$$

и

$$H(\alpha\beta) = -p(A_1B_1)\log p(A_1B_1) - p(A_1B_2)\log p(A_1B_2) - \ldots \\ \ldots - p(A_kB_l)\log p(A_kB_l).$$

С другой стороны, согласно правилу сложения вероятностей (см. стр. 27)

$$p(A_i) = p(A_iB_1) + p(A_iB_2) + \ldots + p(A_iB_l), \quad i = 1, 2, \ldots, k$$

и

$$p(B_j) = p(A_1B_j) + p(A_2B_j) + \ldots + p(A_kB_j), \quad j = 1, 2, \ldots, l,$$

так что

$$-p(A_i)\log p(A_i) = -p(A_iB_1)\log p(A_i) - \\ - p(A_iB_2)\log p(A_i) - \ldots - p(A_iB_l)\log p(A_i),$$

$$-p(B_j)\log p(B_j) = -p(A_1B_j)\log p(B_j) - \\ - p(A_2B_j)\log p(B_j) - \ldots - p(A_kB_j)\log p(B_j).$$

Подставив все эти выражения в исходную формулу, получим

$$I(\alpha, \beta) = -p(A_1B_1)[\log p(A_1) + \log p(B_1) - \log p(A_1B_1)] - p(A_1B_2)[\log p(A_1) + \log p(B_2) - \log p(A_1B_2)] - \ldots\ldots\ldots\ldots\ldots\ldots\ldots - p(A_kB_l)[\log p(A_k) + \log p(B_l) - \log p(A_kB_l)]$$

или, окончательно,

$$I(\alpha, \beta) = p(A_1B_1)\log\frac{p(A_1B_1)}{p(A_1)\,p(B_1)} + p(A_1B_2)\log\frac{p(A_1B_2)}{p(A_1)\,p(B_2)} + \cdots + p(A_kB_l)\log\frac{p(A_kB_l)}{p(A_k)\,p(B_l)}.$$

Эта формула, очевидно, также симметрична относительно опытов α и β.

Формулу

$$I(\alpha, \beta) = I(\beta, \alpha)$$

можно также записать в следующем виде:

$$I(\alpha, \beta) = H(\alpha) - H_\beta(\alpha).$$

Из нее следует, что *информация* $I(\alpha, \beta)$, *содержащаяся в опыте* α *относительно опыта* β, *не превосходит энтропии* $H(\alpha)$ *опыта* α — обстоятельство, которое часто оказывается полезным. Впрочем, последнее предложение, разумеется, нельзя считать неожиданным: естественно, что информация, которую содержит опыт α о другом опыте β, не превосходит информации, содержащейся в α относительно самого себя — энтропии $H(\alpha)$ этого опыта. Таким образом, энтропия $H(\alpha)$ может быть также определена как наибольшая информация, которая может содержаться в опыте α («полная информация», содержащаяся в α).

Из формулы $I(\alpha, \beta) = H(\alpha) - H_\beta(\alpha)$ вытекает также, что *информация* $I(\alpha, \beta)$ *точно равна энтропии* $H(\alpha)$ *опыта* α *в том и только в том случае, когда условная энтропия* $H_\beta(\alpha)$ *равна нулю*, т. е. *когда результат опыта* β *полностью определяет исход вспомогательного опыта* α; именно так будет обстоять дело, например, в задачах, разбираемых в следующей главе. Если же $H_\beta(\alpha) \neq 0$, то

информация $I(\alpha, \beta)$ *будет ровно на величину* $H_\beta(\alpha)$ *меньше энтропии* $H(\alpha)$. В частности, *если опыты* α *и* β *независимы* (*и только в этом случае*) *информация* $I(\alpha, \beta)$ *будет равна нулю*.

Заметим еще, что если условная энтропия $H_\beta(\alpha)$ равна нулю и, следовательно, информация $I(\alpha, \beta)$ относительно β, содержащаяся в α, является наибольшей (т. е. ни про какой другой опыт β_1 опыт α не содержит большей информации), то *информация относительно любого независимого от* β *опыта* γ, *содержащаяся в* α, *равна нулю* — это дает основание говорить, что опыт α «прямо направлен» к выяснению исхода β и не содержит никакой «посторонней» информации. В общем же случае *информация относительно любого независимого от* β *опыта* γ, *содержащаяся в опыте* α, *не превосходит величины* $H_\beta(\alpha) = I(\alpha, \alpha) - I(\alpha, \beta)$; если $H_\beta(\alpha) = 0$, то это утверждение обращается в более частный результат, указанный выше. Доказательство сделанного утверждения требует введения одного важного вспомогательного понятия; оно будет приведено (вместе с доказательством других сформулированных ниже утверждений) в самом конце параграфа.

Пусть теперь α, β и γ — три произвольных опыта. В таком случае всегда

$$I(\beta\gamma, \alpha) \geqslant I(\beta, \alpha);$$

иначе говоря, *сложный опыт* $\beta\gamma$ (т. е. пара опытов β и γ) *всегда содержит не меньшую информацию относительно любого опыта* α, *чем простой опыт* β. Этот факт представляется вполне естественным с точки зрения наших наглядных представлений об «информации»; строгое доказательство этого и подобных ему предложений и делает законным употребление слова «информация» в применении к величине $I(\alpha, \beta)$. При этом равенство $I(\beta\gamma, \alpha) = I(\beta, \alpha)$ будет иметь место лишь в том случае, когда условная вероятность любого исхода опыта α при условии, что опыты β и γ имеют некоторые определенные исходы, не изменяется при изменении исхода γ (т. е. зависит лишь от исхода β). В этом последнем случае совершенно естественно считать, что сложный опыт $\beta\gamma$ не содержит никакой дополнительной информации относительно α по сравнению с опытом β, так что равенство $I(\beta\gamma, \alpha) = I(\beta, \alpha)$ здесь также

находится в полном соответствии с наглядными представлениями о понятии «информация».

Предположим теперь, что равенство $I(\beta\gamma, \alpha) = I(\beta,\alpha)$ имеет место. Можно доказать, что в этом случае всегда

$$I(\gamma, \alpha) \leqslant I(\beta, \alpha).$$

Таким образом, *если сложный опыт* $\beta\gamma$ *не содержит никакой дополнительной информации об* α *по сравнению с опытом* β, *то информация об* α, *содержащаяся в опыте* γ, *не может быть больше информации об* α, *содержащейся в опыте* β. При этом знак «меньше или равно» в последнем неравенстве можно заменить знаком равенства в том и только в том случае, когда $I(\beta\gamma, \alpha) = I(\gamma, \alpha)$, т. е. когда сложный опыт $\beta\gamma$ не содержит дополнительной информации об α также и по сравнению с опытом γ.

Неравенство $I(\gamma, \alpha) \leqslant I(\beta, \alpha)$, о котором здесь идет речь, играет в теории информации значительную роль (см., например, [8] и [47], а также гл. IV этой книги). Оно показывает, что при последовательной передаче информации об опыте α, осуществляемой посредством цепочки опытов $\beta, \gamma, \delta, \ldots$, где только опыт β непосредственно связан с α, а γ всю содержащуюся в нем информацию об α получает из связи с опытом β (так что $\beta\gamma$ уже не содержит об α дополнительной информации по сравнению с β), δ всю информацию об α получает из связи с опытом γ и т. д., информация об α может лишь уменьшаться:

$$H(\alpha) = I(\alpha, \alpha) \geqslant I(\beta, \alpha) \geqslant I(\gamma, \alpha) \geqslant I(\delta, \alpha) \geqslant \ldots$$

Наглядной иллюстрацией этого положения может служить известная детская игра в «испорченный телефон», при которой первый играющий тихо произносит на ухо своему соседу некоторое слово (опыт α); сосед тихо передает расслышанное им слово (которое может и отличаться от первоначально произнесенного) следующему играющему (опыт β); этот играющий также передает услышанное слово соседу (опыт γ) и т. д.; в конце игры все говорят услышанные ими слова, и проигравшим считается тот из участников, кто первым неправильно услышал передаваемое слово. В этой игре может случиться так, что второй играющий передает первоначально сказанное слово неправильно, а третьему в результате повторной ошибки покажется, что он услышал то же слово, которое

передавалось вначале; однако при большом числе повторений той же процедуры второй играющий, разумеется, в среднем будет чаще передавать дальше слово, которое на самом деле произнес первый игрок, чем третий играющий. Но наше понятие информации I как раз и является статистическим понятием, характеризующим соотношения, имеющие место «в среднем»; поэтому для него всегда будет выполняться выписанная выше цепь неравенств. Ясно, что с точки зрения наглядных представлений о передаче информации это обстоятельство также можно считать очевидным.

Неравенства

$$I(\beta\gamma, \alpha) \geqslant I(\beta, \alpha) \text{ и } I(\beta\gamma, \alpha) > I(\gamma, \alpha)$$

(см. стр. 22) можно дополнить еще одним неравенством, несколько менее очевидным с точки зрения интуитивно ожидаемых свойств величины, получившей название «информации». Ясно, что, вообще говоря, вполне может иметь место неравенство

$$I(\beta\gamma, \alpha) < I(\beta, \alpha) + I(\gamma, \alpha);$$

так, например, если $\beta = \gamma$, то и $\beta\gamma = \beta$, — и потому в таком случае, вообще говоря,

$$I(\beta\gamma, \alpha) = I(\beta, \alpha) < I(\beta, \alpha) + I(\gamma, \alpha) = 2I(\beta, \alpha).$$

Однако если опыты β и γ *независимы* (т. е. $I(\beta, \gamma) = I(\gamma, \beta) = 0$), то неравенство $I(\beta\gamma, \alpha) < I(\beta, \alpha) + I(\gamma, \alpha)$ является невозможным: в этом случае обязательно

$$I(\beta\gamma, \alpha) \geqslant I(\beta, \alpha) + I(\gamma, \alpha).$$

Невозможность неравенства $I(\beta, \alpha) + I(\gamma, \alpha) > I(\beta\gamma, \alpha)$ здесь объясняется тем, что независимость опытов β и γ гарантирует отсутствие «общей части» величин $I(\beta, \alpha)$ и $I(\gamma, \alpha)$: ведь здесь опыты β и α доставляют существенно *разную* информацию об опыте α и связанная с выполнением сразу обоих опытов β и γ информация $I(\beta\gamma, \alpha)$ не может оказаться меньше суммы информаций $I(\beta, \alpha)$ и $I(\gamma, \alpha)$. Это можно сравнить с неравенством

площадь $(F_1 + F_2)$ < площадь F_1 + площадь F_2,

где $F_1 + F_2$ — *объединение* фигур F_1 и F_2, которое, очевидно, невозможно, если F_1 и F_2 не имеют общей части. Однако здесь, казалось бы, можно ожидать равенства

$$I(\beta\gamma, \alpha) = I(\beta, \alpha) + I(\gamma, \alpha),$$

ибо неясно, за счет чего информация $I(\beta\gamma, \alpha)$ может оказаться *больше* суммы информаций $I(\beta, \alpha)$ и $I(\gamma, \alpha)$.

Дело, однако, заключается в том, что даже в случае независимости опытов β и γ их совместное осуществление (т. е. опыт $\beta\gamma$), позволяющее *сразу* узнать и исход β, и исход γ, вообще говоря,

может доставить больше информации, чем раздельное осуществление β и γ (с которым связана величина $I(\beta, \alpha) + I(\gamma, \alpha)$). Проиллюстрировать это можно на примере, изложенном мелким шрифтом на стр. 45—46. Рассмотрим опять изображенный на рис. 2 тетраэдр, и пусть α (соответственно β и γ) — опыт, состоящий в проверке того, имеется ли или отсутствует на той грани, на которую упал тетраэдр, цифра 1 (соответственно 2 и 3); таким образом, опыт α может иметь исходы A и $\bar{A}$, опыт β — исходы B и $\bar{B}$, а опыт γ — исходы C и $\bar{C}$. Из приведенных на стр. 46 подсчетов сразу следует, что все три опыта α, β и γ являются попарно независимыми; таким образом, здесь

$$I(\beta, \alpha) = 0 \text{ и } I(\gamma, \alpha) = 0, \text{ так что } I(\beta, \alpha) + I(\gamma, \alpha) = 0.$$

С другой стороны, результат сложного опыта $\beta\gamma$ уже полностью определяет исход α (опыт α имеет исход A, если β и γ имеют «одинаковые» исходы, т. е. и β и γ имеют «положительные» исходы B и, соответственно, C или же и β и γ имеют «отрицательные» исходы $\bar{B}$ и, соответственно, $\bar{C}$; опыт α имеет исход $\bar{A}$, если β и γ имеют «различные» исходы, т. е. B и $\bar{C}$ или $\bar{B}$ и C). Таким образом, здесь

$$I(\beta\gamma, \alpha) = H(\alpha) = 1 \text{ бит},$$

т. е.

$$I(\beta\gamma, \alpha) > I(\beta, \alpha) + I(\gamma, \alpha) \quad (= 0).$$

Более того, здесь опыты β и γ не содержат н и к а к о й информации об α, а опыт $\beta\gamma$ содержит об α «полную» информацию, т. е. н а и б о л ь ш у ю информацию, какую только об α можно иметь.

Доказательство приведенных выше утверждений может быть получено на основе изучения величины

$$I_\beta(\gamma, \alpha) = H_\beta(\alpha) - H_{\beta\gamma}(\alpha),$$

которую мы будем называть с р е д н е й у с л о в н о й и н ф о р м а ц и е й д в у х о п ы т о в γ и α д р у г д р у г а о т н о с и т е л ь н о п р и у с л о в и и в ы п о л н е н и я о п ы т а β или, короче, просто у с л о в н о й и н ф о р м а ц и е й о п ы т о в γ и α п р и у с л о в и и β. Прежде всего отметим, что условная информация $I_\beta(\gamma, \alpha)$ всегда неотрицательна. Этот факт немедленно вытекает из неравенства

$$H_{\beta\gamma}(\alpha) \leqslant H_\beta(\alpha),$$

означающего, что предварительное осуществление сложного опыта $\beta\gamma$ (т. е. двух опытов β и γ) всегда уменьшает степень неопределенности опыта α не в меньшей степени, чем осуществление одного опыта β (строгое доказательство этого неравенства см. в Приложении I в конце книги). Так как, кроме того, всегда $H_{\beta\gamma}(\alpha) \geqslant 0$ (ибо $H_{\beta\gamma}(\alpha)$ — это некоторая условная энтропия), то

$$0 \leqslant I_\beta(\gamma, \alpha) \leqslant H_\beta(\alpha).$$

При этом $I_\beta(\gamma, \alpha) = H_\beta(\alpha)$ лишь если $H_{\beta\gamma}(\alpha) = 0$, т. е. если сложный опыт $\beta\gamma$ однозначно определяет исход опыта α; $I_\beta(\gamma, \alpha) = 0$

лишь если $H_{\beta\gamma}(\alpha) = H_\beta(\alpha)$, а следовательно, и $I(\beta\gamma, \alpha) = I(\beta, \alpha)$, т. е. если условные вероятности всех исходов опыта α при условии, что опыты β и γ имеют некоторые определенные исходы, не зависят от исхода γ (см. конец Приложения I).

Докажем теперь, что условная информация симметрична:

$$I_\beta(\gamma, \alpha) = I_\beta(\alpha, \gamma)$$

(это обстоятельство подчеркивается и самим названием «условная информация опытов γ и α друг относительно друга»). В самом деле, по определению

$$I_\beta(\gamma, \alpha) = H_\beta(\alpha) - H_{\beta\gamma}(\alpha),\ I_\beta(\alpha, \gamma) = H_\beta(\gamma) - H_{\beta\alpha}(\gamma).$$

Но так как сложный опыт $\alpha\beta\gamma$, состоящий в осуществлении трех опытов α, β и γ, можно с одинаковым правом рассматривать как объединение сложного опыта $\alpha\beta$ и опыта γ или же как объединение опыта α и сложного опыта $\beta\gamma$ [1]), то

$$H(\alpha\beta\gamma) = H(\alpha\beta) + H_{\alpha\beta}(\gamma) = H(\beta) + H_\beta(\alpha) + H_{\alpha\beta}(\gamma)$$

и

$$H(\alpha\beta\gamma) = H(\beta\gamma) + H_{\beta\gamma}(\alpha) = H(\beta) + H_\beta(\gamma) + H_{\beta\gamma}(\alpha).$$

Следовательно,

$$H_\beta(\alpha) + H_{\alpha\beta}(\gamma) = H_\beta(\gamma) + H_{\beta\gamma}(\alpha),$$

т. е.

$$I_\beta(\gamma, \alpha) = H_\beta(\alpha) - H_{\beta\gamma}(\alpha) = H_\beta(\gamma) - H_{\alpha\beta}(\gamma) = I_\beta(\alpha, \gamma).$$

Равенство $I_\beta(\gamma, \alpha) = I_\beta(\alpha, \gamma)$ вытекают также из следующей «симметричной записи» условной информации $I_\beta(\gamma, \alpha)$, легко проверяемой непосредственно: если A_i (где $i = 1, 2, \ldots, l$), B_j (где $j = 1, 2, \ldots, m$) и C_k (где $k = 1, 2, \ldots, n$) — всевозможные исходы опытов α, β и γ, то

$$I_\beta(\gamma, \alpha) = p(B_1) I_{B_1}(\gamma, \alpha) + p(B_2) I_{B_2}(\gamma, \alpha) + \ldots + p(B_m) I_{B_m}(\gamma, \alpha),$$

где

$$I_{B_j}(\gamma, \alpha) = p_{B_j}(A_1C_1) \log \frac{p_{B_j}(A_1C_1)}{p_{B_j}(A_1)\, p_{B_j}(C_1)} + \ldots + p_{B_j}(A_lC_n) \log \frac{p_{B_j}(A_lC_n)}{p_{B_j}(A_l)\, p_{B_j}(C_n)}$$

---

[1]) Символически это можно записать равенствами

$$\alpha\beta\gamma = (\alpha\beta)\gamma = \alpha(\beta\gamma)$$

(ср. с «ассоциативным законом» умножения событий на стр. 61 § 4 гл. I).

— взаимная информация опытов α и γ при условии, что опыт β имел исход B_j. Такая запись хорошо поясняет смысл условной информации $I_\beta(\gamma, \alpha)$; нам она, однако, не понадобится.

Из формулы $I_\beta(\gamma, \alpha) = H_\beta(\alpha) - H_{\beta\gamma}(\alpha)$ легко получается важное соотношение

$$I(\beta\gamma, \alpha) = I(\beta, \alpha) + I_\beta(\gamma, \alpha),$$

близкое по форме к равенству $H(\beta\gamma) = H(\beta) + H_\beta(\gamma)$ (это соотношение следует из того, что $I(\beta\gamma, \alpha) = H(\alpha) - H_{\beta\gamma}(\alpha)$ и $I(\beta, \alpha) = H(\alpha) - H_\beta(\alpha)$). Ясно, что наши утверждения, касающиеся величины информации $I(\beta\gamma, \alpha)$, являются автоматическими следствиями этого соотношения и свойств условной информации.

В дальнейшем нам будет полезна еще следующая ф о р м у л а т р о й н о й и н ф о р м а ц и и:

$$I(\beta\gamma, \alpha) + I(\beta, \gamma) = I(\alpha\gamma, \beta) + I(\alpha, \gamma).$$

Для доказательства этой формулы надо только поменять местами опыты β и γ в полученном выражении для $I(\beta\gamma, \alpha)$ и использовать аналогичное выражение для $I(\alpha\gamma, \beta)$, после чего для правой и левой частей нашей формулы получаются одинаковые выражения

$$I(\beta\gamma, \alpha) + I(\beta, \gamma) = I(\gamma, \alpha) + I_\gamma(\beta, \alpha) + I(\beta, \gamma)$$

и

$$I(\alpha\gamma, \beta) + I(\alpha, \gamma) = I(\gamma, \beta) + I_\gamma(\alpha, \beta) + I(\alpha, \gamma).$$

Из формулы тройной информации сразу следует указанный выше результат о содержащейся в третьем опыте α информации относительно опыта γ, независимого от некоторого другого опыта β. В самом деле, независимость опытов β и γ означает, что $I(\beta, \gamma) = 0$; с другой стороны, мы знаем, что всегда $I(\alpha\gamma, \beta) \geqslant I(\alpha, \beta)$. В силу формулы тройной информации отсюда вытекает, что в случае независимости опытов β и γ

$$I(\alpha, \gamma) = I(\beta\gamma, \alpha) - I(\alpha\gamma, \beta) \leqslant I(\beta\gamma, \alpha) - I(\alpha, \beta) = I_\beta(\gamma, \alpha),$$

а $I_\beta(\gamma, \alpha)$ всегда не больше, чем $H_\beta(\alpha)$. С другой стороны, воспользовавшись «симметричностью» информации (т. е. равенством $I(\alpha, \beta) = I(\beta, \alpha)$), мы можем переписать формулу тройной информации так:

$$I(\beta\gamma, \alpha) + I(\beta, \gamma) = I(\beta, \alpha\gamma) + I(\gamma, \alpha),$$

а неравенство $I(\alpha\gamma, \beta) \geqslant I(\alpha, \beta)$ заменить следующим:

$$I(\beta, \alpha\gamma) \geqslant I(\beta, \alpha),$$

откуда сразу следует, что в случае независимости опытов β и γ (т. е. если $I(\beta, \gamma) = 0$)

$$I(\beta\gamma, \alpha) \geqslant I(\beta, \alpha) + I(\gamma, \alpha)$$

(ср. выше, стр. 124).

Неравенство $I(\gamma, \alpha) \leqslant I(\beta, \alpha)$ для случая, когда $I_\beta(\gamma, \alpha) = 0$, также может быть получено из формулы тройной информации.

Для его вывода надо лишь заменить в этой формуле $I(\alpha\gamma, \beta)$ на $I(\gamma,\beta) + I_\gamma(\alpha, \beta)$ и учесть, что в нашем случае $I(\beta\gamma, \alpha) = I(\beta, \alpha)$, и что информация всегда симметрична; после этого мы придем к соотношению

$$I(\beta, \alpha) = I(\gamma, \alpha) \dotplus I_\gamma(\alpha, \beta),$$

сразу показывающему, что наше неравенство имеет место. Мы видим также, что это неравенство обращается в равенство тогда и только тогда, когда $I_\gamma(\alpha, \beta) = 0$. В этом случае $I(\gamma, \alpha) = I(\beta\gamma, \alpha)$, т. е. сложный опыт $\beta\gamma$ не содержит никакой дополнительной информации относительно α и по сравнению с γ — обстоятельство, которое тоже было отмечено выше.

В заключение обратим еще внимание на то, что неравенство $I(\beta\gamma, \alpha) \geqslant I(\beta, \alpha)$ («сложный опыт $\beta\gamma$ содержит о любом опыте α не меньшую информацию, чем простой опыт β») можно в известном смысле сопоставить неравенству для энтроний $H(\beta\gamma) \geqslant H(\beta)$ («степень неопределенности сложного опыта $\beta\gamma$ всегда не меньше степени неопределенности простого опыта β»). Однако в случае энтроний существует еще и оценка величины $H(\beta\gamma)$ с другой стороны: $H(\beta\gamma) \leqslant H(\beta) + H(\gamma)$ («степень неопределенности сложного опыта $\beta\gamma$ всегда не больше суммы степеней неопределенности опытов β и γ»). В случае информации положение оказывается совсем иным: зная количество информации об опыте α, содержащееся в двух опытах β и γ, мы не можем оценить сверху информацию об α, содержащуюся в сложном опыте $\beta\gamma$. Так, в случае, разобранном на стр. 125 (где опыты α, β и γ состояли в выяснении того, имеется ли на грани, на которую упал изображенный на рис. 2 тетраэдр, цифра 1, соответственно, 2 и 3) мы имели

$$I(\beta, \alpha) = I(\gamma, \alpha) = 0, \text{ а } I(\beta\gamma, \alpha) = 1 \ (= H(\alpha));$$

поэтому из того, что информации $I(\beta, \alpha)$ и $I(\gamma, \alpha)$ обе малы, еще никак нельзя заключить, что также и информация $I(\beta\gamma, \alpha)$ будет малой.

§ 4. Определение энтропии перечислением ее свойств

Основным понятием этой главы является введенное в § 1 понятие энтропии или меры степени неопределенности опыта α, могущего иметь в зависимости от случая тот или иной исход. Целесообразность принятого определения энтропии в § 1 как-то мотивировалась; однако приведенная там аргументация имела лишь наводящий характер и настоящим оправданием такого определения меры степени неопределенности явился лишь весь ряд теорем, доказанных в этой главе, а также в гл. IV и в Приложении I к книге. Сейчас мы снова вернемся к определению энтропии и покажем, что оно с необходимостью вытекает из простейших требований, которые естественно наложить на величину, призванную служить количественной характеристикой степени неопределенности.

Естественно считать, что энтропия (мера степени неопределенности) $H(\alpha)$ опыта α, которому отвечает таблица вероятностей

исходы опыта	A_1	A_2	...	A_k
вероятности	$p(A_1)$	$p(A_2)$	...	$p(A_k)$

зависит лишь от величин $p(A_1)$, $p(A_2)$, . . . , $p(A_k)$ (является функцией этих величин). Мы обозначим здесь вероятности $p(A_1)$, $p(A_2)$, . . ., $p(A_k)$ через $p_1, p_2, \ldots, p_k$, а энтропию $H(\alpha)$ — через $H(p_1, p_2, \ldots, p_k)$ (ср. выше, стр. 75).

Сформулируем теперь те условия, выполнения которых естественно требовать от функции $H(p_1, p_2, \ldots, p_k)$. Прежде всего ясно, что эта функция не должна зависеть от п о р я д к а чисел $p_1, p_2, \ldots, p_k$: ведь изменение порядка этих чисел означает лишь изменение столбцов в таблице вероятностей и не связано с каким бы то ни было изменением самого опыта α. Таким образом первое условие гласит:

1°. *Значение функции* $H(p_1, p_2, \ldots, p_k)$ *не меняется при любой перестановке чисел* $p_1, p_2, \ldots, p_k$.

Столь же естественно и второе условие:

2°. *Функция* $H(p_1, p_2, \ldots, p_k)$ *является непрерывной*, т. е. мало меняется при малых изменениях вероятностей p_1, $p_2, \ldots, p_k$ — ведь при малых изменениях вероятностей и степень неопределенности опыта должна мало изменяться.

Третье условие, которое мы введем, является несколько более сложным. Для того чтобы яснее понять, в чем оно заключается, предположим сначала, что рассматриваемый опыт α имеет всего три исхода A_1, A_2, A_3, т. е. что его таблица вероятностей имеет вид

исходы опыта	A_1	A_2	A_3
вероятности	p_1	p_2	p_3

Мера неопределенности $H(\alpha)$ этого опыта, равна $H(p_1, p_2, p_3)$; неопределенность здесь возникает из-за того, что мы не знаем, какой именно из трех исходов опыта α будет иметь место. Будем теперь выяснять, какой из исходов опыта α на самом деле имел место, в два этапа. А именно выясним прежде всего, имел ли место один из первых двух исходов A_1 и A_2 или же последний исход A_3; это означает, что наш опыт α мы заменим новым опытом β с таблицей вероятностей

исходы опыта	B	A_3
вероятности	p_1+p_2	p_3

Мера неопределенности этого нового опыта, очевидно, равна $H(\beta) = H(p_1 + p_2, p_3)$. Ясно, что мера неопределенности α должна быть больше, чем мера неопределенности β — это связано с тем, что знание исхода β еще не определяет полностью исхода α, так как и

после определения исхода β может остаться еще некоторая неопределенность в исходе α.

Нетрудно ответить и на вопрос о том, на с к о л ь к о и м е н н о мера неопределенности α должна быть больше меры неопределенности β. Ясно, что если мы будем повторять опыт α многократно и каждый раз будем сначала выяснять, имел ли опыт β исход B или A_3, то в некоторых случаях — в тех, когда опыт α имеет исход A_3,— это выяснение решит вопрос и об исходе α. В других случаях — а именно, когда опыт α имеет исход A_1 или A_2,— нам после установления исхода β придется определить, какой именно из этих двух исходов имел опыт α, что равносильно выяснению исхода нового опыта β' с таблицей вероятностей

исходы опыта	A_1	A_2
вероятности	$\frac{p_1}{p_1+p_2}$	$\frac{p_2}{p_1+p_2}$

Мера неопределенности этого опыта β', очевидно, равна $H(\beta') = H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$. А так как вероятность (т. е. средняя частота) случаев, в которых приходится после выполнения β дополнительно выяснять еще и исход опыта β', равна p_1+p_2, то естественно считать, что мера неопределенности $H(\alpha)$ опыта α должна превосходить меру неопределенности $H(\beta)$ опыта β на величину $(p_1+p_2)H(\beta')$, т. е. что должно выполняться равенство

$$H(p_1, p_2, p_3) = H(p_1+p_2, p_3) + (p_1+p_2)\cdot H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right).$$

Те же соображения, примененные к опыту α с таблицей вероятностей

исходы опыта	A_1	A_2	A_3	...	A_k
вероятности	p_1	p_2	p_3	...	p_k

приводят к третьему свойству функции $H(p_1, p_2, \ldots, p_k)$:

3°. *Функция* $H(p_1, p_2, \ldots, p_k)$ *удовлетворяет соотношению*

$$H(p_1, p_2, \ldots, p_k) = H(p_1+p_2, p_3, \ldots, p_k) + (p_1+p_2)H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right). \quad (1)$$

Это соотношение означает, что неопределенность $H(\beta)$ опыта β с таблицей вероятностей

исходы опыта	B	A_3	...	A_k
вероятности	p_1+p_2	p_3	...	p_k

получаемого отождествлением двух первых исходов опыта α, меньше неопределенности $H(\alpha)$ этого последнего опыта на умноженную на $p_1 + p_2$ меру неопределенности опыта β', состоящего в выяснении того, какой именно из первых двух исходов опыта α имел место, если известно, что осуществился один из именно этих двух исходов.

Можно доказать, что условия 1°, 2° и 3° уже полностью определяют вид функции $H(p_1, p_2, \ldots, p_k)$: *единственная функция, которая удовлетворяет всем этим условиям, имеет вид* [1])

$$H(p_1, p_2, \ldots, p_k) = c(-p_1 \log p_1 - p_2 \log p_2 - \ldots - p_k \log p_k). \quad (*)$$

Однако доказательство этого факта не очень просто (впервые оно было получено Д. К. Фаддеевым [48]). В дальнейшем было также показано, что условие 2° на самом деле можно даже еще значительно ослабить (например, его можно заменить условием 2°а: *функция* $H(p, 1-p)$ *непрерывна в точке* $p = 0$ (т. е. $H(p, 1-p) \to H(0, 1)$ при $p \to 0$), или условием 2°б: *функция* $H(p, 1-p)$ *не меняет знака и ограничена на интервале* $0 \leqslant p \leqslant 1$ — формула (*) при этом все равно будет однозначно вытекать из условий 1° и 3° (некоторые другие допустимые варианты ослабления условия 2° и ссылки на относящуюся сюда довольно большую литературу могут быть найдены, например, в статье З. Дароци [49]). Но мы не будем здесь гнаться за наибольшей общностью, а, следуя Шеннону [1], не только будем считать справедливыми все три условия 1° — 3°, но и дополним их еще одним условием, справедливость которого из них в действительности вытекает, но введение которого заметно упрощает все рассуждения.

В дальнейшем значительную роль будет играть функция $H(1/k, 1/k, \ldots, 1/k)$ — мера неопределенности опыта α_0, имеющего k равновероятных исходов. Очевидно, что в силу равновероятности всех исходов опыта α_0 степень его неопределенности $H(\alpha_0)$ зависит лишь от числа k исходов, т. е. является функцией одного аргумента k: $H(1/k, 1/k, \ldots, 1/k) = f(k)$. Ясно также, что степень неопределенности опыта α_0 должна быть тем больше, чем больше число k его исходов. Таким образом, можно утверждать, что

4°. *Функция* $H(1/k, 1/k, \ldots, 1/k) = f(k)$ *растет с увеличением числа* k.

Покажем теперь, что функция $H(p_1, p_2, \ldots, p_k)$, удовлетворяющая условиям 1° — 4°, обязательно имеет вид (*) (где c — какое-то положительное число). Для этого нам придется несколько обобщить равенство (1), выполнение которого составляет

---

[1]) Если требовать положительности коэффициента c, то придется еще оговорить, что функция $H(p_1, p_2, \ldots, p_k)$ должна быть неотрицательна (разумеется, достаточно включить в число основных условий требование неотрицательности, например, одной величины $H\left(\frac{1}{2}, \frac{1}{2}\right)$). Заметим еще, что если не фиксировать заранее основания системы логарифмов, то в формуле (*) можно отбросить множитель c (ибо $c \log_a p = \log_b p$, где $b = a^{1/c}$).

содержание условия 3°. Покажем прежде всего, что

$$H(p_1, \ldots, p_k) = H(p_1 + \ldots + p_i, p_{i+1}, \ldots, p_k) +$$

$$+ (p_1 + \ldots + p_i) \times$$

$$\times H\left(\frac{p_1}{p_1 + \ldots + p_i}, \frac{p_2}{p_1 + \ldots + p_i}, \ldots, \frac{p_i}{p_1 + \ldots + p_i}\right), \quad i < k$$

(смысл этого равенства, очевидно, аналогичен смыслу первоначального соотношения (1) с той лишь разницей, что здесь речь идет об объединении i исходов $A_1, A_2, \ldots, A_i$ опыта α в один исход B). При $i = 2$ это равенство совпадает с (1) и, следовательно, справедливо в силу условия 3°. Предположим теперь, что справедливость его для некоторого значения i уже доказана; в таком случае, используя также справедливость его для $i = 2$, имеем

$$H(p_1, p_2, \ldots, p_k) = H(p_1 + p_2 + \ldots + p_i, p_{i+1}, \ldots, p_k) +$$

$$+ (p_1 + p_2 + \ldots + p_i) H\left(\frac{p_1}{p_1 + \ldots + p_i}, \ldots, \frac{p_i}{p_1 + \ldots + p_i}\right) =$$

$$= \left\{H(p_1 + p_2 + \ldots + p_i + p_{i+1}, p_{i+2}, \ldots, p_k) + \right.$$

$$\left. + (p_1 + \ldots + p_i + p_{i+1}) H\left(\frac{p_1 + \ldots + p_i}{p_1 + \ldots + p_{i+1}}, \frac{p_{i+1}}{p_1 + \ldots + p_{i+1}}\right)\right\} +$$

$$+ (p_1 + \ldots + p_i) H\left(\frac{p_1}{p_1 + \ldots + p_i}, \frac{p_2}{p_1 + \ldots + p_i}, \ldots, \frac{p_i}{p_1 + \ldots + p_{i+1}}\right).$$

С другой стороны, поскольку для значения i наше равенство считается доказанным, то

$$H\left(\frac{p_1}{p_1 + \ldots + p_{i+1}}, \ldots, \frac{p_i}{p_1 + \ldots + p_{i+1}}, \frac{p_{i+1}}{p_1 + \ldots + p_{i+1}}\right) =$$

$$= H\left(\frac{p_1 + \ldots + p_i}{p_1 + \ldots + p_{i+1}}, \frac{p_{i+1}}{p_1 + \ldots + p_{i+1}}\right) + \frac{p_1 + \ldots + p_i}{p_1 + \ldots + p_{i+1}} \times$$

$$\times H\left(\frac{p_1}{p_1 + \ldots + p_i}, \frac{p_2}{p_1 + \ldots + p_i}, \ldots, \frac{p_i}{p_1 + \ldots + p_i}\right).$$

Отсюда сразу следует справедливость доказываемого равенства для значения $i + 1$:

$$H(p_1, p_2, \ldots, p_k) = H(p_1 + p_2 + \ldots + p_{i+1}, p_{i+2}, \ldots, p_k) +$$

$$+ (p_1 + p_2 + \ldots + p_{i+1}) H\left(\frac{p_1}{p_1 + \ldots + p_{i+1}}, \ldots, \frac{p_{i+1}}{p_1 + \ldots + p_{i+1}}\right).$$

В силу принципа математической индукции мы можем теперь быть уверены в том, что требуемое равенство выполняется при л ю б о м i.

Так как функция $H(p_1, p_2, \ldots, p_k)$ не зависит от порядка своих аргументов $p_1, p_2, \ldots, p_k$ (условие 1°), то из доказанного также следует, что

$$H(p_1, p_2, \ldots, p_{i-1}, p_i, p_{i+1}, \ldots, p_j, p_{j+1}, \ldots, p_k) =$$
$$= H(p_1, p_2, \ldots, p_{i-1}, p_i + p_{i+1} + \ldots + p_j, p_{j+1}, \ldots, p_k) +$$
$$+ (p_i + p_{i+1} + \ldots + p_j) \times$$
$$\times H\left(\frac{p_i}{p_i + \ldots + p_j}, \frac{p_{i+1}}{p_i + \ldots + p_j}, \ldots, \frac{p_j}{p_i + \ldots + p_j}\right),$$
$$1 \leqslant i < j \leqslant k,$$

и вообще

$$H(p_1, \ldots, p_{i_1}, p_{i_1+1}, \ldots, p_{i_2}, p_{i_2+1}, \ldots, p_{i_3}, \ldots, p_{i_s+1}, \ldots, p_k) =$$
$$= H(p_1 + \ldots + p_{i_1}, p_{i_1+1} + \ldots + p_{i_2}, \ldots, p_{i_s+1} + \ldots + p_k) +$$
$$+ (p_1 + \ldots + p_{i_1}) H\left(\frac{p_1}{p_1 + \ldots + p_{i_1}}, \ldots, \frac{p_{i_1}}{p_1 + \ldots + p_{i_1}}\right) +$$
$$+ (p_{i_1+1} + \ldots + p_{i_2}) H\left(\frac{p_{i_1+1}}{p_{i_1+1} + \ldots + p_{i_2}}, \ldots, \frac{p_{i_2}}{p_{i_1+1} + \ldots + p_{i_2}}\right) +$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$+ (p_{i_s+1} + \ldots + p_k) H\left(\frac{p_{i_s+1}}{p_{i_s+1} + \ldots + p_k}, \ldots, \frac{p_k}{p_{i_s+1} + \ldots + p_k}\right),$$
$$i \leqslant i_1 < i_2 < i_3 < \ldots < i_s < k. \qquad (2)$$

Это довольно сложное по форме равенство выражает в весьма общей форме правило сложения энтропий из § 2 [1]).

Обобщение (2) соотношения (1) нам будет нужно не само по себе, а лишь в применении к функции $f(k)$. Предположим, что $k = lm$, где l и m — какие-то целые числа, и что $k = lm$ вероятностей $p_1, p_2, \ldots, p_k$, фигурирующих в формуле (2), все равны между собой (и следовательно, равны $\frac{1}{lm}$) — в таком случае левая часть

---

[1]) Нетрудно убедиться, что если $i_1 = i$, $i_2 = 2i$, $i_3 = 3i$, $\ldots$ $\ldots, k = (s+1)i$ и величины $p_1, p_2, \ldots, p_{i_1}; p_{i_1+1}, p_{i_1+2}, \ldots, p_{i_2}; \ldots$ суть вероятности исходов $A_1B_1, A_1B_2, \ldots, A_1B_i; A_2B_1, A_2B_2, \ldots, A_2B_i; \ldots$ сложного опыта $\alpha\beta$ (так что суммы $p_1 + p_2 + \ldots + p_{i_1}$, $p_{i_1+1} + p_{i_1+2} + \ldots + p_{i_2}, \ldots$ будут равны вероятностям исходов $A_1, A_2, \ldots$ опыта α), то равенство (2) перейдет в правило сложения энтропий.

этого равенства будет равна $f(lm)$. Далее предположим, что группы $(p_1, \ldots, p_{i_1})$, $(p_{i_1+1}, \ldots, p_{i_2})$, $\ldots$, $(p_{i_{s+1}}, \ldots, p_k)$, фигурирующие в том же равенстве (2), состоят из l чисел каждая; в таком случае число таких групп будет равно m. При этом мы будем иметь

$$p_1 + \ldots + p_{i_1} = p_{i_1+1} + \ldots + p_{i_2} = \ldots = p_{i_s+1} + \ldots + p_k = l \cdot \frac{1}{lm} = \frac{1}{m},$$

и, следовательно, первая строка в правой части равенства (2) обратится в $H\left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right) = f(m)$. Что же касается до остальных членов правой части равенства (2), то число этих членов будет равно m и все они будут равны

$$(p_1 + \ldots + p_{i_1}) H\left(\frac{p_1}{p_1 + \ldots + p_{i_1}}, \ldots, \frac{p_{i_1}}{p_1 + \ldots + p_{i_1}}\right) = \frac{1}{m} H\left(\frac{1/ml}{1/m}, \ldots, \frac{1/ml}{1/m}\right) = \frac{1}{m} H\left(\frac{1}{l}, \ldots, \frac{1}{l}\right) = \frac{1}{m} f(l).$$

Таким образом, в рассматриваемом случае равенство (2) примет следующий простой вид:

$$f(lm) = f(m) + m \cdot \frac{1}{m} f(l) = f(m) + f(l). \qquad (2а)$$

Из (2а), в частности, следует

$$f(k^2) = f(k \cdot k) = f(k) + f(k) = 2f(k),$$

$$f(k^3) = f(k^2 \cdot k) = f(k^2) + f(k) = 3f(k),$$

$$f(k^4) = f(k^3 \cdot k) = 4f(k)$$

и вообще

$$f(k^n) = nf(k). \qquad (2б)$$

Мы знаем, что соотношение (2а) выполняется для функции $f(k) = c \log k$. Нетрудно показать также, что функция $c \log k$ является единственной функцией, удовлетворяющей соотношению (2а) и условию 4°. В самом деле, пусть k и l — два произвольных целых положительных числа. Выберем еще какое-либо большое целое число N и найдем такое число n, что

$$l^n \leqslant k^N < l^{n+1}.$$

Согласно условию 4°,

$$f(l^n) \leqslant f(k^N) < f(l^{n+1})$$

или, в силу (2б),

$$nf(l) \leqslant Nf(k) < (n+1)f(l);$$

отсюда вытекает, что

$$\frac{n}{N} \leqslant \frac{f(k)}{f(l)} \leqslant \frac{n+1}{N}.$$

Заметим теперь, что из равенства $l^n \leqslant k^N < l^{n+1}$ следует

$$n \log l \leqslant N \log k < (n+1) \log l$$

или

$$\frac{n}{N} \leqslant \frac{\log k}{\log l} < \frac{n+1}{N}.$$

Таким образом, отношения $\frac{f(k)}{f(l)}$ и $\frac{\log k}{\log l}$ заключаются в одних и тех же пределах, и, следовательно, должны быть близки между собой:

$$\left| \frac{f(k)}{f(l)} - \frac{\log k}{\log l} \right| < \frac{1}{N}.$$

А так как последнее неравенство имеет место при л ю б о м значении N, то

$$\frac{f(k)}{f(l)} = \frac{\log k}{\log l}$$

или

$$\frac{f(k)}{\log k} = \frac{f(l)}{\log l}.$$

Это соотношение имеет место для к а ж д ы х двух чисел k и l; следовательно,

$$\frac{f(k)}{\log k} = \frac{f(l)}{\log l} = c,$$

где c не зависит от k и l, и, значит,

$$f(k) = c \log k.$$

А так как функция $f(k)$ — возрастающая, то $c > 0$.

Пусть теперь $p_1, p_2, \ldots, p_k$ — произвольные дроби:

$$p_1 = \frac{q_1}{p}, \; p_2 = \frac{q_2}{p}, \; \ldots, \; p_k = \frac{q_k}{p}$$

(p — общий знаменатель всех этих дробей), меньшие единицы и такие, что $p_1 + p_2 + \ldots + p_k = 1$. Согласно формуле (2)

(стр. 133) имеем

$$f(p) = H\Big(\underbrace{\frac{1}{p}, \frac{1}{p}, \ldots, \frac{1}{p}}_{p \text{ раз}}\Big) =$$

$$= H\Big(\underbrace{\frac{1}{p}, \ldots, \frac{1}{p}}_{q_1 \text{ раз}}, \underbrace{\frac{1}{p}, \ldots, \frac{1}{p}}_{q_2 \text{ раз}}, \ldots, \underbrace{\frac{1}{p}, \ldots, \frac{1}{p}}_{q_k \text{ раз}}\Big) =$$

$$= H\left(\frac{q_1}{p}, \frac{q_2}{p}, \ldots, \frac{q_k}{p}\right) + \frac{q_1}{p} H\Big(\underbrace{\frac{1}{q_1}, \ldots, \frac{1}{q_1}}_{q_1 \text{ раз}}\Big) +$$

$$+ \frac{q_2}{p} H\Big(\underbrace{\frac{1}{q_2}, \ldots, \frac{1}{q_2}}_{q_2 \text{ раз}}\Big) + \ldots + \frac{q_k}{p} H\Big(\underbrace{\frac{1}{q_k}, \ldots, \frac{1}{q_k}}_{q_k \text{ раз}}\Big) =$$

$$= H(p_1, p_2, \ldots, p_k) + p_1 f(q_1) + p_2 f(q_2) + \ldots + p_k f(q_k).$$

Отсюда следует, что

$$H(p_1, p_2, \ldots, p_k) = f(p) - p_1 f(q_1) - p_2 f(q_2) - \ldots - p_k f(q_k) =$$
$$= (p_1 + p_2 + \ldots + p_k) f(p) - p_1 f(q_1) - p_2 f(q_2) - \ldots - p_k f(q_k) =$$
$$= p_1 (f(p) - f(q_1)) + p_2 (f(p) - f(q_2)) + \ldots + p_k (f(p) - f(q_k)).$$

А так как

$$f(p) - f(q_1) = c \log p - c \log q_1 = -c \log \frac{q_1}{p} = -c \log p_1,$$
$$f(p) - f(q_2) = -c \log p_2, \ldots, f(p) - f(q_k) = -c \log p_k,$$

то окончательно получаем

$$H(p_1, p_2, \ldots, p_k) = c(-p_1 \log p_1 - p_2 \log p_2 - \ldots - p_k \log p_k).$$

Последнее равенство доказано пока только для р а ц и о н а л ь н ы х значений $p_1, p_2, \ldots, p_k$. Но в силу непрерывности функции $H(p_1, p_2, \ldots, p_k)$ отсюда следует, что оно верно для л ю б ы х $p_1, p_2, \ldots, p_k$. Этим и завершается рассуждение.

Глава III

РЕШЕНИЕ НЕКОТОРЫХ ЛОГИЧЕСКИХ ЗАДАЧ С ПОМОЩЬЮ ПОДСЧЕТА ИНФОРМАЦИИ

§ 1. Простейшие примеры

Для иллюстрации применимости понятий и предложений, введенных в гл. II, мы разберем здесь несколько занимательных задач, типа собранных в книге Б. А. Кордемского «Математическая смекалка» [50] [1]). При этом в §§ 1 и 2, посвященных конкретным примерам таких задач, мы часто будем пользоваться «наглядными представлениями» об информации, отложив более аккуратное обсуждение приведенных здесь рассуждений до заключительного § 3.

Начнем с довольно популярных в школьных математических кружках логических задач о «городах лжецов и честных людей».

Задача 23. *Пусть известно, что жители некоторого города А всегда говорят правду, а жители соседнего города Б всегда обманывают. Наблюдатель* Н. *знает, что он находится в одном из этих двух городов, но не знает в каком именно. Путем опроса встречного ему требуется определить, в каком городе он находится, или в каком городе живет его собеседник* (жители A могут заходить в $Б$ и наоборот), *или то и другое вместе. Спрашивается, каково наименьшее число вопросов, которые должен задать* Н. (на все вопросы Н. встречный отвечает лишь «*да*» или «*нет*»)?

Пусть Н. надо определить, в каком городе он находится. Здесь опыт β, результат которого нас интересует, может иметь два исхода (этот опыт состоит в выяснении того, в каком из двух городов A и $Б$ находится наблюдатель Н.). Если считать, что заранее Н. не имеет никакой информации о том, в какой из двух городов он попал, то эти исходы следует считать равновозможными; следовательно, энтропия $H(\beta)$ опыта β (т. е. «полное» количество информации, которое содержится в исходе этого опыта)

[1]) В дальнейшем мы будем цитировать эту книгу как «М.с.»

равна одному биту. Далее, опыт α, состоящий в том, что Н. задает встречному один вопрос, также может иметь два исхода (собеседник может ответить утвердительно или отрицательно); поэтому энтропия $H(\alpha)$ этого опыта (равная «полному» количеству информации, содержащейся в ответе на поставленный вопрос) самое большее равна одному биту. В задаче спрашивается, можно ли так поставить опыт α, чтобы информация $I(\alpha, \beta)$, содержащаяся в опыте α относительно опыта β, равнялась энтропии $H(\beta) = 1$ опыта β, т. е. чтобы исход α полностью определял исход β. Так как единственная связь между информацией $I(\alpha, \beta)$ и энтропией $H(\alpha)$ заключается в том, что

$$I(\alpha, \beta) \leqslant H(\alpha) \quad (\text{ибо } I(\alpha, \beta) = H(\alpha) - H_\beta(\alpha)),$$

а $H(\alpha)$ может равняться 1, то, вообще говоря, можно надеяться, что при удачном выборе опыта α будет иметь место равенство

$$I(\alpha, \beta) = H(\beta).$$

Для этого необходимо только, чтобы вопрос α был таким, чтобы утвердительный и отрицательный ответ на него были равновероятны [1]) (только в этом случае будут иметь место равенства $H(\alpha) = 1 = H(\beta)$), и чтобы исход опыта β определял исход α (только при этом условии имеет место равенство $I(\alpha, \beta) = H(\alpha)$ или $H_\beta(\alpha) = 0$, указывающее, что вопрос α «прямо направлен» к выяснению исхода β и ответ на этот вопрос не содержит никакой «посторонней» информации). Всем этим условиям удовлетворяет вопрос «*Живете ли Вы в этом городе?*», полностью решающий задачу (положительный ответ на этот вопрос может быть дан только в городе A, а отрицательный — только в $Б$).

Еще проще видеть, что Н. может с помощью одного вопроса установить, в каком городе живет его собеседник: для этого достаточно задать любой вопрос, ответ на который Н. знает заранее (например, «*Нахожусь ли я в городе?*» или «*Равно ли 2·2 четырем?*»).

Если же Н. должен узнать, и в каком городе он находится и в каком городе живет его собеседник, то ему требуется определить исход сложного опыта $\beta_1\beta_2$, где опыт β_1

[1]) При условии равновероятности того, что Н. находится в A и в $Б$ и что его собеседник живет в A и в $Б$.

состоит в выяснении того, где находится Н., а опыт β_2 — в выяснении места жительства его собеседника. Энтропия $H(\beta_1\beta_2)$ этого опыта больше энтропии $H(\beta_1)$ опыта β_1: $H(\beta_1\beta_2) = H(\beta_1) + H_{\beta_1}(\beta_2)$ (см. § 2 гл. II). Иначе говоря, в этом случае требуется получить информацию большую, чем 1 бит (напомним, что $H(\beta_1) = 1$). Так как энтропия $H(\alpha)$ опыта α с двумя исходами, состоящего в постановке вопроса, не может превосходить 1, то один опыт α не дает возможности получить информацию, равную $H(\beta_1\beta_2)$, т. е. не позволяет полностью определить исход опыта $\beta_1\beta_2$ (за исключением того мало интересного случая, когда условная энтропия $H_{\beta_1}(\beta_2)$ равна 0, т. е. когда исход β_1 определяет исход β_2 — так будет обстоять дело в том случае, когда жители A не могут попасть в $Б$, и наоборот). Таким образом, оценки количества информации дают нам строгое доказательство того, что один вопрос (как бы он ни был поставлен!) не позволяет выяснить сразу и то, в каком городе находится Н., и то, в каком городе живет его собеседник. Если же Н. задаст два вопроса (т. е. произведет сложный опыт $\alpha_1\alpha_2$, имеющий 4 возможных исхода), то он, разумеется, может выяснить исход опыта $\beta_1\beta_2$ (с помощью вопроса α_1 можно определить исход β_1, а затем с помощью вопроса α_2 — исход β_2).

Усложним теперь несколько условия задачи 23.

З а д а ч а 24. *Пусть имеются три города A, $Б$ и $В$, причем жители A во всех случаях говорят правду, жители $Б$ — только неправду, а жители $В$ через раз отвечают на вопросы верно и неверно. Наблюдатель* Н. *хочет выяснить, в каком городе он находится и в каком городе живет встреченный им человек. Сколько вопросов ему потребуется задать этому встречному, если на все вопросы его собеседник отвечает лишь «да» или «нет»?*

Здесь требуется определить, какой из д е в я т и возможных исходов имеет интересующий нас опыт β (Н. может находиться в одном из трех городов A, $Б$ и $В$ и, независимо от этого, его собеседник может проживать в одном из этих же трех городов). Если полагать, что заранее у Н. нет никаких сведений, относящихся к опыту β, то все эти девять исходов можно считать равновероятными и энтропия $H(\beta)$ опыта β (а, следовательно, и количество информации, получаемой при выяснении исхода этого опыта) будет равна $\log 9$. Пусть сложный опыт $A_k = \alpha_1\alpha_2 \ldots \alpha_k$ состоит

в том, что Н задает k вопросов. Так как на каждый вопрос он может получить утвердительный или отрицательный ответ, то энтропия каждого из опытов $\alpha_1, \alpha_2, \ldots, \alpha_k$ не превосходит одного бита. С другой стороны,

$$H(\alpha_1\alpha_2) = H(\alpha_1) + H_{\alpha_1}(\alpha_2) \leqslant H(\alpha_1) + H(\alpha_2)$$

(ибо $H_{\alpha_1}(\alpha_2) \leqslant H(\alpha_2)$) и аналогично

$$H(A_k) = H(\alpha_1\alpha_2\ldots\alpha_k) \leqslant H(\alpha_1) + H(\alpha_2) + \ldots + H(\alpha_k) \leqslant k$$

(строгое доказательство этого неравенства легко получить, воспользовавшись методом математической индукции). Иначе это можно выразить так: если ответ на каждый вопрос дает нам информацию, не превосходящую одного бита, то, задав k вопросов, мы можем получить информацию не бóльшую, чем k бит. Поэтому, если $k = 3$, то полученная информация будет меньше чем $\log 9$ (она может быть, самое большее, равна $3 = \log 8 < \log 9$) и, значит, три вопроса не могут обеспечить выяснения и местонахождения Н. и места проживания его собеседника. Четыре же удачно поставленных вопроса, может быть, позволят выяснить все, что требуется (ибо можно лишь утверждать, что $H(A_4) \leqslant 4 = \log 16$). И действительно, легко видеть, что следующие 4 вопроса:

1) Нахожусь ли я в одном из городов A и $Б$?
2) Нахожусь ли я в городе B?
3) Живете ли Вы в городе B?
4) Нахожусь ли я в городе A?

обеспечивают выяснение всего, что интересует Н.

В самом деле, утвердительные или отрицательные ответы на о б а вопроса 1) и 2) сразу указывают, что собеседник Н. живет в B. Пусть, например, ответы на оба эти вопроса являются утвердительными (аналогично разбирается и случай, когда оба ответа — отрицательные). В этом случае отрицательный (неправильный) ответ на вопрос 3) означает, что ответ на вопрос 2) был верен, и четвертый вопрос уже не нужен; положительный (правильный) ответ на вопрос 3) означает, что верен ответ на вопрос 1), и для того, чтобы выяснить, в каком городе находится Н., требуется задать вопрос 4) (ответ на который будет заведомо неверен). Утвердительный же ответ на вопрос 1) и отрицательный — на вопрос 2) или обратная ситуация указывают, что собеседник Н. живет в A или $Б$. При этом отрицательный (правильный) ответ на вопрос 3) означает, что отвечающий живет в A и вопрос 4) оказывается нужным лишь в том случае, если ответ на вопрос 2) был отрицатель-

ным; положительный (неправильный) ответ на вопрос 3) означает, что собеседник Н. живет в E и вопрос 4) оказывается нужным лишь в том случае, если ответ на вопрос 2) был положительным.

Вот еще одна задача подобного рода (см. «М. с.», задача 283):

З а д а ч а 25. *Сколько вопросов надо задать, чтобы отгадать задуманное собеседником целое положительное число, не превосходящее* 10 (*или* 100, *или* 000, *или произвольного целого положительного числа* n), *если спрашиваемый на все вопросы отвечает лишь «да» или «нет»?*

Пусть известно, что задуманное число не превосходит 10. В таком случае опыт β, состоящий в выяснении этого числа, может иметь 10 различных исходов. До ответа на первый поставленный вопрос все эти исходы можно считать равновероятными, так что энтропия $H(\beta)$ опыта β (т. е. требуемая информация) равна $\log 10 \approx 3{,}32$ бита. Рассмотрим сложный опыт $A_k = \alpha_1\alpha_2 \ldots \alpha_k$, заключающийся в том, что спрашивающий задает k вопросов. Энтропия опыта α_1, заключающегося в постановке одного вопроса, не превосходит одного бита, так как α_1 может иметь два исхода (положительный и отрицательный ответы на вопрос); энтропия опыта A_k не превосходит k бит (см. предыдущую стр.). С другой стороны, информация $I(A_k, \beta)$ относительно опыта β, содержащаяся в опыте A_k, не может превосходить полной информации, содержащейся в исходе последнего опыта — энтропии $H(A_k)$. Для того чтобы исход опыта A_k полностью определял исход β, необходимо, чтобы имело место равенство $I(A_k, \beta) = H(\beta)$. Отсюда заключаем, что в этом случае

$$\log 10 = H(\beta) = I(A_k, \beta) \leqslant H(A_k) \leqslant k$$

т. е.

$$k \geqslant \log 10 \approx 3{,}32,$$

или, так как k — целое число,

$$k \geqslant 4.$$

Покажем теперь, что с помощью четырех вопросов действительно можно полностью определить исход β, т. е. обнаружить загаданное число x. Легко понять, как для этого следует поступать. Прежде всего естественно добиваться, чтобы информация, содержащаяся в ответе на

первый вопрос (т. е. энтропия $H(\alpha_1)$), была возможно большей, т. е. чтобы она действительно равнялась одному биту; для этого надо, чтобы оба исхода нашего опыта α_1 были равновероятны. Далее следует потребовать, чтобы информация $I(\alpha_1, \beta)$ относительно β, заключенная в α_1, равнялась энтропии $H(\alpha_1)$ опыта α_1, а не была бы меньше этой величины. Для этого надо, чтобы ответ на первый вопрос не содержал «посторонней» информации, т. е. чтобы условная энтропия $H_\beta(\alpha_1)$ равнялась нулю (другими словами: чтобы исход опыта β полностью определял исход α_1). Эти соображения ясно указывают, как следует поставить первый вопрос. Разобьем множество всех возможных значений x (т. е. множество целых положительных чисел от 1 до 10) на две равные по численности части (так как исходы опыта α_1 должны быть равновероятны) и спросим, относится ли x к одной или к другой из них; так, например, можно спросить, будет ли x больше 5. Очевидно, что в этом случае

$$I(\alpha_1, \beta) = H(\beta) - H_{\alpha_1}(\beta) = 1,$$

т. е.

$$H_{\alpha_1}(\beta) = p(A_1) H_{A_1}(\beta) + p(A_2) H_{A_2}(\beta) = H(\beta) - 1$$

(A_1 и A_2 — исходы опыта α_1; $p(A_1) = p(A_2) = \frac{1}{2}$); кроме того,

$$H_{A_1}(\beta) = H_{A_2}(\beta) = H(\beta) - 1,$$

так что при любом исходе опыта α_1 энтропия интересующего нас опыта β уменьшится на 1 бит. Далее следует точно таким же образом разбить новое множество допустимых значений x на две возможно более близкие по численности части и выяснить, к какой из них x принадлежит (если обнаружилось, что x больше 5, то можно спросить, больше ли это число, чем 7; если же x не превосходит 5, то можно спросить, больше ли x, чем 3) и т. д. Если каждый раз разбивать множество допустимых значений x на возможно более близкие по численности части, то мы, наверное, определим x с помощью четырех вопросов [1]).

[1]) Разумеется, после того как уже выяснено, что число x имеет одно из m значений, где m нечетно (например, $m = 5$), мы не можем добиться строгой равновероятности исходов последу-

Совершенно так же показывается, что наименьшее число k вопросов, позволяющее определить загаданное число x, которое может иметь 100 или 1000 значений, определяется неравенствами $k \geqslant \log 100 \approx 6{,}64$ и, соответственно, $k \geqslant \log 1000 \approx 9{,}97$; так как во всех случаях k — целое число, то отсюда получаем

$$k \geqslant 7 \text{ и } k \geqslant 10.$$

Вообще *наименьшее число k вопросов, позволяющее найти загаданное число x, имеющее одно из n допустимых значений, определяется неравенствами*

$$k - 1 < \log n \leqslant k \quad (\text{или } 2^{k-1} < n \leqslant 2^k). \qquad (1)$$

Заметим еще, что независимо от значения n

$$k \geqslant \log n;$$

при этом $k = \log n$ только в том случае, когда число n является целой степенью числа 2 и, следовательно, $\log n$ есть целое число. Однако при весьма больших n разница между числами k и $\log n$ оказывается очень малой по сравнению с самими этими числами (ибо при больших n и величина $\log n$ будет большой, а разность $k - \log n$ всегда не превосходит единицы). Таким образом, можно считать, что при больших n отношение $\log n$ энтропии рассматриваемого опыта β к (равной 1 биту) информации относительно β, содержащейся в опыте α, состоящем в выяснении ответа на один вопрос, весьма точно указывает число k опытов, требующихся для того, чтобы определить исход β.

Задача 25 на первый взгляд представляется столь же искусственной, как и две ей предшествующие; впоследствии, однако, мы увидим, что она имеет серьезные

ющего опыта α_{i+1}, поскольку m возможных значений x здесь нельзя разбить на равные по численности части; следовательно, энтропия $H(\alpha_{i+1})$ опыта α_{i+1} будет м е н ь ш е 1. Это означает, что наш опрос не будет наиболее выгоден с точки зрения полученной информации, т. е. что с помощью того же числа вопросов можно найти загаданное число и тогда, когда множество его возможных значений имеет большую численность (так, с помощью 4 вопросов можно обнаружить загаданное число, имеющее не одно из 10, а даже одно из $2^4 = 16$ возможных значений).

технические приложения [1]). Более подробное обсуждение приведенного здесь решения этой задачи (включающее также и более общую формулировку ее условия) мы отложим до § 3 этой главы.

Очень близка к задаче 25 и следующая

З а д а ч а 26. *Некто задумал два (различных) числа, не превосходящих* 100. *Сколько надо задать ему вопросов для того, чтобы определить эти числа, если на каждый вопрос спрашиваемый отвечает лишь «да» или «нет»?*

В этом случае опыт β, исход которого нам требуется определить, может иметь $C^2_{100} = 4950$ различных исходов; если, как всегда, считать все эти исходы равновероятными, то энтропия $H(\beta)$ опыта β (т. е. информация, которую мы получим, определив исход β) будет равна $\log 4950$. А так как информация, которую может дать ответ на один вопрос, не превосходит одного бита (ибо опыт α, состоящий в постановке одного вопроса, может иметь два исхода: *«да»* и *«нет»*), то наименьшее число вопросов, с помощью которых всегда можно определить исход β, никак не может быть меньше чем

$$\log 4950 \approx 12{,}27$$

(ср. с решением задачи 25). Таким образом, если мы зададим меньше тринадцати вопросов, то наверное может случиться, что оба загаданных числа нам определить не удастся.

Нетрудно видеть также, что 13 удачно поставленных вопросов всегда позволяют найти загаданные числа. Для того чтобы достичь этого, надо добиваться, чтобы информация $I(\alpha, \beta)$ относительно исхода опыта β, содержащаяся в исходе опыта α — ответе на один вопрос (точнее — на каждый из задаваемых вопросов), была как можно ближе к одному биту. Отсюда ясно, что вопросы надо ставить так, чтобы оба ответа *«да»* и *«нет»* имели возможно более близкие вероятности. А для этого достаточно разбить сначала 4950 исходов β на две возможно более близкие по численности части (так, чтобы каждая часть содержала 2475 исходов) и выяснить, к какой из этих частей относится тот

---

[1]) Следует, впрочем, указать, что на самом деле и за шуточными формулировками задач 23—24 скрывается достаточно серьезное содержание (ср. стр. 163—165).

исход β, который имеет место (т. е. прежде всего следует спросить, принадлежат ли или не принадлежат загаданные два числа к первой группе, содержащей 2475 пар чисел). Вслед за этим надо точно так же разбить на две по возможности близкие по численности части ту группу исходов β, к которой оказался принадлежащим интересующий нас исход, и выяснить, к какой из этих двух меньших частей он относится, и т. д. Ясно, что при этом мы всегда определим загаданную пару чисел с помощью не более чем тринадцати вопросов.

Заметим еще, что отличие задачи 26 от задачи 25 можно считать чисто словесным. Ясно, что в решении задачи 25 играет роль только о б щ е е к о л и ч е с т в о n тех чисел, одно из которых загадано. При этом, разумеется, всегда можно считать, что эти n чисел являются номерами каких угодно объектов — например, номерами n каких-то предметов, или n пар чисел, или n каких-то других групп чисел и т. д.— на решение задачи это никак не повлияет. Но если считать, что число n в задаче 25 равно 4950 и что соответствующие 4950 чисел — это номера всевозможных п а р чисел, каждое из которых не превосходит 100, то мы придем к задаче 26.

Точно так же показывается, что наименьшее число вопросов, с помощью которых можно определить загаданные m чисел, не превосходящих n, равно наименьшему целому числу k, такому, что $k \geqslant \log C_n^m$. Если же, например, мы знаем, что загадано или одно число, не превосходящее n, или ни одного числа, то для того, чтобы выяснить, было ли число загадано и если да, то какое именно, требуется не меньше чем $\log(n+1)$ и не больше чем $\log(n+1)+1$ вопросов: ведь в этом случае число возможных исходов соответствующего опыта β равно $n+1$ (единица в этой сумме соответствует случаю, когда никакое число не было загадано). Наконец, если предположить, что было загадано н е б о л е е m чисел, где $m \leqslant \frac{n}{2}$, каждое из которых не превосходит n, то число вопросов, нужных для выяснения того, сколько чисел было загадано и какие именно, будет заключено между

$$\log(C_n^m + C_n^{m-1} + \ldots + C_n^1 + 1)$$

и

$$\log(C_n^m + C_n^{m-1} + \ldots + C_n^1 + 1) + 1.$$

В самом деле, рассматриваемый здесь опыт β может иметь $C_n^m + C_n^{m-1} + \ldots + C_n^1 + 1$ разных исходов (поскольку могут оказаться загаданными: или одна из C_n^m групп из m чисел, или одна из C_n^{m-1} групп из $m-1$ чисел, ..., или одно из $C_n^1 = n$ отдельных чисел, или же вообще ни одно из чисел). Перенумеровав эти $N = C_n^m + C_n^{m-1} + \ldots + C_n^1 + 1$ исходов опыта β числами от 1 до N, мы придем к задаче 25 (в которой лишь число n заменено на N). Ниже мы еще воспользуемся этим замечанием.

§ 2. Задачи на определение фальшивых монет с помощью взвешиваний

Этот параграф мы начнем со следующей задачи, весьма близкой к задаче 25.

З а д а ч а 27. *Имеется 25 монет одного достоинства; 24 из них имеют одинаковый вес, а одна — фальшивая — несколько легче остальных. Спрашивается, сколькими взвешиваниями на чашечных весах без гирь можно обнаружить эту фальшивую монету* (ср. «М. с.», задачи 277, 1) и 2)).

Опыт β, результат которого требуется определить, имеет в этом случае 25 возможных исходов (фальшивой может оказаться любая из 25 монет); эти исходы естественно считать равновероятными, так что $H(\beta) = \log 25$. Иначе говоря, определение фальшивой монеты в данном случае связано с получением информации, измеряющейся числом $\log 25$. Опыт α_1, состоящий в одном (каком угодно) взвешивании, может иметь три исхода (может перевесить левая или правая чашка весов и весы могут остаться в равновесии); поэтому $H(\alpha_1) \leqslant \log 3$ и информация $I(\alpha_1, \beta)$, получаемая при проведении такого опыта, не превосходит $\log 3$. Рассмотрим теперь сложный опыт $A_k = \alpha_1\alpha_2 \ldots \alpha_k$, заключающийся в k последовательных взвешиваниях; он дает информацию, не превосходящую $k \log 3$ (ср. выше, стр. 40). Если опыт A_k позволяет полностью определить исход опыта β, то должно быть

$$H(A_k) \geqslant I(A_k, \beta) \geqslant H(\beta) \text{ или } k \log 3 \geqslant \log 25.$$

Отсюда заключаем, что $3^k \geqslant 25$, т. е.

$$k \geqslant \log_3 25 = \frac{\log 25}{\log 3}$$

или, так как k — целое число,

$$k \geqslant 3.$$

Нетрудно показать, что с помощью трех взвешиваний всегда можно определить фальшивую монету. Для того чтобы информация, получаемая при проведении опыта α_1, была возможно большей, надо, чтобы исходы этого опыта имели возможно более близкие вероятности. Предположим, что на каждую чашку весов нами положено по m монет (ясно, что не имеет смысла класть на чашки разное число монет: в этом случае исход соответствующего опыта будет заранее известен и полученная информация будет равна нулю); не положены на весы будут $25-2m$ монет. Так как вероятность того, что фальшивая монета окажется в данной группе из n монет, равна $\frac{n}{25}$ (ибо все исходы опыта β мы считаем равновероятными!), то три исхода опыта α_1 будут иметь вероятности $\frac{m}{25}$, $\frac{m}{25}$ и $\frac{25-2m}{25}$; наиболее близки одна к другой эти вероятности будут в том случае, когда $m = 8$ и $25 - 2m = 9$. Если мы положим на каждую чашку весов по 8 монет, то первое взвешивание (опыт α_1) позволит нам выделить группу в 9 монет (если весы окажутся в равновесии) или в 8 монет (если одна из чашек перетянет), в которой находится фальшивая монета. В обоих случаях при втором взвешивании (опыт α_2) для получения наибольшей информации на обе чашки весов следует положить по 3 монеты из этой группы; при этом сложный опыт $\alpha_1\alpha_2$ позволяет выделить группу в 3 (или в 2) монеты, среди которых находится фальшивая. При третьем взвешивании (опыт α_3) мы положим на обе чашки весов по одной из оставшихся подозрительными монет и легко обнаружим фальшивую.

Точно так же показывается, что *наименьшее число k взвешиваний, позволяющих обнаружить одну фальшивую* (более легкую!) *монету, имеющуюся в группе из n монет, определяется неравенствами*

$$3^{k-1} < n \leqslant 3^k \text{ или } k - 1 < \frac{\log n}{\log 3} \leqslant k. \qquad (2)$$

Если n — большое число, то это число k с большой степенью точности дается отношением $\frac{\log n}{\log 3}$, т. е. отношением

энтропии опыта β, состоящего в определении фальшивой монеты, к наибольшей информации, которую можно получить при одном взвешивании (ср. стр. 143).

В дальнейшем нам будет полезен также аналогичный результат, относящийся к несколько более общей постановке задачи. Прежде всего ясно, что если мы имеем n монет, одна из которых является фальшивой — несколько более т я ж е л о й, чем остальные,— то наименьшее число k взвешиваний на чашечных весах без гирь, позволяющее обнаружить эту фальшивую монету, определяется теми же неравенствами (2): замена более легкой монеты более тяжелой практически не меняет наших рассуждений. Рассмотрим теперь более общий случай, когда наши n *монет разбиты на две группы — группу* A *из* a *монет и группу* B *из* $b = n - a$ *монет, причем известно, что одна из этих* n *монет является фальшивой и что, если эта монета принадлежит к группе* A, *то она легче остальных, а если она принадлежит к группе* B, *то тяжелее остальных*, и покажем, что и здесь *наименьшее число* k *взвешиваний, позволяющих обнаружить фальшивую монету, дается неравенствами* (2) [1]); при $b = 0$ это утверждение переходит в сделанное выше.

В самом деле, так как интересующий нас опыт β, очевидно, может иметь n различных исходов, то $3^k \geqslant n$ — в противном случае опыт $A_k = \alpha_1\alpha_2 \ldots \alpha_k$, состоящий в k-кратном взвешивании, никак н е м о ж е т однозначно определить исход опыта β (ибо в этом случае $I(A_k, \beta) \leqslant H(A_k) \leqslant k \log 3 = \log 3^k < \log n = H(\beta)$; исходы β мы, как и всегда, считаем равновероятными). С другой стороны, при $n \leqslant 3^k$ фальшивую монету всегда м о ж н о выделить k взвешиваниями; это легко показать, воспользовавшись, например, методом математической индукции. В самом деле, если $k = 1$, т. е. $n = 1$, 2 или 3, то наше утверждение почти очевидно (с одним ограничением, указанным в подстрочном примечании на этой странице): при $n = 1$ фальшивая монета известна заранее, а при $n = 2$ (и $a = 2$ или $b = 2$) и при $n = 3$ для ее определения достаточно сравнить вес двух монет из одной группы. Пред-

---

[1]) Это утверждение имеет одно очевидное исключение: если $n = 2$, $a = b = 1$, то фальшивую монету, разумеется, вовсе невозможно выделить.

положим теперь, что мы уже доказали, что при $n \leqslant 3^k$ фальшивую монету всегда можно выделить при помощи не больше чем k взвешиваний и пусть $3^k < n \leqslant 3^{k+1}$. Легко видеть, что при этом всегда можно будет отобрать четное число $2x$ монет из группы A и четное число $2y$ монет из группы B так, чтобы числа x и y удовлетворяли условиям:

$$2x + 2y \leqslant 2 \cdot 3^k, \quad n - (2x + 2y) \leqslant 3^k,$$

т. е.

$$3^k \geqslant x + y \geqslant \frac{n - 3^k}{2}.$$

Поместим теперь на каждую чашку весов по x монет из группы A и по y монет из группы B; не использованными у нас останутся $n_1 = n - 2x - 2y \leqslant 3^k$ монет. Если весы при таком взвешивании (опыт α_1) останутся в равновесии, то значит фальшивая монета находится среди n_1 отложенных монет (т. е. среди $a_1 = a - 2x$ не участвующих в первом взвешивании монет группы A или среди $b_1 = b - 2y$ не использованных монет группы B); если одна из чашек перетянет, то фальшивая монета находится среди x монет группы A, лежащих на более легкой чашке, или среди y монет группы B, лежащих на более тяжелой чашке. Но так как $n_1 \leqslant 3^k$ и $x + y \leqslant 3^k$, то согласно сделанному предположению мы в обоих случаях сможем выделить фальшивую монету, произведя еще не более чем k взвешиваний [1]); следовательно, из наших $n \leqslant 3^{k+1}$ монет одну фальшивую, наверное, можно выделить при помощи не больше чем $k + 1$ взвешиваний. Это рассуждение и завершает доказательство сделанного выше утверждения.

Рассмотрим теперь следующую, несколько более сложную задачу такого же типа, пользующуюся большой популярностью в школьных математических кружках:

З а д а ч а 28. *Имеется 12 монет одного достоинства; 11 из них имеют одинаковый вес, а одна — фальшивая —*

[1]) Если $n > 2$, то случай, когда $x = y = 1$ или $a_1 = b_1 = 1$, теперь уже не представляет исключения: ведь помимо одной сомнительной монеты из группы A и одной — из группы B мы имеем теперь еще некоторое число заведомо не фальшивых («настоящих») монет; сравнив вес одной их них с весом одной из сомнительных монет, мы сможем одним взвешиванием выделить фальшивую монету.

отличается по весу от остальных (причем неизвестно, легче ли она или тяжелее настоящих). Каково наименьшее число взвешиваний на чашечных весах без гирь, которое позволяет обнаружить фальшивую монету и выяснить, легче ли она, чем остальные монеты, или тяжелее? Решить тот же вопрос для случая 13 монет (ср. «М. с.», задача 277(3) или Д. О. Шклярский, Н. Н. Ченцов, И. М. Яглом [51], задача 6а)).

Здесь рассматривается опыт β, имеющий 24 (или 26) возможных исходов (каждая из 12 или из 13 имеющихся монет может оказаться фальшивой, причем она может быть или легче или тяжелее настоящих). Если считать все эти исходы равновероятными, то энтропия $H(\beta)$ опыта β будет равна $\log 24$ или $\log 26$. Таким образом, требуется получить $\log 24$ или, соответственно, $\log 26$ единиц информации. Так как, произведя сложный опыт $A_k = \alpha_1\alpha_2 \dots \alpha_k$, состоящий в k взвешиваниях, мы можем получить информацию, не бóльшую, чем $k \log 3 = \log 3^k$, а $3^3 = 27$, то с первого взгляда кажется правдоподобным, что и в случае 12 и в случае 13 монет трехкратное взвешивание может позволить найти фальшивую монету и выяснить, легче ли она или тяжелее других. На самом деле, однако, в случае 13 монет трех взвешиваний может оказаться недостаточно; этот факт весьма просто доказывается с помощью несколько более тщательного вычисления информации, доставляемой первым взвешиванием.

В самом деле, первое взвешивание может заключаться в том, что на обе чашки весов кладется по 1, по 2, по 3, по 4, по 5, или, наконец, по 6 монет; соответствующие опыты обозначим через $\alpha_1^{(i)}$, где i может быть равно 1, 2, 3, 4, 5 или 6. Если i равно 1, 2, 3 или 4 и весы в результате первого взвешивания остаются в равновесии, то опыт $\alpha_1^{(i)}$ указывает, что фальшивой является одна из $13-2i$ отложенных монет; так как это число не меньше 5, то остаются возможными 10 (или еще больше) исходов и два последующих взвешивания не могут гарантировать выявления фальшивой монеты и выяснения того, легче ли она или тяжелее остальных (ибо $2 \log 3 = \log 9 < \log 10$). Если i равно 5 или 6 и в опыте $\alpha_1^{(i)}$ одна (например, правая) чашка весов перевесила, то опыт $\alpha_1^{(i)}$ указывает, что либо фальшивой и более тяжелой является одна из i «правых»

монет, либо же фальшивая и более легкая — одна из i «левых» монет. Таким образом, и здесь у нас остается еще $i + i = 2i \geqslant 10$ возможных исходов опыта β — и опять двух взвешиваний недостаточно для того, чтобы выяснить, какой из них на самом деле имеет место.

Перейдем теперь к случаю 12 монет. Пусть при первом взвешивании мы положили на обе чашки по i монет (опыт $\alpha_1^{(i)}$). Если при этом чашки весов остались в равновесии (исход P опыта $\alpha_1^{(i)}$; подобные обозначения мы будем употреблять и в дальнейшем), то фальшивой является одна из $12-2i$ отложенных монет, что отвечает $2(12-2i)$ равновероятным исходам рассматриваемого опыта β (из общего числа 24 исходов). Если перевесила правая чашка (исход $П$), то либо фальшивой и более тяжелой является одна из i «правых» монет, либо фальшивой и более легкой является одна из i «левых» монет — эти случаи отвечают $2i$ исходам β; точно так же случаю, когда перевесила левая чашка (исход $Л$) отвечают еще $2i$ исходов β. Таким образом, три исхода опыта $\alpha_1^{(i)}$ имеют вероятности

$$\frac{2(12-2i)}{24} = \frac{6-i}{6}, \quad \frac{2i}{24} = \frac{i}{12} \text{ и } \frac{i}{12}.$$

Отсюда сразу следует, что из шести опытов $\alpha_1^{(1)}$, $\alpha_1^{(2)}$, $\alpha_1^{(3)}$, $\alpha_1^{(4)}$, $\alpha_1^{(5)}$ и $\alpha_1^{(6)}$ наибольшую энтропию имеет опыт $\alpha_1^{(4)}$, три исхода которого равновероятны; поэтому в этом случае мы получим наибольшую информацию и наиболее целесообразно начинать именно с него. Далее рассмотрим отдельно два случая.

А. *При первом взвешивании чашки весов остались в равновесии.* В таком случае фальшивой является одна из 4 отложенных монет. Нам надо при помощи двух взвешиваний определить, какая именно из них является фальшивой, и выяснить, легче ли она или тяжелее остальных; так как у нас осталось $2 \cdot 4 = 8$ возможных исходов опыта β, а $2 \log 3 = \log 9 > \log 8$, то можно ожидать, что это возможно. Если, однако, положить на каждую чашку весов по одной из наших четырех монет, а две монеты отложить (опыт $\alpha_2^{(1)}$) и чашки весов останутся в равновесии, то последним взвешиванием нам надо будет определить, какой именно из четырех исходов, остающихся еще

возможными, имеет место — а этого сделать нельзя (ибо $4 > 3$). Если же положить на каждую чашку по две из наших четырех монет (опыт $\alpha_2^{(2)}$) и одна из двух чашек перетянет, то у нас снова остаются возможными еще четыре исхода опыта β—и опять нам будут нужны по крайней мере еще два взвешивания, чтобы полностью определить, какой из них имеет место. Таким образом, создается впечатление, что и в случае 12 монет трех взвешиваний недостаточно для решения задачи.

Однако это заключение является преждевременным. Ведь у нас в запасе есть еще $4+4=8$ заведомо настоящих монет, которые могут участвовать во втором взвешивании; поэтому у нас имеется значительно больше двух возможных вариантов опыта α_2. Обозначим через $\alpha_2^{(i,\,j)}$ опыт, состоящий в том, что на правую чашку весов кладутся i из наших четырех подозрительных монет, а на левую $j \leqslant i$ из этих монет и еще $i-j$ заведомо настоящих монет (разумеется, не имеет смысла класть настоящие монеты на обе чашки весов); в таком случае $\alpha_2^{(1,\,1)}$ и $\alpha_2^{(2,\,2)}$ — это те опыты $\alpha_2^{(1)}$ и $\alpha_2^{(2)}$, которые рассматривались выше. Через $p(P)$, $p(П)$ и $p(Л)$ мы обозначим соответственно вероятности того, что при опыте $\alpha_2^{(i,\,j)}$ чашки весов останутся в равновесии и что перетянет правая или левая чашка весов. Эти вероятности легко подсчитать; они равны отношению числа тех исходов β, при которых $\alpha_2^{(i,\,j)}$ имеет исход P, соответственно $П$ или $Л$, к общему числу оставшихся возможными исходов β (это число равно 8). Так как, очевидно, $i+j \leqslant 4$, то все опыты $\alpha_2^{(i,\,j)}$ легко перечислить; отвечающие им значения вероятностей $p(P)$, $p(П)$ и $p(Л)$ собраны в таблице на следующей странице, в которой указана также энтропия (в битах) $H(\alpha_2^{(i,\,j)})$ опыта $\alpha_2^{(i,\,j)}$ (равная $-\,p(P)\log p(P) - p(П)\log p(П) - p(Л)\log p(Л)$).

Из этой таблицы видно, что наибольшую энтропию имеют опыты $\alpha_2^{(2,\,1)}$ и $\alpha_2^{(3,\,0)}$; поэтому для получения наибольшей информации следует в процессе второго взвешивания либо положить на одну чашку весов две из четырех сомнительных монет, а на вторую чашку — одну из сомнительных монет и одну заведомо настоящую, либо положить на одну чашку три сомнительные монеты, а на вторую — три заведомо настоящие. Нетрудно видеть, что в обоих

i	j	$p(P)$	$p(\Pi)$	$p(Л)$	$H(\alpha_2^{(i,\ j)})$
1	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	1,50
1	0	$\frac{3}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	1,06
2	2	0	$\frac{1}{2}$	$\frac{1}{2}$	1,00
2	1	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	1,56
2	0	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	1,50
3	1	0	$\frac{1}{2}$	$\frac{1}{2}$	1,00
3	0	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	1,56
4	0	0	$\frac{1}{2}$	$\frac{1}{2}$	1,00

случаях мы можем затем третьим взвешиванием полностью определить исход β. Действительно, если опыт $\alpha_2^{(2,\ 1)}$ или опыт $\alpha_2^{(3,\ 0)}$ имеет исход P, то фальшивой является единственная сомнительная монета, не участвующая во втором взвешивании; при этом для того, чтобы выяснить, легче она или тяжелее остальных, надо сравнить вес ее с весом одной из 11 заведомо настоящих монет (3-е взвешивание). Если опыт $\alpha_2^{(2,\ 1)}$ имеет исход Π, то либо фальшивой является одна из двух «правых» монет, причем эта монета тяжелее остальных, либо фальшивой является единственная сомнительная монета, лежащая на левой чашке, причем она легче настоящих; сравнив вес двух «правых» монет (3-е взвешивание), мы узнаем исход β (если эти монеты имеют одинаковый вес, то фальшивой является третья из подозреваемых монет; в противном случае — более тяжелая из двух взвешиваемых). Если опыт $\alpha_2^{(3,\ 0)}$ имеет исход Π, то фальшивой является одна из трех лежащих на правой чашке монет, причем она тяжелее настоящих; сравнивая вес двух из этих монет (3-е взвешивание), мы узнаем исход β (фальшивой является более тяжелая из

сравниваемых монет, а если они одинаковы, то третья монета). Аналогично разбираются и случаи, когда опыты $\alpha_2^{(2,1)}$ или $\alpha_2^{(3,0)}$ имеют исход $Л$.

Б. *При первом взвешивании одна из двух чашек весов* (например, *правая*) *перетянула*. В таком случае либо одна из четырех «правых» монет является фальшивой и более тяжелой, чем остальные, либо одна из четырех «левых» монет является фальшивой и более легкой. При втором взвешивании мы можем на правую чашку весов положить i_1 «правых» монет и i_2 «левых», а на левую чашку — j_1 «правых» монет, j_2 «левых» и $(i_1 + i_2) - (j_1 + j_2)$ заведомо настоящих монет из числа не участвующих в первом взвешивании (опыт $\alpha_2^{(i_1, i_2; j_1, j_2)}$; мы считаем, что $i_1 + i_2 \geqslant j_1 + j_2$). Здесь тоже можно было бы составить таблицу энтропий опытов $\alpha_2^{(i_1, i_2; j_1, j_2)}$ при всевозможных значениях i_1, i_2, j_1 и j_2; однако, так как число возможных вариантов тут довольно велико, то некоторые из них целесообразно исключить с самого начала.

Заметим, что так как информация, которую можно получить об исходе β, произведя третье взвешивание (опыт α_3), не превосходит $\log 3$ (ибо $H(\alpha_3) \leqslant \log 3$), то после двух взвешиваний у нас должны остаться н е б о л е е т р е х возможных исходов опыта β: в противном случае опыт α_3 не даст возможности однозначно определить исход β. Отсюда, прежде всего, следует, что число сомнительных монет, не участвующих во втором взвешивании, не должно превосходить 3, так как в случае исхода $Р$ опыта α_2 под подозрением останутся именно эти монеты. Таким образом, имеем

$$8 - (i_1 + i_2 + j_1 + j_2) \leqslant 3, \text{ т. е. } i_1 + i_2 + j_1 + j_2 \geqslant 5$$

или, так как $i_1 + i_2 \geqslant j_1 + j_2$,

$$i_1 + i_2 \geqslant 3, \; j_1 + j_2 \geqslant 5 - (i_1 + i_2).$$

Далее, если опыт $\alpha_2^{(i_1, i_2; j_1, j_2)}$ имеет исход $П$, то либо одна из i_1 «правых» монет, лежащих на правой чашке, является фальшивой и более тяжелой, либо одна из j_2 «левых» монет, лежащих слева, является фальшивой и более легкой. Совершенно так же в случае исхода $Л$ можно подозревать, что фальшивой является одна из i_2 «левых» монет, лежащих справа, или одна из j_1 «правых» монет,

лежащих слева. Отсюда получаем еще два неравенства:

$$i_1 + j_2 \leqslant 3 \quad \text{и} \quad i_2 + j_1 \leqslant 3,$$

выполнения которых естественно требовать. Наконец, ясно, что должны выполняться также неравенства

$$i_1 + j_1 \leqslant 4,\ i_2 + j_2 \leqslant 4 \quad \text{и} \quad (i_1 + i_2) - (j_1 + j_2) \leqslant 4.$$

Перечислим теперь все случаи, удовлетворяющие нашим условиям:

i_1	i_2	j_1	j_2	$p\,(P)$	$p\,(\Pi)$	$p\,(Л)$	$H\,(\alpha_2(i_1, i_2; j_1, j_2))$
2	1	2	1	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	1,56
2	1	2	0	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	1,56
2	1	1	1	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{4}$	1,56
1	2	1	2	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	1,56
1	2	0	2	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{4}$	1,56
1	2	1	1	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	1,56
3	1	1	0	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{4}$	1,56
1	3	0	1	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	1,56
2	2	1	1	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	1,56
2	2	1	0	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{3}{8}$	1,56
2	2	0	1	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{4}$	1,56
3	2	1	0	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	1,56
2	3	0	1	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	1,56

Таким образом, мы видим, что здесь имеются уже не 2, как в предыдущем случае, а целых 13 вариантов опыта α_2, при которых этот опыт содержит одну и ту же наибольшую информацию относительного опыта β (совершенно ясно, что здесь информация $I(\alpha_2, \beta)$ равна энтропии $H(\alpha_2)$). При любом выборе опыта α_2 этой информации оказывается достаточно для того, чтобы иметь возможность полностью определить исход β с помощью еще одного, 3-го, взвешивания. Так, например, в случае исхода P опыта $\alpha_2^{(2,1;2,1)}$ фальшивой является одна из двух не участвующих во втором взвешивании «левых» монет; так как мы знаем, что эта монета легче настоящей, то для того, чтобы найти ее, достаточно сравнить вес этих двух монет (или сравнить одну из них с заведомо настоящей). В случае исхода $П$ того же опыта либо фальшивой и более тяжелой окажется одна из двух лежащих на правой чашке «правых» монет, либо фальшивой и более легкой будет единственная «левая» монета, лежащая на левой чашке; для того чтобы найти фальшивую монету, нам достаточно сравнить вес двух подозреваемых «правых» монет. Точно так же исследуется случай, когда опыт $\alpha_2^{(2,1;2,1)}$ имеет исход $Л$.

Этим полностью завершается исследование случая 12 монет. Теперь мы можем вернуться к случаю 13 монет и доказать, что в этом случае достаточно четырех взвешиваний (ранее мы показали только, что трех взвешиваний в этом случае может не хватить). Положим на обе чаши весов по 4 монеты; пять монет отложим в сторону. Если одна из двух чашек весов перетянет, то мы будет иметь ту же ситуацию, с которой встретились, разбирая случай исхода $П$ первого взвешивания в задаче о 12 монетах (с той несущественной разницей, что теперь у нас имеются уже не 4, а 5 заведомо настоящих монет); и этом случае уже тремя взвешиваниями можно обнаружить фальшивую монету и узнать, легче ли она или тяжелее остальных. Если же весы окажутся в равновесии, то нам надо будет выделить фальшивую монету не из 4, а из 5 сомнительных. Здесь мы можем начать со сравнения веса какой-нибудь одной из подозреваемых монет и заведомо настоящей: если их веса окажутся различными, то наша задача сразу будет разрешена; в противном случае мы снова придем к случаю 4 сомнительных монет, при котором двумя взвешиваниями можно определить фальшивую монету и выяснить, легче

ли она или тяжелее остальных (см. случай А на стр. 151 и след.).

Обобщим теперь условия предыдущей задачи:

З а д а ч а 29. *Имеется n монет одного достоинства; одна из этих монет — фальшивая, более легкая или более тяжелая, чем остальные. Каково наименьшее число k взвешиваний на чашечных весах без гирь, которое позволяет найти фальшивую монету и определить, легче ли она или тяжелее, чем остальные* (ср. Д. О. Шклярский и др. [51], задача 66)).

Прежде всего, поскольку энтропия рассматриваемого здесь опыта β (все исходы которого мы, как всегда, считаем равновероятными) равна $\log 2n$, а энтропия опыта $A_k = \alpha_1\alpha_2 \dots \alpha_k$, состоящего в k-кратном взвешивании, не превосходит $k \log 3 = \log 3^k$, то должно быть

$$2n \leqslant 3^k, \text{ т. е. } n \leqslant \frac{3^k}{2}$$

или, так как n и k — целые числа и 3^k — нечетно,

$$n \leqslant \frac{3^k - 1}{2}.$$

Следовательно, можно утверждать, что

$$k \geqslant \log_3 (2n + 1) = \frac{\log (2n + 1)}{\log 3}.$$

Так, например, если $n > \frac{3^3 - 1}{2} = 13$, то фальшивая монета н е м о ж е т быть определена менее чем тремя взвешиваниями.

Нетрудно видеть также, что даже и в том случае, когда $n = \frac{3^k - 1}{2}$, k взвешиваний н е в с е г д а п о з в о л я ю т обнаружить фальшивую монету и определить, легче ли она или тяжелее остальных (так, при $n = 13$ фальшивая монета не во всех случаях может быть определена тремя взвешиваниями). Доказательство этого в общем случае принципиально не отличается от приведенного выше доказательства для частного случая $n = 13$ и $k = 3$

(см. начало решения задачи 28). Действительно, при оценке энтропии опыта $A_k = \alpha_1\alpha_2 \ldots \alpha_k$ мы до сих пор исходили из того, что энтропия каждого отдельного взвешивания может равняться $\log 3$; в нашем случае, однако, из-за того, что $n = \frac{3^k - 1}{2}$ не делится на 3, уже энтропия первого взвешивания (опыт α_1) не может достигнуть этого значения (ибо три исхода первого взвешивания никак не могут быть равновероятны). Поскольку $n - 1 = \frac{3(3^{k-1} - 1)}{2}$ делится на 3, то ясно, что выгоднее всего при первом взвешивании на каждую чашку весов положить по

$$\frac{n-1}{3} = \frac{3^{k-1} - 1}{2}$$

монет, а остальные

$$\frac{n+2}{3} = \frac{3^{k-1} + 1}{2}$$

монет отложить в сторону; в этом случае вероятности трех исходов опыта α_1 $\left(\text{равные } \frac{n-1}{3} : n = \frac{1}{3} - \frac{1}{3n},\ \frac{n-1}{3} : n = \right.$ $\left. = \frac{1}{3} - \frac{1}{3n} \text{ и } \frac{n+2}{3} : n = \frac{1}{3} + \frac{2}{3n}\right)$ будут ближе всего друг к другу и, следовательно, энтропия $H(\alpha_1)$ соответствующего опыта будет больше, чем в любом другом случае. Но легко убедиться, что остающаяся после этого степень неопределенности такова, что она не может быть полностью уничтожена при помощи $k - 1$ взвешиваний. Проще всего это показывается так: предположим, что при первом взвешивании чашки весов окажутся в равновесии; в таком случае фальшивая монета находится среди $\frac{n+2}{3} = \frac{3^{k-1} + 1}{2}$ монет, отложенных в сторону при этом взвешивании, так что у нас останутся еще $3^{k-1} + 1$ равновероятных исходов интересующего нас опыта β (фальшивой может оказаться любая из $\frac{3^{k-1} + 1}{2}$ отложенных монет и она может быть или легче или тяжелее настоящих). Выяснив, какая из этих возможностей на самом деле имеет место, мы получим информацию, равную $\log(3^{k-1} + 1)$, что превосходит наи-

большую информацию $\log 3^{k-1} = (k-1)\log 3$, которую можно получить в результате $k-1$ взвешиваний. Аналогично показывается, что при любом другом выборе опыта α_1 (первого взвешивания) этот опыт может иметь такой исход, при котором оставшихся $k-1$ взвешиваний будет недостаточно для однозначного выяснения исхода опыта β.

Итак, мы видим, что *если*

$$n \geqslant \frac{3^k-1}{2},$$

то k взвешиваний может оказаться недостаточно. Покажем теперь, что *если* $n < \frac{3^k-1}{2}$ (т. е. если $n \leqslant \frac{3^k-3}{2}$; другими словами, если $k \geqslant \log_3(2n+3) = \log\frac{(2n+3)}{\log 3}$), *то k взвешиваний уже будет достаточно* [1]); этим решение нашей задачи будет завершено.

Начнем со следующей вспомогательной задачи: пусть, кроме n монет, одна из которых фальшивая, у нас имеется по крайней мере одна заведомо настоящая монета; требуется выделить фальшивую монету и определить, легче ли она или тяжелее остальных. В этом случае мы по-прежнему можем утверждать, что если $n > \frac{3^k-1}{2}$, то k взвешиваний будет недостаточно (ибо степень неопределенности исходного опыта от добавления настоящих монет, разумеется, не изменится). Однако теперь мы уже не можем быть уверены, что и при $n = \frac{3^k-1}{2}$ фальшивую монету заведомо нельзя определить при помощи k взвешиваний. В самом деле, использовав дополнительную настоящую монету, мы можем добиться большей, чем раньше, близости вероятностей трех исходов первого взвешивания и, следовательно, получить при этом взвешивании бóльшую информацию; для этого надо только положить на каждую чашку весов по $\frac{n+2}{3} = \frac{3^{k-1}+1}{2}$ монет (одна из использованных $3^{k-1}+1$ монет — имеющаяся у нас

[1]) Это утверждение имеет два очевидных исключения: если $n = 1$, то нельзя определить, легче или тяжелее фальшивая монета настоящих (которых в этом случае нет совсем); если $n = 2$, то фальшивую монету невозможно выделить.

настоящая монета), а остальные $\frac{n-1}{3} = \frac{3^{k-1}-1}{2}$ сомнительных монет отложить в сторону. В таком случае вероятности отдельных исходов первого взвешивания будут, как легко видеть, равны

$$\left[\frac{n+2}{3} + \left(\frac{n+2}{3} - 1\right)\right] : 2n = \frac{1}{3} + \frac{1}{6n}, \quad \frac{1}{3} + \frac{1}{6n}$$

и

$$\frac{n-1}{3} : n = \frac{1}{3} - \frac{1}{3n},$$

т. е. они действительно будут несколько более близки друг к другу, чем раньше; следовательно, и энтропия $H(\alpha_1)$ опыта α_1 здесь будет больше. Этой небольшой разницы уже оказывается достаточно для того, чтобы обеспечить возможность выделения фальшивой монеты и определения того, легче ли она или тяжелее других, при помощи k взвешиваний.

Для доказательства того, что при наличии в нашем распоряжении хоть одной заведомо настоящей монеты при $n \leqslant \frac{3^k-1}{2}$ можно обойтись k взвешиваниями, удобно воспользоваться методом математической индукции. Это утверждение совершенно очевидно при $k = 1$ (т. е. при $n = 1$). Предположим теперь, что оно уже доказано для некоторого значения k и покажем, что в таком случае при $\frac{3^k-1}{2} < n \leqslant \frac{3^{k+1}-1}{2}$ оказывается достаточно $k+1$ взвешиваний; отсюда уже будет следовать справедливость нашего утверждения во всех случаях. При первом взвешивании на одну чашку весов положим какое-то число x из наших n монет, а на вторую $x-1$ из n монет и еще одну заведомо настоящую монету; неиспользованными при этом окажутся $n_1 = n - (2x-1)$ монет. Число x выберем так, чтобы было

$$2x - 1 \leqslant 3^k \quad \text{и} \quad n - (2x-1) \leqslant \frac{3^k-1}{2},$$

т. е.

$$3^k \geqslant 2x - 1 \geqslant n - \frac{3^k-1}{2};$$

ясно, что при $n \leqslant \frac{3^{k+1}-1}{2}$ это сделать можно (ибо $n - \frac{3^k-1}{2} \leqslant \frac{3^{k+1}-1}{2} - \frac{3^k-1}{2} = 3^k$). Если при первом взвешивании чашки весов окажутся в равновесии, то в дальнейшем нам останется только выделить фальшивую монету из числа $n_1 \leqslant \frac{3^k-1}{2}$ отложенных монет; так как у нас имеются, кроме того, заведомо настоящие монеты, то (в силу предположения индукции) это можно сделать k взвешиваниями. Если же одна из чашек перевесит, то у нас останутся $2x - 1 \leqslant 3^k$ сомнительных монет; при этом мы будем знать, что если фальшивой является одна из некоторых a монет, то она является более легкой, чем остальные, а если одна из b остальных ($a + b \leqslant 3^k$), то более тяжелой (если перетянула первая чашка, то $a = x - 1$, $b = x$; если перетянула вторая, то $a = x$, $b = x - 1$). В этом случае также k последующими взвешиваниями всегда можно выделить фальшивую монету (см. выше, стр. 148—149).

Вернемся теперь к нашим первоначальным $n \leqslant \frac{3^k-3}{2}$ монетам, одна из которых является фальшивой. При первом взвешивании положим на обе чашки весов по $\frac{3^{k-1}-1}{2}$ монет; неиспользованными при этом останутся

$$n_1 = n - 2\frac{3^{k-1}-1}{2} \leqslant \frac{3^k-3}{2} - (3^{k-1}-1) = \frac{3^{k-1}-1}{2}$$

монет [1]). Если весы останутся в равновесии, то сомнительными будут $n_1 \leqslant \frac{3^{k-1}-1}{2}$ неиспользованных монет; так как, кроме того, мы будем иметь еще $3^{k-1} - 1$ заведомо настоящих монет, то, согласно доказанному выше, при помощи последующих $k - 1$ взвешиваний мы сможем выделить фальшивую монету и определить, легче ли она или

---

[1]) В том случае, когда n р а в н о $\frac{3^k-3}{2}$, информация $I(\alpha_1, \beta)$ относительно β, содержащаяся в нашем опыте α_1 (1-е взвешивание), будет точно равна $\log 3$.

тяжелее настоящих. Если же одна из двух чашек перетянет, то мы будем иметь $3^{k-1}-1<3^{k-1}$ фальшивых монет и будем знать, что если фальшивой является одна из определенных $a=\frac{3^{k-1}-1}{2}$ монет, то она легче настоящих, а если одна из $b=\frac{3^{k-1}-1}{2}$ $(=a)$ других, то тяжелее; в силу сказанного на стр. 148 здесь мы тоже при помощи $k-1$ последующих взвешиваний сможем выделить фальшивую монету. Тем самым доказательство сделанного ранее утверждения о требующемся числе взвешиваний полностью завершено.

Отметим еще, что при большом n число k, определяемое из неравенств

$$k-1<\frac{\log(2n+3)}{\log 3}\leqslant k,$$

с большой степенью точности можно заменить отношением $\frac{\log 2n}{\log 3}$ (в том смысле, что отношение $k:\frac{\log 2n}{\log 3}$ при возрастании n быстро стремится к единице).

Условия задач, связанных с определением фальшивых монет с помощью взвешиваний на чашечных весах без гирь, можно, разумеется, весьма разнообразить. До сих пор мы всегда считали, что лишь о д н а из имеющихся монет является фальшивой (имеет вес, отличающийся от веса остальных монет); можно, однако, также предположить, что среди заданных монет имеются две или большее число фальшивых; еще более трудны задачи, в которых само число фальшивых монет предполагается неизвестным [1]). Можно также считать, что фальшивые монеты могут быть двух или более различных весов; представление о возникающих при этом новых задачах может дать следующая проблема, поставленная польским математиком Г. Ш т е й н г а у з о м ([56], стр. 42):

---

[1]) По поводу случая двух или более фальшивых задач см., например, [52] (ср. также [53]); общему случаю посвящены статьи [54] и [55], в последней из которых подробно обсуждаются несколько разных вариантов задачи о фальшивой монете (с указанием на их возможное прикладное значение) и приведена обширная библиография.

Задача 30. *Имеется 4 предмета разного веса и чашечные весы без гирь, на которых можно сравнивать веса любых двух предметов. Укажите способ, позволяющий с помощью самое большее пяти взвешиваний установить последовательность весов этих предметов. Докажите, что не существует способа, гарантирующего возможность установления очередности весов предметов при помощи не более чем четырех взвешиваний.*

Для 10 предметов попарно разного веса существует способ установления очередности весов предметов с помощью не более 24 взвешиваний (найдите его); можно ли это число взвешиваний уменьшить?

Полное решение этой задачи (в которой, разумеется, число предметов на самом деле может быть произвольным) до сих пор неизвестно; некоторые частные относящиеся сюда результаты можно найти, например, в работах [57] и [58] [1]). Существует и ряд других задач подобного рода (об этом будет подробнее сказано в следующем параграфе); как правило, они являются весьма трудными, но теория информации помогает отыскать хотя бы общий подход к их исследованию.

§ 3. Обсуждение

В §§ 1 и 2 этой главы понятия энтропии и информации, введенные в гл. II, применялись к анализу некоторых логических задач типа «математических развлечений». Из дальнейшего будет видно, что рассуждения того же рода оказываются полезными также и при решении ряда достаточно серьезных вопросов. Тем более целесообразно обсудить здесь подробнее общую идею всех рассмотренных примеров; при этом мы, естественно, придем также к несколько более общей постановке задач, весьма важной для следующей главы.

Все примеры в §§ 1 и 2 были построены по одной схеме. Во всех этих примерах нас интересовал некоторый объект из конечного множества M однотипных объектов; так, в

[1]) Ср. также рассчитанную на учащихся средней школы статью: Г. М. Адельсон-Вельский, И. Н. Бернштейн, М. Л. Гервер, Кто поедет в Рио?, журнал «Квант», № 8, 1972, стр. 2—8, посвященную близкой тематике.

задачах 23—24 множество M состояло из нескольких *городов* — и требовалось определить, в каком городе находится наблюдатель Н.; в задаче 25 множество M состояло из целых положительных *чисел*, а в задаче 26 — из $C_{100}^2 = 4950$ *пар чисел*; в задачах 27—29 множество M состояло из *монет* — и нам требовалось выделить одну из этих монет, а именно, фальшивую монету; наконец, в задаче 30 множество M состояло из всевозможных *упорядоченных наборов* имеющихся в нашем распоряжении *предметов* (так что в случае 4-х предметов множество M содержало $4! = 24$ всевозможных упорядочиваний наших предметов) — и задача состояла в том, чтобы выяснить, какой из этих порядков соответствует последовательности весов предметов, начиная с самого тяжелого из них и кончая самым легким. Употребляя привычную нам из первых двух глав книги терминологию, можно сказать, что мы говорили об о п ы т е β, могущем иметь n различных исходов $B_1, B_2, \ldots, B_n$; множество всех этих исходов мы и обозначили через M. Для выделения интересующего нас объекта (исхода опыта β) использовались в с п о м о г а т е л ь н ы е о п ы т ы α; каждый из этих опытов мог иметь $m < n$ возможных исходов (этими опытами α были либо вопросы, могущие иметь 2 разных ответа: «*да*» и «*нет*», либо взвешивания на чашечных весах без гирь, могущие иметь 3 разных исхода P, $П$ и $Л$)—и эти исходы выделяли те или иные подмножества множества M исходов β, позволяя отбрасывать ряд из исходов $B_1, B_2, \ldots, B_n$ как «ложные» или «не выполняющиеся». Требовалось указать наименьшее число вспомогательных опытов α, необходимых для выяснения правильного ответа на интересующий нас вопрос (т. е. для установления исхода опыта β) и описать, как именно можно быстрее всего найти этот ответ.

Сходную с описанной структуру имеют не только «игрушечные» задачи §§ 1—2, но и многие жизненно важные проблемы; к ним относятся, например, в первую очередь интересующие нас в этой книге задачи рационального кодирования сообщений (см. гл. IV), задачи сортировки предметов по тем или иным критериям, задачи отыскания слова в словаре или требуемой книги в большой библиотеке, задачи составления рациональных программ контроля тех или иных объектов, например станков на фабрике и т. д.

Такая широта возможных приложений вызвала в наше время большой интерес к тематике §§ 1 и 2 и привела к созданию разработанной терминологии. Системы опытов α, приводящие к обнаружению интересующего нас объекта (исхода опыта β), называются в о п р о с н и к а м и, а сами опыты α — в о п р о с а м и; при этом вопросы могут различаться как *числом* возможных *ответов* [1]), так и — в ряде случаев — «ценой вопроса», характеризующей затраты, которых требует соответствующий опыт α, или усилия, которые надо приложить для «получения ответа» (т. е. для выяснения исхода α). Задача состоит в том, чтобы отыскать такую процедуру «постановки вопросов» (т. е. такую последовательность опытов α), которая приводит к требуемому ответу (к исходу опыта β) с помощью наименьшей (по числу или по общей «цене») цепочки «вопросов». Теории вопросников посвящена сегодня огромная литература, из которой мы назовем только монографии [59] французского математика К. П и к а р а и обзорную статью [60].

Ясно, что во всех задачах рассматриваемого рода фактически требуется наиболее целесообразно использовать ту информацию об исходе опыта β, которая содержится в результатах вспомогательных опытов α. Однако представляется, что слово «информация» здесь употребляется в обычном, «житейском» смысле, а вовсе не в том более специальном смысле, которое мы придавали ему в гл. II. В самом деле, введенная в гл. II величина I имела чисто с т а т и с т и ч е с к и й смысл — ведь само ее определение базировалось на понятии вероятности. В наших же задачах не фигурируют никакие многократно повторяющиеся испытания и не участвуют никакие вероятности; поэтому возможность применения к этим задачам развитой в гл. II теории на первый взгляд может показаться удивительной.

---

[1]) В принципе не исключена также и ситуация, когда разные опыты α имеют р а з н о е число возможных ответов; так, например, можно представить себе такой вариант задачи об отгадывании фальшивой монеты, когда ищущее эту монету лицо может либо задавать вопросы человеку, знающему, какая монета является фальшивой (такой опыт может иметь д в а ответа: «*да*» и «*нет*»), либо прибегнуть к взвешиванию монет (этот опыт может иметь т р и ответа: *Р*, *П* и *Л*).

Выход из создавшегося положения, которым мы фактически все время пользовались, состоит в следующем. Предположим, что мы много раз решаем одну и ту же задачу (т. е. ищем правильный ответ на один и тот же вопрос), причем правильные ответы оказываются в разных случаях различными и каждый из ответов имеет определенную вероятность оказаться правильным; соответствующие вероятности $p(B_1)$, $p(B_2)$, ..., $p(B_n)$ мы считаем произвольными, но заданными нам заранее. В таком случае мы можем говорить об «опыте β, состоящем в выяснении правильного ответа», употребляя слово «опыт» в точности в том самом смысле, в каком употреблялось это слово в предыдущей главе. Опыту β отвечает таблица вероятностей

исходы опыта	B_1	B_2	...	B_n
вероятности	$p(B_1)$	$p(B_2)$	...	$p(B_n)$

и энтропия $-p(B_1)\log p(B_1) - p(B_2)\log p(B_2) - \ldots - p(B_n)\log p(B_n)$, которую мы, как обычно, обозначим через $H(\beta)$. Так как вспомогательные опыты α у нас всегда были «прямо направлены» к выяснению исхода β в том смысле, что знание исхода β полностью определяло и исход α, то задание вероятностей n исходов опыта β позволяет определить и вероятности m исходов любого такого опыта α_1; поэтому и в применении к α_1 слово «опыт» можно употреблять в том же смысле, что и в гл. II. Далее, из того, что исход β полностью определяет исход α_1, следует, что условная энтропия $H_\beta(\alpha_1)$ равна нулю, а условная энтропия $H_{\alpha_1}(\beta)$ — разности $H(\beta) - H(\alpha_1)$ энтропий опытов β и α_1 (см. стр. 95). Но условная энтропия $H_{\alpha_1}(\beta)$ равна среднему значению энтропий $H_{A_1}(\beta)$, ..., $H_{A_m}(\beta)$ опыта β, отвечающих различным возможным исходам $A_1, \ldots, A_m$ опыта α_1. Поэтому хоть при одном исходе A_i из числа этих m исходов энтропия $H_{A_i}(\beta)$ окажется не меньше, чем $H(\beta) - H(\alpha_1)$; таким образом, наверное возможны случаи, когда после выяснения результата испытания α_1 остающаяся энтропия (степень неопределенности) опыта β будет не меньше разности $H(\beta) - H(\alpha_1)$.

Ясно, как можно обобщить последнее рассуждение. Выберем какую угодно последовательность вспомогательных опытов (испытаний) $\alpha_1, \alpha_2, \ldots, \alpha_k$, т. е. рассмотрим некоторый *сложный опыт* $A_k = \alpha_1\alpha_2 \ldots \alpha_k$. Будем считать при этом, что отдельные опыты $\alpha_1, \alpha_2, \ldots, \alpha_k$ не должны быть независимы, т. е. что результаты предшествующих испытаний могут отражаться на условиях проведения последующих; возможно даже, что при некоторых специальных исходах нескольких первых опытов α все последующие опыты становятся ненужными, т. е. могут пониматься как опыты, имеющие единственный строго определенный исход (это означает, что сложный опыт A_k состоит из *не более чем* k опытов α, но не обязательно точно из k таких опытов). В рассмотренных выше примерах знание исхода β всегда определяло исход сложного опыта A_k, так что по вероятностям отдельных исходов β можно было найти и вероятности различных исходов сложного опыта A_k; поэтому употребление слова «опыт» в применении к A_k также не может вызвать недоразумений.

Заметим еще, что если каждый из опытов $\alpha_1, \alpha_2, \ldots, \alpha_k$ может иметь не более чем m исходов, то общее число различных исходов A_k не превосходит m^k. Из того, что исход β определяет исход A_k, следует, что средняя условная энтропия $H_{A_k}(\beta)$ опыта β при условии осуществления сложного опыта A_k равна разности $H(\beta) - H(A_k)$ энтропий опытов β и A_k; поэтому хотя бы при одном исходе A_k (т. е. при каких-то определенных исходах k испытаний $\alpha_1, \alpha_2, \ldots, \alpha_k$) «остаточная энтропия» β будет не меньше, чем $H(\beta) - H(A_k)$.

Предположим теперь, что разность $H(\beta) - H(A_k)$ больше нуля. В таком случае хотя бы при одном исходе сложного опыта A_k у нас будет оставаться еще некоторая неопределенность в исходе опыта β; другими словами, при многократном повторении всей серии из k опытов α и выделении лишь тех случаев, когда эти опыты имели некоторые определенные заранее результаты, правильным будет оказываться иногда один, а иногда другой из ответов на наш основной вопрос. Отсюда вытекает, что в случаях, когда сложный опыт A_k имеет указанный исход, мы не сможем по этому исходу однозначно выяснить, какой именно из ответов на рассматриваемый в задаче вопрос

является правильным; значит, k опытов α здесь не хватает для такого выяснения.

Именно это рассуждение и использовалось выше для решения задач 23—29. При этом учитывалось еще то обстоятельство, что подобного рода вывод о невозможности выяснения исхода β по исходам k опытов α может быть сделан всегда, когда х о т я б ы п р и о д н о м в ы б о р е вероятностей $p(B_1)$, $p(B_2)$, ..., $p(B_n)$ исходов β имеет место неравенство $H(\beta) - H(A_k) > 0$. Обычно оказывается достаточно рассмотреть лишь «самый невыгодный» случай, когда энтропия опыта β принимает наибольшее значение, т. е. когда все исходы этого опыта являются равновероятными: $p(B_1) = p(B_2) = \ldots = p(B_n) = \frac{1}{n}$; именно так мы и поступали выше, когда говорили, что «из-за отсутствия каких-либо сведений о возможных исходах β мы будем считать все эти исходы равновероятными». Ясно, что при таком выборе вероятностей исходов β будет справедливо равенство: $H(\beta) = \log n$. Что же касается сложного опыта A_k, то точный подсчет его энтропии в конкретных задачах часто не прост; однако во многих случаях удается ограничиться простейшей оценкой $H(A_k) \leqslant \log m^k = k \log m$, вытекающей из того факта, что число различных исходов A_k не может превосходить m^k. В более сложных случаях мы точно подсчитывали наибольшую «остаточную энтропию» опыта β, отвечающую наиболее «неудачному» исходу п е р в о г о опыта α_1, и лишь после этого в отношении последующих опытов α_2, ..., α_k просто использовали то, что энтропия каждого из них не превосходит $\log m$ (ср. выше, стр. 150—151 и 157—158). Отметим еще, что оценка $H(A_k) \leqslant k \log m$ сразу приводит к важному неравенству

$$k \geqslant \frac{\log n}{\log m}, \tag{1}$$

которое, разумеется, можно вывести и не используя понятий теории информации: оно означает лишь, что при наличии n разных возможностей нельзя однозначно выделить одну из них при помощи сложного опыта, могущего иметь меньше чем n различных исходов [1]). Выше наша оценка

[1]) Подчеркнем, что подсчет числа имеющихся возможностей здесь равносилен использованию простейшего понятия степени неопределенности «в смысле Хартли» (ср. стр. 79).

нужного числа опытов α часто сводилась к использованию лишь этого простейшего неравенства.

Наш основной вывод о невозможности однозначного определения исхода β по исходу сложного опыта A_k в случаях, когда $H(\beta) - H(A_k) > 0$, можно обосновать и несколько иначе. Если исход нашего сложного опыта A_k во всех случаях однозначно определяет исход β, то $H_{A_k}(\beta) = 0$ и, значит, в силу равенства $I(A_k, \beta) = H(\beta) - H_{A_k}(\beta)$ инфорамация $I(A_k, \beta)$ об опыте β, содержащаяся в опыте A_k, должна быть равна степени неопределенности β, т. е. $I(A_k, \beta) = H(\beta)$. С другой стороны, так как в наших случаях исход опыта β однозначно определял исход сложного опыта A_k, то одновременно $I(\beta, A_k) = H(A_k)$. Таким образом, *если сложный опыт* A_k (состоящий не более чем из k испытаний α) *позволяет во всех случаях однозначно указать правильный ответ на поставленный вопрос* (т. е. найти исход опыта β), *то должно иметь место равенство* $H(A_k) = H(\beta)$. Например, в условиях задачи 29 легко видеть, что $H(\alpha_1) = \log 3 \approx 1{,}58$ бит (все исходы первого взвешивания у нас были равновероятны); далее, при любом исходе первого взвешивания второе из них (опыт α_2) выбиралось так, что его три исхода имели вероятности $\frac{1}{4}$, $\frac{3}{8}$ и $\frac{3}{8}$ и, следовательно, $H_{\alpha_1}(\alpha_2) = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{8}\log\frac{3}{8} - \frac{3}{8}\log\frac{3}{8} \approx 1{,}56$ бит (см. стр. 153 и 155); наконец, третье взвешивание (опыт α_3) в случае, когда α_2 имело исход с вероятностью $\frac{1}{4}$, сводилось к сравнению на чашечных весах двух монет заведомо разного веса, т. е. имело энтропию $\log 2 = 1$, а в остальных $\frac{3}{4}$ всех случаев (при любом из двух исходов α_2 с вероятностями $\frac{3}{8}$) оно могло иметь три равновероятных исхода, т. е. имело энтропию $\log 3$. Поэтому здесь $H_{\alpha_1\alpha_2}(\alpha_3) = \frac{1}{4}\log 2 + \frac{3}{4}\log 3 \approx 1{,}44$ бит, и так как $H(\beta) = \log 24 \approx 4{,}58$ бит, то

$$H(A_3) = H(\alpha_1\alpha_2\alpha_3) = H(\alpha_1) + H_{\alpha_1}(\alpha_2) + H_{\alpha_1\alpha_2}(\alpha_3) \approx$$
$$\approx 1{,}58 + 1{,}56 + 1{,}44 = 4{,}58 \text{ бит} = H(\beta),$$

как это и должно быть. *Если же* равенство $H(A_k) = H(\beta)$ не выполняется, а *имеет место неравенство* $H(A_k) < H(\beta)$, *то* это означает, что *опыт* A_k *наверное не позволяет однозначно указать правильный ответ.*

Легко понять также, что предположение о том, что *исход* β *полностью определяет исходы испытаний* α, не является необходимым для справедливости последнего вывода. Если это предположение не выполняется, то задание вероятностей отдельных исходов β не позволяет однозначно судить о вероятностях всех исходов вспомогательных опытов α. Поэтому, предполагая, что опыты по определению исхода β с помощью испытания α производятся м н о г о к р а т н о, здесь следует дополнительно задать и эти последние вероятности (разумеется так, чтобы их значения не противоречили заданным раньше значениям вероятностей исходов β). При этом по-прежнему, если сложный опыт $A_k = \alpha_1\alpha_2 \ldots \alpha_k$, состоящий не более чем из k испытаний α, полностью определяет исход β, то информация $I(A_k, \beta) = H(\beta) - H_{A_k}(\beta)$ равна энтропии $H(\beta)$; с другой стороны, поскольку всегда $I(A_k, \beta) = H(A_k) - H_\beta(A_k) \leqslant H(A_k)$, то должно иметь место неравенство $H(\beta) \leqslant H(A_k)$. Таким образом по-прежнему, *если*

$$H(A_k) < H(\beta),$$

то исход сложного опыта $A_k = \alpha_1\alpha_2 \ldots \alpha_k$ *не может во всех случаях однозначно определить исход* β; отсюда уже можно получить определенную оценку наименьшего числа k испытаний α, позволяющих определить исход β. Однако в рассматриваемом здесь случае получаемая таким путем оценка оказывается заметно менее точной, чем в случае, когда исход β однозначно определяет исходы всех испытаний α. Это связано с тем, что теперь испытания α уже не направлены прямо к выяснению исхода β и, следовательно, информация $I(A_k, \beta)$ относительно β, содержащаяся в k испытаниях $\alpha_1, \alpha_2, \ldots, \alpha_k$, уже не равна энтропии $H(A_k)$, а меньше этой энтропии.

Для примера предположим, что в условии задачи 29 (см. стр. 157) нам не требуется выяснить, является ли фальшивая монета более тяжелой или более легкой, чем настоящие (надо лишь у к а з а т ь эту фальшивую монету). Будем считать, что каждая из n имеющихся монет с определенной вероятностью может оказаться фальшивой; в

таком случае мы можем подсчитать вероятности всех исходов опыта β. Если кроме того, считать, что фальшивая монета имеет определенную вероятность оказаться более тяжелой или более легкой, чем остальные, то можно будет определить и вероятности всех исходов любых испытаний α, что позволяет с полным правом говорить об энтропии опытов α и β и об информации, содержащейся в одном из них относительно другого. В частности, если считать все исходы опыта β равновероятными (т. е. считать, что каждая из n монет имеет одинаковую вероятность оказаться фальшивой), то энтропия $H(\beta)$ опыта β будет равна $\log n$; с другой стороны, энтропия каждого из опытов α не превосходит $\log 3$ (ибо подобный опыт по-прежнему может иметь 3 разных исхода: *Р*, *П* и *Л*), а энтропия сложного опыта $A_k = \alpha_1\alpha_2 \ldots \alpha_k$ не превосходит $k \log 3$. Отсюда вытекает, что наименьшее число k взвешиваний, необходимое для определения фальшивой монеты, должно удовлетворять неравенству

$$k \geqslant \frac{\log n}{\log 3} \tag{2}$$

Эта оценка приводит к меньшему числу k, чем аналогичная оценка наименьшего числа взвешиваний, необходимых для обнаружения фальшивой монеты и выяснения того, легче ли она или тяжелее остальных, имеющая вид:

$$k \geqslant \frac{\log 2n}{\log 3} \tag{3}$$

(ибо здесь опыт β имеет $2n$ разных исходов, так как каждая монета может оказаться и легче и тяжелее остальных). Но оценка (3) является довольно точной: так, при $k = 3$ она дает $n \leqslant 13$, а на самом деле, как мы знаем, наибольшее число монет, из которых можно тремя взвешиваниями выделить фальшивую монету и обнаружить, легче ли она или тяжелее остальных, равно 12 (см. выше задачу 28). В противоположность этому оценка (2) очень неточна: из нее вытекает лишь, что при $k = 3$ число $n \leqslant 27$; на самом же деле, однако, можно проверить, что наибольшее число монет, из числа которых можно тремя взвешиваниями выделить фальшивую монету, не выясняя, легче ли она или тяжелее остальных, равно лишь 13. Причина

этого кроется в том, что здесь опыты α (т. е. взвешивания монет) не будут прямо направлены к определению исхода β (они будут содержать «постороннюю» информацию, а именно информацию о весе фальшивой монеты); поэтому каждый такой опыт вносит заметно меньший, чем $\log 3$, вклад в накапливаемую информацию об исходе β и, следовательно, число опытов α должно быть б о л ь ш и м, чем $\frac{\log n}{\log 3}$.

Обратимся теперь к вопросу о том, как можно доказать, что при помощи не более чем k вспомогательных опытов α действительно м о ж н о однозначно определить исход интересующего нас опыта β; до сих пор мы говорили лишь о д о к а з а т е л ь с т в а х н е в о з м о ж н о с т и выяснения исхода β с помощью слишком малого числа испытаний α. Подобное «доказательство возможности» требует явного указания наиболее целесообразной цепочки α_1, α_2, ..., α_k вспомогательных опытов, или, другими словами, указания соответствующего сложного опыта A_k. Разумеется, получаемое при этом «решение» не содержит непосредственно понятий энтропии и информации. Однако эти понятия играют важную эвристическую роль, помогая быстрее всего отыскать соответствующую цепочку испытаний. В самом деле, цель наших испытаний состоит в определении исхода опыта β, т. е. в получении полной информации об этом опыте; поэтому естественно подбирать эти испытания таким образом, чтобы они содержали возможно большую информацию об исходе β. Строгий метод решения задачи заключается в перечислении всех возможных сложных опытов $A_k = \alpha_1\alpha_2 \ldots \alpha_k$, вычислении для каждого из них информации $I(A_k, \beta)$ и отборе тех A_k, для которых $I(A_k, \beta) = H(\beta)$; в случае, когда исход β однозначно определяет исход всех испытаний α, вычисление информации значительно облегчается тем обстоятельством, что здесь $I(A_k, \beta) = H(A_k)$. Так как, однако, оперировать сразу со сложными опытами A_k довольно неудобно, то практически обычно начинают с определения того из вспомогательных опытов α_1 (1-е испытание), который содержит наибольшую информацию $I(\alpha_1, \beta)$ об исходе опыта β; далее подбирают второе испытание α_2 (зависящее, вообще говоря, от исхода α_1) так, чтобы информация $I(\alpha_1\alpha_2, \beta)$ была возможно большей и

т. д.; именно так мы и поступали выше при решении задач 23—29 [1]).

В §§1 и 2 все время предполагалось, что все исходы опыта β являются р а в н о в е р о я т н ы м и. Это предположение означает, что все исходы β считаются равноправными; оно является вполне естественным, так как нам надо, чтобы обнаружение исхода β не требовало большого числа испытаний, к а к и м б ы н и б ы л э т о т исход. Ясно, что удовлетворяющий этому условию путь определения исхода β приводит, вообще говоря, к сложному опыту A_k, состоящему во всех случаях (т. е. при любом исходе β) примерно из одного и того же числа отдельных испытаний α. Вспомним, например, задачу 25 из § 1, в которой требовалось с помощью наименьшего числа вопросов установить, какое из чисел от 1 до 10 загадало некоторое лицо. В решении этой задачи предлагалось выяснить прежде всего, не превосходит ли загаданное число x числа 5 (испытание α_1); затем, в зависимости от исхода α_1, рекомендовалось установить, не больше ли x чем 7 или чем 3 (испытание α_2); далее, учитывая исход α_2, можно было спросить, не больше ли x, чем число 8, или чем 6, или чем 4, или чем 1 (испытание α_3); наконец, если три испытания α_1, α_2 и α_3 не приводили к выяснению значения x, следовало

---

[1]) Приведем один поучительный пример, иллюстрирующий осложнения, которые могут встретиться при осуществлении этой программы в тех случаях, когда $H_\beta(\alpha) \neq 0$ и испытания α не направлены целиком к выяснению исхода опыта β. Пусть нам требуется при помощи взвешиваний на чашечных весах без гирь определить, является ли имеющаяся среди данных четырех монет одна фальшивая монета более легкой или более тяжелой, чем остальные (но не требуется найти эту фальшивую монету!). Ясно, что здесь любое взвешивание α_1 будет содержать н у л е в у ю информацию относительно интересующего нас опыта β (так как при любом исходе опыта α_1 вероятности того, что фальшивая монета легче и что она тяжелее настоящих никак не изменятся), т. е. любой выбор α_1 приводит к одному и тому же на первый взгляд малоутешительному результату. Однако на самом деле то обстоятельство, что всегда $I(\alpha_1, \beta) = 0$, вовсе не означает, что вспомогательные опыты α вообще бесполезны: опыт α_1 не доставляет непосредственно никакой информации об β, но он зато увеличивает пригодность последующих испытаний для этой цели. В самом деле, легко видеть, что, положив на чашки весов по одной или по две монеты (т. е. выбрав опыт α_1 произвольно), мы сразу же придем к положению, когда при помощи еще одного взвешивания (опыт α_2) уже можно однозначно определить исход β.

еще спросить, не больше ли x чем 9 или чем 2 (испытание α_4). Во всех случаях для обнаружения числа x здесь приходится использовать не более четырех вопросов; при этом, если x равно одному из чисел 2, 3, 9 или 10, то число вопросов будет точно равно 4, а в остальных шести случаях оно будет равно 3. Ясно, что если бы мы с самого начала спросили, не равно ли число x, скажем, 10, то мы имели бы определенные шансы обойтись о д н и м е д и н с т в е н н ы м вопросом; однако в большинстве случаев нам пришлось бы затратить больше четырех вопросов, что делает такой путь выяснения исхода β менее выгодным.

Заметим теперь, что если бы мы начали с вопроса о том, не превосходит ли загаданное число x числа 8, то мы имели бы шансы обнаружить x при помощи всего двух вопросов (если это число x равнялось 9 или 10), и в то же время нам ни в каком случае не потребовалось бы использовать более четырех вопросов (ибо если после первого вопроса окажется, что число x не превосходит $2^3 = 8$, то далее мы можем обнаружить его при помощи трех вопросов; см. решение задачи 25). Таким образом, на первый взгляд представляется, что такой путь обнаружения загаданного числа x даже более удобен, чем предложенный в § 1. Однако это заключение является слишком поспешным. В самом деле, если мы не будем считать длину с а м о й д л и н н о й цепочки испытаний единственным критерием, определяющим ценность какого-либо метода обнаружения x, а будем учитывать также и то, что в некоторых случаях этот метод приводит к цели быстрее, то и в отношении старого метода мы должны будем считаться с тем, что он во многих случаях позволяет найти x с помощью трех, а не четырех вопросов.

Для того чтобы сравнить «выгодность» обоих методов решения задачи 25 при таком новом подходе к ней, предположим, что мы м н о г о к р а т н о повторяем попытку обнаружить загаданное число x, причем вероятность быть загаданным для всех десяти чисел по-прежнему предполагается одинаковой. При первом методе решения задачи нам примерно в $\frac{6}{10} = \frac{3}{5}$ всех случаев придется задать три вопроса и в $\frac{4}{10} = \frac{2}{5}$ всех случаев (когда x равно 2, 3, 9 или 10) — четыре вопроса; таким образом, с р е д-

нее значение числа заданных вопросов здесь будет равно

$$\frac{3}{5} \cdot 3 + \frac{2}{5} \cdot 4 = \frac{17}{5} = 3{,}4.$$

Второй метод решения задачи обеспечивает нахождение x с помощью двух вопросов в $\frac{2}{10} = \frac{1}{5}$ от общего числа всех попыток (когда x равно 9 или 10), зато в остальных $\frac{8}{10} = \frac{4}{5}$ случаев придется задать четыре вопроса; поэтому здесь среднее значение числа заданных вопросов будет равно

$$\frac{1}{5} \cdot 2 + \frac{4}{5} \cdot 4 = \frac{18}{5} = 3{,}6.$$

Таким образом, в среднем второй метод обнаружения x несколько менее выгоден, чем первый. Это обстоятельство имеет общий характер — можно показать, что *каково бы ни было число* n, *не существует метода решения задачи* 25, *который был бы в среднем выгоднее метода, намеченного на стр.* 142—143.

Последнее заключение позволяет подойти по-новому к задачам, рассмотренным в §§ 1 и 2; оно проливает также дополнительный свет на смысл использования при решении этих задач понятий энтропии и информации. Ясно, что применение этих понятий, носящих существенно статистический характер, является вполне уместным лишь в тех случаях, когда и сама решаемая задача имеет статистический характер, т. е. связана с многократным повторением одного и того же испытания. Все дело в том, что именно так можно понимать и приведенные выше задачи 23—29, если интересоваться не точным числом испытаний α, требующихся для однократного выяснения исхода какого-либо опыта β, а средним значением этого числа при многократном повторении указанного опыта. Если при этом еще условиться считать все исходы β равновероятными, то при выборе испытаний $\alpha_1, \alpha_2, \ldots, \alpha_k$ так, чтобы среднее значение их числа было наименьшим, число этих испытаний для всех исходов β оказывается примерно одинаковым; поэтому и наибольшее значение

требуемого числа испытаний здесь будет, вообще говоря, наименьшим возможным.

Попробуем теперь *отказаться от условия равновероятности исходов* β. Для примера вернемся снова к той же задаче 25, но теперь осложним несколько ее формулировку. Предположим, что некто задумал определенное число x, которое может принимать одно из n значений; нам требуется отгадать это число, задавая задумавшему его какие-либо вопросы, на которые он отвечает «*да*» или «*нет*». При этом будем считать, что мы заранее имеем определенную информацию о числе x, заставляющую нас считать, что что n возможных значений этого числа не являются равновероятными, т. е. что одни из них вернее окажутся загаданными, чем другие [1]). Как в этом случае следует задавать вопросы?

Ясно, что если ни одно из n значений x не исключается полностью имеющейся у нас информацией (в противоположном случае следовало бы говорить не о n, а о меньшем числе возможных значений x), то наименьшее число вопросов, которое в о в с е х с л у ч а я х гарантирует нахождение загаданного числа x, по-прежнему определяется неравенствами (1) § 1 (стр. 143), причем вопросы здесь надо задавать так же, как указано выше. Действительно, если бы существовала последовательность из меньшего числа вопросов, позволяющая во всех случаях (т. е. независимо от ответов на эти вопросы) однозначно определить число x, то это противоречило бы результату задачи 25. Отсюда, однако, еще не следует, что всегда целесообразно поступать точно так же, как в случае равновероятности всех значений x; после сказанного выше это должно быть совершенно ясно. Так, например, если имеется весьма большая вероятность того, что загаданное число имеет какое-то определенное значение x_0 (скажем, если эта вероятность равна 0,99 или еще больше), то, разумеется, прежде всего следует спросить, не равно ли x этому числу x_0, несмотря на то, что в случае отрицательного ответа мы потратим один

[1]) Для конкретности можно представить себе, что загаданное число было записано, а отгадывающий подсмотрел эту запись, но не вполне уверен в том, что он увидел (однако, разумеется, строгий смысл этого условия связан с предположением о том, что в процессе м н о г о к р а т н о г о повторения процедуры отгадывания одни числа оказываются загаданными чаще, чем другие).

вопрос с очень малой пользой для себя (множество возможных значений x уменьшится всего лишь на единицу). В общем же случае следует *каждый раз разбивать множество возможных значений x на две такие части, чтобы вероятности того, что загаданное число принадлежит к одной и к другой из этих частей, были возможно более близки*. Такое разбиение обеспечивает наибольшую возможную энтропию опыта α, состоящего в постановке вопроса о том, не принадлежит ли x к одной из этих частей, а следовательно, и наибольшую возможную информацию, содержащуюся в α относительно интересующего нас опыта β. Правда, при этом мы не сможем уже обеспечить минимум наибольшего числа вопросов, которое нам может понадобиться в самом неблагоприятном случае, но зато среднее значение общего числа вопросов здесь будет, вообще говоря, меньше (или во всяком случае — не больше), чем при любой другой постановке вопросов.

Вместо строгого доказательства последнего утверждения мы ограничимся тем, что немного позже проверим его на одном простом частном примере (см. текст, напечатанный мелким шрифтом в конце этого параграфа). Что же касается до самого общего случая, то для него сравнительно легко доказывается лишь, что среднее значение l числа требуемых для определения x вопросов всегда будет не меньше чем $H(\beta)$ (где $H(\beta)$, как обычно,— энтропия интересующего нас опыта β) [1]. Этот результат представляет собой обобщение неравенства $k \geqslant \log n$, относящегося к случаю равновероятности всех возможных значений x; он может быть обоснован при помощи рассуждений, близких к тем, которые привели нас к указанному неравенству. В самом деле, информация, доставляемая ответом на один вопрос, очевидно, во всех случаях не может превосходить одного бита; поэтому, задав k вопросов, мы получаем информацию, не превосходящую k бит. Если теперь мы многократно (скажем, 10 000 раз) будем определять загаданное число, задавая вопросы в

---

[1]) Для случая, когда число n очень велико, а вероятность каждого отдельного значения x мала, можно показать также, что это среднее значение будет весьма близко к $H(\beta)$ (ср. гл. IV).

соответствии с каким-то выбранным нами методом, и при этом вероятности того, что загаданным оказывается то или иное из n чисел, будут иметь заданные значения, то с р е д н я я и н ф о р м а ц и я, получаемая при одном определении числа x, будет равна $H(\beta)$, а общая информация, полученная после 10 000 повторений отгадывания, будет близка к 10 000 $H(\beta)$. При этом число задаваемых вопросов здесь может существенно меняться от случая к случаю в зависимости от того, какое именно число x было загадано (достаточно вспомнить про случай, когда существует определенное число x_0, для которого очень велика вероятность быть загаданным). Однако, по самому определению среднего значения l общее число вопросов, заданное во всех 10 000 опытов по обнаружению x, будет близко к 10 000 l (это означает, что в с р е д н е м одно обнаружение x требует как раз l вопросов). Отсюда можно заключить, что должно выполняться неравенство

$$10\,000\,H(\beta) \leqslant 10\,000\,l,$$

т. е.

$$l \geqslant H(\beta), \qquad (4')$$

—что нам и требовалось доказать. Учитывая большую важность неравенства (4′) для теории передачи сообщений (см. по этому поводу § 2 гл. IV), мы в дальнейшем приведем также совсем другое, очень изящное его доказательство, являющееся более формальным, но идейно более простым (см. заключительную часть параграфа).

Все сказанное относительно задачи 25 может быть без труда перенесено и на задачу 27 (стр. 146). Здесь также можно несколько обобщить условия задачи, считая, что разные монеты имеют различные вероятности оказаться фальшивыми (это можно понимать, например, в том смысле, что внешний вид различных монет внушает подозрения в различной степени). В таком случае наиболее целесообразным будет при каждом взвешивании делить подозрительные монеты на три части так, чтобы в е р о я т н о с т и для фальшивой монеты оказаться в двух равных по числу монет частях, положенных на правую и на левую чашки весов, и в третьей части, отложенной в сторону, были все время в о з м о ж н о б о л е е б л и з к и о д н а к д р у г о й. Правда, при таком образе действий общее число взвешиваний, нужное для определения фальшивой

монеты, может в неудачном случае оказаться и большим того, которое дается неравенствами (2) из § 2 (стр. 147); однако среднее значение требуемого числа взвешиваний при этом будет наименьшим. Можно показать также, что это среднее значение l всегда будет не меньше чем $\frac{H(\beta)}{\log 3}$, где $H(\beta)$ — энтропия опыта, состоящего в определении фальшивой монеты:

$$l > \frac{H(\beta)}{\log 3}. \qquad (4'')$$

(см., в частности, заключительную часть настоящего параграфа); при большом числе монет и малой вероятности того, что каждая из них окажется фальшивой, это среднее значение l всегда будет весьма близко к $\frac{H(\beta)}{\log 3}$.

Приведем теперь простой пример, иллюстрирующий тот факт, что при обнаружении загаданного числа x (не превосходящего некоторого n) выгоднее всего каждый раз разбивать множество n возможных значений x на две части так, чтобы вероятности для x принадлежать к той или к другой части были возможно более близки одна к другой.

Пусть число n возможных значений x равно 4; в этом случае число k, определяемое неравенствами (1) (стр. 143), равно 2. Предположим теперь, что у нас есть основание считать одно значение x_0 числа x более вероятным, чем остальные три x_1, x_2 и x_3; пусть p есть вероятность того, что x равно x_0, а q — вероятности того, что x равно x_i (здесь i — любое из чисел 1, 2, 3; $p > q$, $p + 3q = 1$). В качестве первого вопроса можно спросить, совпадает ли x с одним из чисел x_0 или x_1; можно также сразу задать вопрос о том, не равно ли x числу x_0. Опыты, состоящие в постановке этих двух вопросов, обозначим через $\alpha_1^{(1)}$ и $\alpha_1^{(2)}$; так как исходы опыта $\alpha_1^{(1)}$ имеют вероятности $p + q$ и $2q$, то $H(\alpha_1^{(1)}) = -(p + q)\log(p + q) - 2q\log(2q)$; два же исхода опыта $\alpha_1^{(2)}$ имеют вероятности p и $3q$, так что $H(\alpha_1^{(2)}) = -p\log p - 3q\log(3q)$. Если $p > \frac{1}{2}$, то, разумеется, исходы опыта $\alpha_1^{(2)}$ имеют более близкие вероятности, чем исходы опыта $\alpha_1^{(1)}$; если же $\frac{1}{2} > p > q$, то следует сравнить разности $(p+q) - 2q = p - q$ и $3q - p$ вероятностей двух исходов для опытов $\alpha_1^{(1)}$ и $\alpha_1^{(2)}$. Так как $p - q > 3q - p$, если $p > 2q$, т. е. если $p > \frac{2}{5}$ $\left(\text{ибо } q = \frac{1-p}{3}, \text{ а } p > \frac{2}{3}(1-p) \text{ при } p > \frac{2}{5}\right)$, то мы заключаем, что при $p > \frac{2}{5}$ следует начать с опыта $\alpha_1^{(2)}$, а при

$p < \frac{2}{5}$ с опыта $\alpha_1^{(1)}$; при $p = \frac{2}{5}$, по-видимому, безразлично, с какого из этих двух опытов мы начнем.

Если мы начинаем с вопроса «*Не равно ли x одному из чисел x_0 и x_1?*», то тем самым мы разбиваем множество возможных значений x на две равные по численности части; в таком случае при любом ответе на первый вопрос мы находим x с помощью двух вопросов. Если же мы начинаем с вопроса «*Не равно ли x числу x_0?*», то мы имеем определенные шансы обнаружить x одним вопросом; вероятность, что это будет именно так, равна вероятности того, что x совпадает с x_0, т. е. равна p. Однако если x не равно x_0, то мы уже не можем гарантировать возможность обнаружить x следующим вопросом; на вопрос «*Не равно ли x числу x_1?*» может последовать положительный ответ (вероятность этого равна q), но может последовать и отрицательный ответ (вероятность этого равна вероятности того, что x совпадает с x_2 или с x_3, т. е. равна $2q$),—и в этом последнем случае нам понадобится еще один, третий вопрос. Таким образом, в том случае, когда мы начинаем с опыта $\alpha_1^{(2)}$, мы имеем вероятность p определить x одним вопросом, вероятность q того, что нам потребуется два вопроса, и вероятность $2q$ того, что надо будет задать три вопроса. Отсюда видно, что среднее значение числа вопросов здесь равно

$$p \cdot 1 + q \cdot 2 + 2q \cdot 3 = p + 8q.$$

Нетрудно проверить, что $p + 8q < 2$, если $p > \frac{2}{5}$ $\left(\text{ибо } p + 8q = \frac{8 - 5p}{3}, \text{ поскольку } q = \frac{1 - p}{3}\right)$; таким образом, мы убеждаемся, что действительно с опыта $\alpha_1^{(2)}$ целесообразно начинать в том случае, когда $p > \frac{2}{5}$.

В заключение параграфа приведем еще строгое доказательство неравенств (4′) и (4″), не использующее никаких результатов гл. II, кроме определения энтропии опыта. При этом нам понадобится следующий факт. Пусть $p_1, p_2, \ldots, p_n$ — *какие-то n положительных чисел, сумма которых равна 1, а $q_1, q_2, \ldots, q_n$ — какие угодно другие n положительных чисел, сумма которых не превосходит 1; в таком случае всегда*

$$-p_1 \log p_1 - p_2 \log p_2 - \ldots - p_n \log p_n \leqslant -p_1 \log q_1 - \\ - p_2 \log q_2 - \ldots - p_n \log q_n. \quad (*)$$

Полное доказательство неравенства (*) мы отложим до Приложения I в конце книги (см. стр. 454); здесь же заметим лишь, что при $n = 2$, $p_1 = p_2 = \frac{1}{2}$, $q_1 + q_2 = 1$, это неравенство принимает вид

$$-\frac{1}{2} \log q_1 - \frac{1}{2} \log q_2 \geqslant 1,$$

или иначе

$$\frac{1}{2}\log q_1+\frac{1}{2}\log q_2\leqslant -1=\log\frac{1}{2},\ \text{т. е. } q_1q_2\leqslant\frac{1}{2}=\frac{q_1+q_2}{2}.$$

Таким образом, если $p_1=p_2=\frac{1}{2}$ и $q_1+q_2=1$, то оно сводится к известному неравенству между средним арифметическим и средним геометрическим двух чисел.

Вернемся теперь снова к опыту β с n исходами $B_1, B_2, \ldots, B_n$ и таблицей вероятностей

исходы опыта	B_1	B_2	...	B_n
вероятности	p_1	p_2	...	p_n

Пусть для выяснения того, какой из исходов β осуществился на самом деле, производятся последовательные испытания (вспомогательные опыты) α, каждое из которых может иметь m различных исходов; наибольшее число испытаний, которое может потребоваться для определения исхода β, мы, как и раньше, обозначим через k. Пусть далее n_1 — число тех исходов β, которые могут быть обнаружены при помощи одного испытания α_1, n_2 — число исходов β, которые могут быть обнаружены при помощи двух испытаний α_1 и α_2, ..., наконец, n_k — число исходов β, которые могут быть обнаружены лишь при помощи k испытаний $\alpha_1, \alpha_2, \ldots, \alpha_k$; очевидно, что $n_1+n_2+\ldots+n_k=n$.

Заметим, что число n исходов β, которые могут быть обнаружены с помощью одного испытания α_1, очевидно, не превосходит числа m исходов α_1:

$$n_1\leqslant m;$$

при этом $n_1=m$ лишь в том (разумеется, мало интересном) случае, когда $n=m$ и каждому исходу испытания α_1 отвечает единственный исход β (например, когда в условии задачи 25 число возможных значений загаданного числа равно 2). Если же существуют такие исходы α_1, которые не определяют однозначно исход β, т. е. если имеются случаи, когда оказывается необходимым произвести последующее испытание α_2, то, наверное, $n_1<m$. При этом число исходов опыта α_1, не определяющих однозначно исход β, будет равно $m-n_1$; так как число исходов опыта α_2 равно m, то число n_2 тех исходов β, которые могут быть обнаружены с помощью двух испытаний α_1 и α_2, наверное удовлетворяет неравенству

$$n_2\leqslant(m-n_1)\cdot m=m^2-n_1m.$$

Аналогично этому, если в некоторых случаях требуется произвести еще и третий вспомогательный опыт α_3, то $n_2<(m-n_1)m$, причем не более чем при $(m-n_1)m-n_2$ исходах опыта α_2 требуется произвести опыт α_3. Так как к тому же сам опыт α_3 имеет всего m разных исходов, то очевидно

$$n_3\leqslant[(m-n_1)m-n_2]\cdot m=m^3-n_1m^2-n_2m.$$

Точно также показывается, что

$$n_4\leqslant[(m^3-n_1m^2-n_2m)-n_3]m=m^4-n_1m^3-n_2m^2-n_3m$$

и т. д.; наконец, для числа n_k исходов β, обнаружение которых требует ровно k испытаний, по индукции легко получаем

$$n_k \leqslant [(m^{k-1} - n_1 m^{k-2} - n_2 m^{k-3} - \ldots - n_{k-2} m) - n_{k-1}]\, m =$$
$$= m^k - n_1 m^{k-1} - n_2 m^{k-2} - \ldots - n_{k-2} m^2 - n_{k-1} m.$$

Перенесем здесь все члены правой части, кроме первого члена m^k, влево и разделим обе части полученного неравенства на m^k; тогда будем иметь:

$$\frac{n_k}{m^k} + \frac{n_{k-1}}{m^{k-1}} + \ldots + \frac{n_2}{m^2} + \frac{n_1}{m} \leqslant 1.$$

Обозначим через l_i (где $i = 1, 2, \ldots, n$) число испытаний α, которые приходится произвести для обнаружения исхода β в том случае, когда оказывается, что этим исходом является исход B_i. В таком случае n_1 из n чисел l_i будут равны 1, n_2 из этих чисел будут равны 2, . . ., n_k из этих чисел будут равны k. Поэтому последнее неравенство можно переписать также в следующем виде:

$$\frac{1}{m^{l_1}} + \frac{1}{m^{l_2}} + \ldots + \frac{1}{m^{l_n}} \leqslant 1.$$

Напомним теперь, что для справедливости выписанного выше неравенства (*) надо лишь, чтобы сумма всех чисел p_i была равна 1, а сумма всех чисел q_i ($i = 1, 2, \ldots, n$) не превосходила 1. Поэтому мы можем положить в этом неравенстве, в частности, p_i равным вероятности i-го исхода B_i опыта β, а q_i равным $\frac{1}{m^{l_i}}$, так что

$$-p_1 \log p_1 - p_2 \log p_2 - \ldots - p_n \log p_n \leqslant$$
$$\leqslant -p_1 \log \frac{1}{m^{l_1}} - p_2 \log \frac{1}{m^{l_2}} - \ldots - p_n \log \frac{1}{m^{l_n}}.$$

В левой части последнего неравенства, очевидно, стоит энтропия $H(\beta)$ опыта β. Заменив теперь в правой части $-\log \frac{1}{m^{l_i}}$ (где i равно 1, 2, . . ., n) на $l_i \cdot \log m$, получим

$$H(\beta) \leqslant [p_1 l_1 + p_2 l_2 + \ldots + p_n l_n] \log m.$$

Но по самому определению среднего значения (см. стр. 24) сумма $p_1 l_1 + p_2 l_2 + \ldots + p_n l_n$ равна как раз среднему значению l числа требуемых испытаний α. Таким образом, мы получаем основное неравенство

$$l \geqslant \frac{H(\beta)}{\log m}. \qquad (4)$$

Это и есть тот результат, который мы хотели доказать; при $m = 2$ (например, в случае, когда опыты α — это вопросы, на которые отвечают лишь «*да*» или «*нет*») он переходит в неравенство (4′) (ибо $\log 2 = 1$), а при $m = 3$ (например, в случае, когда α — это взвешивания на чашечных весах без гирь) — в неравенство (4″).

Глава IV

ПРИЛОЖЕНИЕ ТЕОРИИ ИНФОРМАЦИИ К ВОПРОСУ О ПЕРЕДАЧЕ СООБЩЕНИЙ ПО ЛИНИЯМ СВЯЗИ

§ 1. Основные понятия. Экономность кода

Для того чтобы проиллюстрировать пользу введенных в гл. II понятий энтропии и информации, мы разобрали в главе III ряд «занимательных задач» типа тех, которые обычно рассматриваются в школьных математических кружках. В настоящей главе мы рассмотрим некоторые простейшие, но сами по себе достаточно серьезные приложения тех же понятий к практическому вопросу о передаче сообщений по линиям связи. При этом окажется, что применения эти имеют очень много общего с рассмотренными выше «игрушечными задачами» об отгадывании задуманного числа или об определении фальшивой монеты с помощью взвешиваний, так что ряд приведенных в предыдущих параграфах рассуждений может быть непосредственно перенесен на решение практических вопросов техники связи.

Рассмотрим прежде всего общую схему передачи сообщений по линиям связи; для определенности будем говорить, например, о телеграфии. На одном конце линии отправитель подает некоторое *сообщение*, записанное при помощи 33 букв русского алфавита (исключая букву *ё*, но включая сюда и «нулевую букву» — пустой промежуток между словами), или 27 букв латинского алфавита, или при помощи 10 цифр (числовое сообщение), или при помощи букв и цифр вместе взятых. Для передачи этого сообщения в случае обычного проводного телеграфа используется постоянный ток, некоторые характеристики которого телеграфист может менять по своему усмотрению; при этом он создает определенную последовательность *сигналов*, воспринимаемых вторым телеграфистом на приемном конце линии. Простейшими различимыми сигналами, широко используемыми на практике, являются посылка тока (т. е. включение его на некоторое вполне определенное время) и отсутствие посылки — пауза

(выключение тока на то же время); при помощи одних только этих двух сигналов уже можно передать любое сообщение, если условиться заменять каждую букву или цифру определенной комбинацией посылок тока и пауз.

В технике связи правило, сопоставляющее каждому передаваемому сообщению некоторую комбинацию сигналов, обычно называется *кодом* (в случае телеграфии, например, телеграфным кодом), а сама операция перевода сообщения в последовательность различимых сигналов — *кодированием* сообщения. При этом коды, использующие только два различных элементарных сигнала (например, посылку тока и паузу), называются *двоичными кодами*; коды, использующие три различных элементарных сигнала — *троичными кодами* и т. д. В телеграфии, в частности, применяется целый ряд различных кодов, важнейшими из которых являются *код Морзе* («азбука Морзе») и *код Бодо*. В коде Морзе каждой букве или цифре сообщения сопоставляется некоторая последовательность кратковременных посылок тока («точек») и в три раза более длинных посылок тока («тире»), разделяемых кратковременными паузами той же длительности, что и «точки»; пробел между буквами (или цифрами) при этом отмечается специальным разделительным знаком — длинной паузой (той же длительности, что и «тире»), а пробел между словами — еще в 2 раза более длинной паузой. Хотя этот код использует лишь посылки тока и паузы, его можно считать троичным, так как каждое закодированное сообщение здесь естественно разлагается в совокупность следующих трех сравнительно крупных «элементарных сигналов» — точек, к каждой из которых добавляется всегда следующая за точкой кратковременная пауза, тире со следующей за каждым тире кратковременной паузой и длинных пауз, разделяющих отдельные буквы. В настоящее время код Морзе обычно используется лишь при повреждении основных телеграфных линий, а также в коротковолновой радиотелеграфии, имеющей многие важные применения. В обычных же буквопечатающих телеграфных аппаратах, стоящих на всех больших телеграфных линиях, чаще всего применяется двоичный код Бодо, сопоставляющий каждой букве некоторую последовательность из пяти простейших элементарных сигналов — посылок тока и пауз одинаковой

длительности. Так как при этом все буквы передаются комбинациями сигналов одной и той же длительности (коды, обладающие этим последним свойством, называются р а в н о м е р н ы м и), то в коде Бодо не требуется специального знака, отделяющего одну букву от другой — и без того известно, что через каждые пять элементарных сигналов кончается одна буква и начинается следующая (в приемных аппаратах такое разделение на комбинации из пяти сигналов обычно производится автоматически). Поскольку комбинируя две возможности для первого сигнала с двумя возможностями для второго, двумя — для третьего, двумя — для четвертого и двумя — для пятого, мы можем составить всего $2^5 = 32$ различных комбинаций, то код Бодо в его простейшей форме позволяет передавать 32 различные буквы [1]).

В некоторых телеграфных аппаратах кроме простого включения и выключения тока можно также изменять его направление на обратное; при этом появляется возможность вместо посылок тока и пауз использовать в качестве основных сигналов посылки тока в двух различных направлениях или же использовать сразу три различных элементарных сигнала одной и той же длительности — посылку тока в одном направлении, посылку в другом направлении и паузу. Возможны также еще более сложные телеграфные аппараты, в которых посылки тока различаются не только по направлению, но и по силе тока; тем самым мы получаем возможность сделать число различных элементарных сигналов еще бо́льшим. Увеличение числа разных элементарных сигналов позволяет сделать код более сжатым (т. е. уменьшить число элементарных сигналов, требующихся для передачи данного сообщения или же передавать при помощи сигналов той же длительности значительно больше различных «букв»). Однако вместе с тем оно усложняет и удорожает систему передачи, так что в

[1]) Так как 32 комбинаций для передачи всех букв и цифр оказывается недостаточно, то в аппаратах, работающих на коде Бодо, имеются два регистра; после перевода регистра та же комбинация используется для передачи еще одного знака. При этом число возможностей почти удваивается, что позволяет передавать все буквы, цифры и знаки препинания. В случае одного регистра такие же возможности предоставляют коды, сопоставляющие каждой букве или цифре комбинацию ш е с т и элементарных сигналов; подобные коды также иногда используются в телеграфии.

технике все же предпочтительно используются коды с малым числом элементарных сигналов.

В радиотелеграфе вместо изменений силы тока изменениям подвергаются некоторые параметры радиоволны — синусоидального колебания высокой частоты,— т. е. элементарные сигналы здесь имеют другой смысл; однако и в этом случае каждая передаваемая буква заменяется некоторой последовательностью элементарных сигналов, воспринимаемых на приемном конце линии. Аналогично обстоит дело и в большинстве других линий связи; подробнее об этом мы еще будем говорить ниже (см. §§ 3 и 4).

Отвлечемся теперь от технических подробностей и сформулируем основную математическую задачу, с которой приходится иметь дело в технике связи. Пусть имеется сообщение, записанное при помощи некоторого «алфавита», содержащего n «букв» (например, 33 русские буквы, или 10 цифр, или 43 буквы и цифры, или буквы, цифры и знаки препинания и т. д.). Требуется «закодировать» это сообщение, т. е. указать правило, сопоставляющее каждому такому сообщению определенную последовательность из m различных «элементарных сигналов», составляющих «алфавит» передачи. Как выгоднее всего это сделать?

Прежде всего надо объяснить, в каком смысле здесь понимается слово «выгоднее». *Мы будем считать кодирование тем более выгодным, чем меньше элементарных сигналов приходится затратить на передачу сообщения.* Если считать, что каждый из элементарных сигналов продолжается одно и то же время, то наиболее выгодный код позволит затратить на передачу сообщения меньше всего времени. Так как сооружение и содержание линии связи обычно обходится очень дорого (а в случае радиосвязи, где дело обстоит несколько иначе, чрезмерное увеличение числа линий связи является невозможным, поскольку при этом такие линии начинают мешать друг другу), то переход к более выгодному коду, позволяющий увеличить эффективность использования данной линии связи, имеет несомненное практическое значение.

Постараемся теперь несколько подробнее разобраться в том, какие вообще бывают коды. Будем для определенности пока считать, что $m = 2$ (т. е. что наш код — двоичный). Кроме того ограничимся лишь случаем побуквенного кодирования, т. е. случаем кодов, приспособленных

для передачи каждой буквы сообщения по отдельности (о возможностях, доставляемых отказом от этого последнего условия, будет говориться позже). В таком случае кодирование, очевидно, состоит в том, что каждой из n «букв» нашего «алфавита» сопоставляется какая-то последовательность двух элементарных сигналов — *кодовое обозначение* соответствующей «буквы». Отвлекаясь от физической природы используемых элементарных сигналов, мы можем заменить их цифрами 0 и 1, т. е. рассматривать все кодовые обозначения как некоторые последовательности этих двух цифр. Для задания кода надо перечислить n таких последовательностей, которые сопоставляются n имеющимся «буквам». При этом не всякие n различных последовательностей цифр 0 и 1 определяют пригодный для практического использования двоичный код; требуется еще, чтобы закодированное сообщение можно было *однозначно декодировать*, т. е. чтобы в длинной последовательности цифр 0 и 1, сопоставляемой многобуквенному сообщению, всегда можно было понять, где кончается кодовое обозначение одной буквы и начинается обозначение следующей. Проще всего добиться этого, если, как в коде Морзе, ввести специальный разделительный знак (в технической литературе такой знак иногда называют «запятой»), отличающийся от всех других кодовых обозначений и легко различимый, и передавать этот знак между кодовыми обозначениями каждых двух «букв». Ясно, однако, что этот путь вряд ли может быть выгодным, так как здесь число «букв» в передаваемом сообщении практически удваивается (за счет добавления $(n + 1)$-й разделительной «буквы», вставляемой между каждыми двумя другими буквами); поэтому ниже мы будем интересоваться лишь однозначно декодируемыми кодами без разделительного знака (т. е. «кодами без запятой»). Примерами таких кодов являются, в частности, те, в которых кодовые обозначения всех букв имеют одну и ту же длину (т. е. равномерные коды; ср. выше описание кода Бодо). Кроме того, существуют также и многие неравномерные коды (содержащие кодовые обозначения различной длины), которые могут быть однозначно декодированы и поэтому не требуют разделительного знака. Так, например, в случае двухбуквенного алфавита (при $n = 2$) простейшим кодом без запятой является равномерный код

с кодовыми обозначениями 0 и 1; если, однако, мы заменим кодовое обозначение 1 совокупностью двух цифр 11, или 10, или 01 (но, разумеется, не 00), то такой код все равно будет без труда однозначно декодироваться (во всех этих случаях кодовые обозначения второй буквы легко идентифицируются в длинной последовательности кодовых обозначений обоих типов по входящей в них цифре 1). Общее необходимое и достаточное условие, выделяющее однозначно декодируемые коды среди всех других совокупностей n последовательностей цифр 0 и 1, может быть найдено в статье А. Сардинаса и Дж. Паттерсона [61] (см. в этой связи также работу [62], посвященную общей теории двоичных неравномерных кодов). Для нас здесь, однако, достаточно будет лишь отметить, что неравномерный код наверное может быть однозначно декодирован, если *никакое кодовое обозначение не совпадает с началом какого-либо другого более длинного кодового обозначения* (так что, например, если «101» — это кодовое обозначение какой-то буквы, то уже не может быть букв, имеющих обозначение «1», «10» или же «1011О»). В самом деле, если это условие выполняется, то, читая подряд кодовую запись сообщения и имея перед собой список всех кодовых обозначений, всегда можно точно сказать, в каком месте кончается обозначение одной буквы и начинается обозначение следующей (так как здесь последовательность элементарных сигналов, начинающаяся вслед за окончанием очередного кодового обозначения, сама будет образовывать кодовое обозначение только в случае, если мы оборвем ее в одном-единственном строго определенном месте) [1]). Заметим еще, что равномерный код также, ра-

[1]) Коды, удовлетворяющие указанному условию, иногда называются мгновенными (или мгновенно декодируемыми), поскольку в случае других однозначно декодируемых кодов для того, чтобы установить, что мы дошли до конца очередного кодового обозначения, иногда (или даже всегда) приходится ознакомиться и с несколькими последующими элементарными сигналами (т. е. декодирование осуществляется с запаздыванием по сравнению с передачей сообщения). В рассмотренных выше трех примерах неравномерного кода для двухбуквенного алфавита с кодовыми обозначениями 0 и 11, или 0 и 10, или 0 и 01, первые два, очевидно, являются примерами мгновенных кодов, а третий — нет (в этом третьем случае для выяснения смысла цифры 0 в длинной последовательности цифр 0 и 1, образующей закодированное сообщение, необходимо знать также и следующую цифру).

зумеется, удовлетворяет напечатанному курсивом условию. Коды же, не удовлетворяющие этому условию, мы, как правило, вообще не будем рассматривать; поэтому в дальнейшем всюду, где не оговорено обратное, *под «кодом» будет пониматься такая совокупность n кодовых обозначений, сопоставляемых n буквам алфавита, для которой выполняется указанное выше условие.*

Перейдем теперь к вопросу о связи двоичного кодирования с условиями задачи 25 об отгадывании загаданного числа, не превосходящего n, при помощи вопросов, на которые отвечают только «*да*» или «*нет*». Связь эта является самой непосредственной. В самом деле, пусть мы имеем некоторый двоичный код; будем считать, что n «букв», которым сопоставляются наши кодовые обозначения, это всевозможные числа от 1 до n. Пусть нам надо отгадать какое-то загаданное число. В качестве первого вопроса спросим: «Является ли первой цифрой кодового обозначения задуманного числа цифра 1?», в качестве второго — спросим: «Является ли второй цифрой этого кодового обозначения цифра 1?» и т. д. При этом мы последовательно определим все цифры кодового обозначения задуманного числа: поскольку никакое из этих обозначений не совпадает с началом другого из них, то как только мы придем к комбинации цифр, являющейся одним из используемых кодовых обозначений, мы с полной уверенностью сможем остановиться и назвать загаданное число. Таким образом, *каждому двоичному коду для n-буквенного алфавита отвечает некоторый метод обнаружения одного из n задуманных чисел при помощи вопросов, на которые отвечают только «да» и «нет».* Обратно, *любой метод обнаружения загаданного числа* позволяет сопоставить каждому из n чисел последовательность цифр 1 и 0, где первая цифра показывает, будет ли в случае, когда загадывается данное число, ответ на первый вопрос гласить «*да*» или «*нет*», вторая цифра точно так же указывает ответ на второй вопрос, третья цифра — ответ на третий вопрос и т. д., т. е. *приводит к двоичному коду.* Сформулированное выше условие здесь, очевидно, всегда выполняется, так как из того, что наш метод позволяет по ответам на поставленные вопросы однозначно указать загаданное число, сразу следует, что никакое из полученных кодовых обозначений не может являться продолжением другого обозначения

(например, наличие среди кодовых обозначений последовательности «101» означает, что ответы «*да*», «*нет*» и «*да*» уже полностью определяют число, и исключает возможность существования обозначения «10110»).

Итак, мы видим, что возможные двоичные коды для n-буквенного алфавита точно соответствуют всевозможным методам определения одного из n задуманных чисел при помощи вопросов, на которые отвечают только «*да*» или «*нет*». Теперь нетрудно уже понять, какой код будет наиболее выгодным. Будем пока измерять в ы г о д н о с т ь (или, лучше сказать, э к о н о м н о с т ь) данного двоичного кода при помощи максимального числа элементарных сигналов (иначе — цифр 1 и 0), требующегося для передачи (или записи) одной буквы: чем меньше это максимальное число, тем более экономен наш код (более точное определение «степени экономности» кода, исходящее из подсчета с р е д н е г о ч и с л а элементарных сигналов, приходящихся на одну букву, будет рассмотрено в следующем параграфе). В таком случае вопрос о построении наиболее экономного кода будет совпадать с содержанием задачи 25. Согласно решению этой задачи наибольшее число k элементарных сигналов, приходящееся на одну букву, не может быть меньше, чем $\log n$, т. е. в лучшем случае оно определяется неравенствами (1) на стр. 143. Тот факт, что всегда $k \geqslant \log n$ легко объясняется соображениями теории информации: одна буква n-буквенного алфавита может содержать информацию, равную $\log n$ (для этого надо только, чтобы все «буквы» сообщения были независимыми друг от друга и каждая из них могла принимать все значения с одинаковой вероятностью), а каждый передаваемый элементарный сигнал, принимающий одно из двух значений (например, являющийся или посылкой тока, или паузой), может содержать информацию, не большую, чем 1 бит; поэтому для передачи одной буквы надо не меньше чем $\log n$ элементарных сигналов.

Для построения н а и б о л е е э к о н о м н о г о двоичного кода мы можем воспользоваться решением задачи 25. А именно, разобьем наши n «букв» на две возможно более близкие по своей численности группы и для всех букв первой группы примем за первую цифру кодового обозначения цифру 1, а для всех букв второй группы — цифру 0; далее, каждую из этих двух групп снова разо-

бьем на две возможно более близкие по численности группы и примем за вторую цифру кодового обозначения цифру 1, если соответствующая буква входит в первую из двух полученных более мелких групп, и цифру 0, если она входит во вторую из этих групп; затем разобьем каждую из четырех уже имеющихся групп на две еще более мелкие группы по возможности близкой численности и в зависимости от этого разбиения выберем третью цифру кодового обозначения и т. д. Согласно сказанному в § 1 гл. III при этом мы придем к двоичному коду, для которого максимальное число k цифр в одном кодовом обозначении определяется неравенствами (1) на стр. 143, так что никакой код не может быть более экономным, чем этот.

Разумеется, это еще не означает, что не существует также и других столь же экономных кодов, т. е. что наиболее экономный код может быть только один. В частности ясно, что, оценивая экономность кода количеством цифр 0 и 1 в наиболее длинном кодовом обозначении, мы можем вовсе не рассматривать неравномерных кодов; добавив, например, в каждом из них в конце кодовых обозначений, длина которых меньше максимальной, некоторое число произвольно выбранных цифр (например, одних лишь цифр 0), мы придем к равномерному коду, имеющему ту же максимальную длину кодового обозначения, что и исходный неравномерный код. Это обстоятельство существенно для приложений, так как равномерные коды имеют заметные практические преимущества: они значительно проще декодируются, причем декодирование здесь легко может быть автоматизировано. Отметим еще, что и равномерных кодов с наименьшей возможной длиной кодовых обозначений может быть несколько. В связи с их большой практической важностью мы опишем здесь еще один метод построения такого кода, по существу довольно близкий к описанному выше.

Метод, о котором пойдет речь, связан с использованием двоичной системы счисления. Обычно мы пользуемся десятичной системой счисления, в которой каждое число представляется в виде суммы степеней числа 10:

$$n = a_k \cdot 10^k + a_{k-1} \cdot 10^{k-1} + \ldots + a_1 \cdot 10 + a_0,$$

где a_k, a_{k-1}, ..., a_1, a_0 — *цифры* числа, могущие прини-

мать значения от 0 до 9; число n при этом обозначается последовательностью своих цифр, т. е. как $a_k a_{k-1} \dots a_1 a_0$. Аналогично этому число n можно представить и в виде суммы степеней числа 2:

$$n = b_l \cdot 2^l + b_{l-1} \cdot 2^{l-1} + \dots + b_1 \cdot 2 + b_0;$$

здесь «цифры» b_l, b_{l-1}, ..., b_1, b_0 уже должны быть все меньше 2, т. е. могут принимать лишь значения 1 и 0. В двоичной системе счисления число обозначается последовательностью соответствующих «двоичных цифр»; так, например, поскольку

$$6 = 1 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0, \quad 9 = 1 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0,$$

то в двоичной системе счисления числа 6 и 9 будут обозначаться, соответственно, как 110 и 1001. Можно, разумеется, представлять числа и в виде суммы степеней любого другого числа m; при этом мы придем к m-и ч н о й с и с т е м е с ч и с л е н и я, в которой «цифры» могут принимать m значений 0, 1, 2, ..., $m - 1$ (такие системы нам еще понадобятся ниже).

Число k цифр в обычной («десятичной») записи числа n, очевидно, определяется неравенствами

$$10^{k-1} \leqslant n < 10^k;$$

так числа в промежутке между $10^1 = 10$ и $10^2 - 1 = 99$ будут двузначными, числа между $10^2 = 100$ и $10^3 - 1 = 999$ — трехзначными и т. д. Аналогично этому в двоичной системе счисления число k «цифр» в записи числа n определяется неравенствами

$$2^{k-1} \leqslant n < 2^k$$

(отсюда, в частности, сразу следует, что число 6 — «трехзначное», а 9 — «четырехзначное»). Поэтому, если мы выпишем первые n целых чисел, начиная с числа 0 (т. е. числа 0, 1, 2, ..., $n - 1$), то окажется, что при

$$2^{k-1} < n \leqslant 2^k$$

двоичная запись всех этих чисел содержит не более k знаков, причем точно k знаков нам, наверное, хоть раз понадобится. Добавив теперь в начало двоичной записи всех

менее чем k-значных чисел некоторое число нулей, мы придем к равномерному двоичному коду для n-буквенного алфавита с минимальной возможной длиной кодовых обозначений. Так, например, при $n = 10$ соответствующими кодовыми обозначениями будут следующие комбинации, представляющие собой запись в двоичной системе счисления всех чисел от 0 до 9, дополненную, если надо, до четырех знаков нулями в начале: 0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111, 1000, 1001. Столь же просто строятся по этому методу все кодовые обозначения и при любом другом n; никакого предварительного разбиения совокупности n чисел на мелкие группы здесь не требуется [1]).

Выше было показано, что в случае n-буквенного алфавита длина кодовых обозначений (т. е. число входящих в них элементарных сигналов) для самого экономного равномерного двоичного кода равна наименьшему целому числу k, удовлетворяющему неравенству $k \geqslant \log n$. Отметим теперь, что если число $\log n$ не является целым, то кодовые обозначения такой длины могли бы быть, вообще говоря, использованы для передачи *бо́льшего количества информации*, чем то, которое реально передается в случае кодирования сообщений, записанных с помощью n-буквенного алфавита. Рассмотрим для примера случай $n = 10$ (скажем, случай передачи числового сообщения). Каждая цифра передаваемого сообщения (записанного в обычной десятичной системе счисления) может принимать одно из десяти значений, т. е. может содержать информацию, равную самое большее $\log 10 \approx \approx 3^1/_3$ битам — такое значение информации будет достигаться в случае, когда все цифры сообщения являются независимыми друг от друга и каждая из них может принимать все значения с одинаковой вероятностью. Каждая цифра закодированного сообщения (т. е. каждый передаваемый элементарный сигнал — например, посылка

[1]) Нетрудно видеть, что в случае, когда n является целой степенью числа 2 (скажем, при $n = 8$, $n = 16$, или $n = 32$), код, получаемый с помощью двоичной системы счисления, точно совпадает с тем, который задается решением задачи 25. [При $n = 10$ «двоичный код» приводит к решению задачи 25, начинающемуся с вопроса «*не превосходит ли загаданное число восьми?*»; ср. выше, стр. 174.]

тока или пауза) может принимать одно из двух значений, т. е. может содержать информацию, самое большее равную 1 дв. ед. (одному биту). Но при использовании равномерного двоичного кода мы затрачиваем на передачу одной цифры сообщения 4 элементарных сигнала, а на передачу сообщения из N цифр — $4N$ элементарных сигналов. Между тем при помощи $4N$ двоичных сигналов мы могли бы передать информацию, равную $4N$ битам, т. е. информацию, примерно на $\frac{2}{3}N$ бит бо́льшую, чем наибольшая информация, которая только может содержаться в числе из N цифр (равная N десятичным единицам информации).

Нетрудно понять, чем это объясняется. Дело в том, что при $n = 10$ в закодированном сообщении все знаки никогда н е б у д у т взаимно независимыми и принимающими оба возможных значения с одинаковой вероятностью: эти условия могут выполняться лишь при $n = 2^k$. В частности, если пользоваться кодом, построенным с помощью разложения чисел от 0 до 9 по двоичной системе счисления, то в случае, когда в исходном сообщении все цифры встречаются одинаково часто, в закодированном сообщении цифра 0 будет встречаться в $\frac{25}{15} = \frac{5}{3}$ раза чаще чем цифра 1 (так как легко проверить, что в выписанных на предыдущей странице десяти кодовых обозначениях цифра 0 встречается 25 раз, а цифра 1 — лишь 15 раз). Между тем для того, чтобы последовательность из данного числа цифр 0 и 1 содержала наибольшую информацию, требуется, чтобы все цифры этой последовательности принимали оба значения с о д и н а к о в о й вероятностью (и были взаимно независимы).

Для передачи длинных числовых сообщений можно построить, однако, и более выгодный двоичный код. Для этого надо только отказаться от побуквенного кодирования (под «буквами», из которых состоят наши сообщения, разумеется, понимаются цифры 0, 1, ..., 9), а вместо этого использовать так называемые б л о к о в ы е к о д ы, в которых кодовые обозначения сопоставляются «блокам», состоящим из фиксированного числа последовательных «букв». Начнем со случая простейших блоков из двух «букв», т. е. разобьем наше сообщение на последователь-

ные п а р ы ц и ф р [1]) и будем переводить в двоичную систему счисления не отдельно каждую цифру, а каждое из двузначных чисел, полученных при таком разбиении. Число двоичных знаков, требуемых для записи всех двузначных чисел (от 00 до 99 включительно), равно числу вопросов, нужных для отгадывания задуманного числа в пределах первой сотни, т. е. равно 7 (см. задачу 25, стр. 143). Таким образом, при такой системе кодирования на две цифры сообщения тратится 7 элементарных сигналов (а не $2 \cdot 4 = 8$, как раньше), т. е. для передачи числа из N цифр (будем для простоты считать N четным) надо передать $3{,}5 \cdot N$ элементарных сигналов — на $\frac{1}{2}N$ сигналов меньше, чем при первоначальной системе кодирования. При необходимости передавать много цифр (случай большого N) выгода оказывается весьма ощутимой.

Еще выгоднее было бы разбить передаваемое число на блоки из т р е х цифр и переводить в двоичную систему счисления лишь получаемые при этом трехзначные числа. Для передачи трехзначного числа, очевидно, надо затратить 10 элементарных сигналов (ср. стр. 143), так что при таком способе кодирования число из N цифр (в случае N кратного трем) передается при помощи $\frac{10}{3}N = 3\frac{1}{3}N$ элементарных сигналов. Выгода от перехода к разбиению сообщения на еще более крупные блоки и перевода в двоичную систему каждого из этих блоков в отдельности практически оказывается уже совсем небольшой (при переходе от блоков из трех цифр к блокам из четырех цифр экономность кодирования даже уменьшается: на передачу четырех цифр, как легко видеть, требуется $14 = 3{,}5 \cdot 4$ элементарных сигналов). Тем не менее интересно отметить, что, применяя разбиения на достаточно крупные блоки, мы можем еще более «сжать» наш код и *сделать отношение числа элементарных сигналов в закодированном сообщении к числу цифр в исходном* (обыкновенном, т. е. десятичном) *числе сколь угодно близким к предельному значению, равному* $\log 10 = 3{,}32193\ldots$ В самом деле, воспользовавшись разбиением на блоки, скажем, из N цифр, мы

[1]) Такое разбиение сообщения на последовательные пары цифр, очевидно, равносильно переводу его в с т о и ч н у ю систему счисления.

придем к коду, в котором на каждые N цифр сообщения приходится k элементарных сигналов, где k — целое число, удовлетворяющее неравенствам

$$k - 1 < \log 10^N \leqslant k,$$

или, что то же самое,

$$N \log 10 \leqslant k < N \log 10 + 1.$$

Отсюда видно, что среднее число $\frac{k}{N}$ элементарных сигналов, приходящихся на одну десятичную цифру, в таком коде не может отличаться от величины $\log 10$ больше, чем на $\frac{1}{N}$; выбрав N достаточно большим, мы всегда можем сделать эту разницу сколь угодно малой (ср. стр. 143).

Разумеется, в приведенных рассуждениях почти ничего не изменится, если исходное сообщение не является числовым, а состоит из «букв» произвольного n-буквенного «алфавита» (например, из обычных русских букв, или из латинских букв, или из букв и цифр, или из букв, цифр и знаков препинания и т. д.). В этом случае, воспользовавшись кодированием сразу крупных блоков из N таких «букв» (для чего надо только разложить по двоичной системе счисления первые n^N чисел), можно добиться того, чтобы *среднее число элементарных сигналов, приходящихся на одну букву сообщения, стало сколь угодно близким к величине* $\log n$ (меньше этой величины наше среднее число никогда быть не может, как это следует из простого подсчета количества информации). Лишь, в том случае, когда n является целой степенью 2 (скажем, равно 2^k), такое разбиение на крупные блоки оказывается ненужным: здесь уже код, сопоставляющий некоторое кодовое обозначение каждой отдельной букве, может быть сделан предельно экономным, так что переход к кодированию по блокам не дает никакой выгоды. Отметим в этой связи, что в некоторых отношениях «кодирование по блокам» всегда является менее удобным, чем «кодирование по отдельным буквам»: при кодировании по блокам декодирование, естественно, оказывается более сложным и громоздким (в тем большей степени, чем длиннее соответствующие блоки) и, кроме того, оно производится здесь всегда с не-

которым запаздыванием (получив закодированное сообщение, мы не имеем возможности выяснить, какой была первая переданная буква, пока не будут переданы следующие $N - 1$ букв).

Все приведенные рассуждения без труда переносятся и на случай, когда при передаче используются не 2, а m элементарных сигналов (случай m-ичного кода); только здесь при построении наиболее экономного равномерного кода надо пользоваться не двоичной, а m-ичной системой счисления. Если n равно целой степени m, то вполне можно ограничиться кодированием каждой буквы сообщения в отдельности; уже при этом число элементарных сигналов, требующихся для передачи одной буквы, может быть сделано равным наименьшему возможному значению, а именно — значению $\frac{\log n}{\log m}$. Если же n не является целой степенью m, то при сопоставлении кодового обозначения каждой букве сообщения в отдельности нам придется затратить на каждую букву $k > \frac{\log n}{\log m}$ элементарных сигналов; здесь k — наименьшее целое число, превосходящее $\frac{\log n}{\log m}$. В этом случае мы можем построить более экономный код, перейдя к кодированию сразу целых N-буквенных блоков; выбрав N достаточно большим, мы всегда можем добиться того, чтобы *среднее число элементарных сигналов, затрачиваемых для передачи одной буквы сообщения, стало сколь угодно близким к* $\frac{\log n}{\log m}$. В частном случае $m = 3$ соответствующие рассуждения будут близки к тем, которые приводились в § 2 гл. III при определении числа взвешиваний на чашечных весах без гирь, нужного для нахождения фальшивой монеты (см. стр. 146 и след.): так как каждое взвешивание может иметь три исхода, то результат последовательности таких взвешиваний можно представить в виде последовательности цифр, каждая из которых принимает одно из трех значений [1]), т. е. в виде некоторого числа, записанного в троичной системе счисления.

---

[1]) Эти значения можно обозначить, как это принято в троичной системе счисления, цифрами 0, 1 и 2, а можно также вместо этого писать, например, буквы $Р$, $П$ и $Л$ (ср. с § 2 гл. III).

§ 2. Коды Шеннона — Фано и Хафмана. Основная теорема о кодировании

Основной результат предыдущего параграфа заключается в следующем: *если число букв в «алфавите» равно n, а число используемых элементарных сигналов равно m, то* при любом методе кодирования *среднее число элементарных сигналов, приходящихся на одну букву алфавита, не может быть меньше чем* $\frac{\log n}{\log m}$; *однако оно* всегда *может быть сделано сколь угодно близким к этому отношению*, если только отдельные кодовые обозначения сопоставлять сразу достаточно длинным «блокам», состоящим из большого числа букв. С идейной точки зрения этот результат, очевидно, примыкает к простейшим соображениям, высказанным в свое время Хартли: он никак не связан с теоретико-вероятностными рассмотрениями (слово «вероятность» в § 1 даже ни разу не упоминалось) и фактически опирается лишь на элементарный подсчет числа «различных последовательностей из N букв n-буквенного алфавита» и «различных последовательностей из N_1 элементарных сигналов». Поэтому вряд ли можно считать, что результаты предыдущего параграфа доказывают важность теории информации для технической задачи передачи сообщений, о чем говорилось в предисловии к настоящей книге.

На самом деле, однако, результаты § 1 могут быть значительно улучшены, если воспользоваться введенным в гл. II понятием энтропии и учесть статистические свойства реальных сообщений. В самом деле, ведь в § 1 экономность кода мы весьма грубо характеризовали лишь н а и б о л ь ш и м ч и с л о м элементарных сигналов, приходящихся на одну букву кодируемого сообщения, и в связи с этим рассматривали только простейшие коды — равномерные. Если в конце параграфа мы и говорили о с р е д н е м ч и с л е сигналов, приходящихся на одну букву сообщения, то это было связано лишь с тем, что там рассматривались равномерные коды сразу для многобуквенных блоков и отношение числа элементарных сигналов в кодовом обозначении к числу букв в соответствующем блоке (которое мы и называли средним числом элементарных сигналов, приходящихся на одну букву) могло не быть целым числом. Между тем на практике обычно приходится

иметь дело с сообщениями, в которых относительные частоты различных букв значительно отличаются друг от друга (достаточно сравнить, например, частоты букв *о* и *щ* в любом русском тексте; подробно об этом мы еще будем говорить в следующем параграфе). Поэтому основное значение здесь должно играть теоретико-вероятностное среднее значение числа элементарных сигналов, приходящихся на одну букву сообщения, определяемое в соответствии с реальными статистическими закономерностями, характеризующими передаваемые сообщения.

Посмотрим теперь, что же можно сказать о кодировании сообщений, подчиняющихся определенным статистическим закономерностям. Мы рассмотрим здесь лишь простейший случай сообщений, записанных при помощи некоторых n «букв», частоты появления которых на любом месте сообщения полностью характеризуются вероятностями $p_1, p_2, \ldots, p_n$, где, разумеется, $p_1 + p_2 + \ldots + p_n = 1$. Упрощение, которым мы при этом пользуемся, состоит в том, что вероятность p_i появления i-й буквы на любом месте сообщения предполагается одной и той же, вне зависимости от того, какие буквы стояли на всех предыдущих местах; иначе говоря, последовательные буквы сообщения предполагаются независимыми друг от друга. На самом деле в реальных сообщениях это чаще всего бывает не так; в частности, в русском языке вероятность появления той или иной буквы существенно зависит от предыдущей буквы (см. ниже, стр. 240 и след.). Однако строгий учет взаимной зависимости букв сделал бы все дальнейшие рассмотрения очень сложными; в то же время естественно думать, что он не должен изменить приведенные ниже результаты, так как, если угодно, под «буквами» мы можем сразу понимать блоки из многих букв, зависимость которых друг от друга является уже сравнительно слабой [1]).

Мы будем пока рассматривать только двоичные коды; обобщение полученных при этом результатов на коды, использующие произвольное число m элементарных сигналов, является, как всегда, крайне простым и о нем

[1]) Действительно, можно доказать, что все эти результаты сохраняются для весьма широкого класса случаев, в которых последовательные буквы сообщения являются зависимыми друг от друга (ср. ниже, стр. 216—217).

будет достаточно сказать лишь несколько слов в самом конце параграфа. Начнем с простейшего случая кодов, сопоставляющих отдельное кодовое обозначение — последовательность цифр 0 и 1 — каждой «букве» сообщения. Выше мы уже отмечали, что каждому двоичному коду для n-буквенного алфавита может быть сопоставлен некоторый метод отгадывания загаданного числа x, не превосходящего n, при помощи вопросов, на которые отвечается лишь «*да*» или «*нет*»; обратно — любой метод отгадывания такого числа приводит нас к определенному двоичному коду. При заданных вероятностях $p_1, p_2, \ldots, p_n$ отдельных букв передача многобуквенного сообщения соответствует как раз положению вещей, описанному на стр. 176 и след.; наиболее экономный код здесь будет сопоставляться тому методу отгадывания числа x, для которого при этих именно вероятностях n значений x среднее значение числа задаваемых вопросов оказывается наименьшим. Само это среднее значение может рассматриваться также как среднее значение числа двоичных знаков (цифр 0 и 1) в одном кодовом обозначении; иначе говоря, оно точно равно среднему значению числа элементарных сигналов, приходящихся на одну его букву при передаче многобуквенного сообщения.

Мы можем теперь непосредственно применить к нашей задаче результаты, изложенные на стр. 176 и след. Согласно этим результатам прежде всего *среднее число двоичных элементарных сигналов, приходящихся в закодированном сообщении на одну букву исходного сообщения, не может быть меньше* H, где $H = -p_1 \log p_1 - p_2 \log p_2 - \ldots - p_n \log p_n$ — энтропия опыта, состоящего в распознавании одной буквы текста (или, короче, просто энтропия одной буквы). Отсюда сразу следует, что *при любом методе кодирования для записи длинного сообщения из* M *букв требуется не меньше чем* MH *двоичных знаков*. Эти факты немедленно вытекают из того, что информация, содержащаяся в отрывке текста, содержащем M букв, в нашем случае равна MH (напомним, что отдельные буквы мы считаем взаимно независимыми); в то же время информация, содержащаяся в одном элементарном сигнале (двоичном знаке), никак не может превосходить одного бита (ср. стр. 177—178; другой вывод того же самого результата указан мелким шрифтом на стр. 180—182).

Если вероятности $p_1, p_2, \ldots, p_n$ не все равны между собой, то $H < \log n$; поэтому естественно думать, что учет статистических закономерностей сообщения может позволить построить код более экономный, чем наилучший равномерный код, требующий, согласно результатам § 1, не менее $M \log n$ двоичных знаков для записи текста из M букв. Как именно следует поступать для получения наиболее экономного кода, ясно из сказанного на стр. 176—177. Удобно при этом начать с того, что расположить все имеющиеся n букв в один столбик в порядке убывания вероятностей. Затем все эти буквы следует разбить на две группы — верхнюю и нижнюю — так, чтобы вероятности для буквы сообщения принадлежать к каждой из этих групп были возможно более близки одна к другой; для букв первой группы в качестве первой цифры кодового обозначения используется цифра 1, а для букв второй группы — цифра 0. Далее, каждую из двух полученных групп снова надо разделить на две части возможно более близкой суммарной вероятности; в качестве второй цифры кодового обозначения мы будем использовать цифру 1 или 0 в зависимости от того, принадлежит ли наша буква к первой или ко второй из этих более мелких групп. Затем каждая из содержащих более одной буквы групп снова делится на две части возможно более близкой суммарной вероятности и т. д.; процесс повторяется до тех пор, пока мы не придем к группам, каждая из которых содержит по одной-единственной букве. Такой метод кодирования сообщений был впервые предложен в 1948—1949 гг. независимо друг от друга Р. Фано и К. Шенноном; поэтому соответствующий код обычно называется кодом Шеннона — Фано (иногда также просто кодом Фано[1]). Так, например, если наш алфавит содержит всего шесть букв, вероятности которых (в порядке убывания) равны 0,4, 0,2, 0,2, 0,1, 0,05 и 0,05, то на первом этапе деления букв на группы мы отщепим лишь одну первую букву (1-я группа), оставив во 2-й группе все остальные. Далее, вторая буква составит 1-ую подгруппу 2-й группы; 2-я же подгруппа той же группы, состоящая из оставшихся четырех букв, будет и далее последовательно делиться на

[1]) На самом деле точно этот метод кодирования был предложен лишь Р. Фано; К. Шенноном же был предложен несколько другой метод, близкий к описанному здесь.

части так, что каждый раз 1-я часть будет состоять из одной лишь буквы (см. таблицу ниже). Аналогично в приведенной на следующей странице таблице разобран случай более богатого буквами «алфавита», включающего 18 букв, имеющих следующие вероятности: 0,3; 0,2; 0,1 (2 буквы); 0,05; 0,03 (5 букв); 0,02 (2 буквы); 0,01 (6 букв).

№ буквы	вероятность	разбиение на подгруппы (римские цифры обозначают номера групп и подгрупп)	кодовое обозначение
1	0,4	I	1
2	0,2	II I	01
3	0,2	II I	001
4	0,1	II II I	0001
5	0,05	II II I	00001
6	0,05	II II II	00000

Основной принцип, положенный в основу кодирования по методу Шеннона — Фано, заключается в том, что при выборе каждой цифры кодового обозначения мы стараемся, чтобы содержащееся в ней количество информации было наибольшим, т. е. чтобы независимо от значений всех предыдущих цифр эта цифра принимала оба возможных для нее значения 0 и 1 по возможности с одинаковой вероятностью. Разумеется, количество цифр в различных обозначениях при этом оказывается различным (в частности, во втором из разобранных примеров оно меняется от двух до семи), т. е. код Шеннона — Фано является неравномерным. Нетрудно понять, однако, что никакое кодовое обозначение здесь не может оказаться началом другого, более длинного обозначения (это ясно, в частности, и из того, что такой код фактически совпадает с описанным на стр. 176 и след. методом решения задачи об отгадывании загаданного числа; ср. стр. 189—190); поэтому закодированное сообщение всегда может быть однозначно декодировано. Весьма существенно, что буквам, имеющим большую вероятность, в коде Шеннона — Фано соответствуют более короткие кодовые обозначения, чем сравнительно маловероятным буквам (ибо при последовательном делении на группы буквы, имеющие большую вероятность, быстрее оказываются выделенными в от-

№ буквы	вероятность	разбиение на подгруппы	кодовое обозначение
1	0,3	I I	11
2	0,2	II	10
3	0,1	I	011
4	0,1	II I	0101
5	0,05	II	0100
6	0,03	I I	00111
7	0,03	I II	00110
8	0,03	II I	00101
9	0,03	II II	00100
10	0,03	I	00011
11	0,02	II I	000101
12	0,02	II II	000100
13	0,01	I	000011
14	0,01	II I I	0000101
15	0,01	II II	0000100
16	0,01	I	000001
17	0,01	II I	0000001
18	0,01	II II	0000000

дельную группу из одного элемента; ср. приведенные выше примеры). В результате, хотя некоторые кодовые обозначения здесь и могут иметь весьма значительную длину, среднее значение длины такого обозначения все же оказывается лишь немногим большим минимального значения H, допускаемого соображениями сохранения количества информации при кодировании. Так, для рассмотренного выше примера 6-буквенного алфавита наилучший равномерный код состоит из трехзначных кодовых обозначений (ибо $2^2 < 6 < 2^3$), и потому в нем на каждую букву исходного сообщения приходится ровно 3 элементарных сигнала; при использовании же кода Шеннона — Фано среднее число элементарных сигналов, приходящихся на одну букву сообщения, равно

$$1 \cdot 0{,}4 + 2 \cdot 0{,}2 + 3 \cdot 0{,}2 + 4 \cdot 0{,}1 + 5 \cdot (0{,}05 + 0{,}05) = 2{,}3.$$

Это значение заметно меньше, чем 3, и не очень далеко от энтропии

$$H = -0{,}4 \log 0{,}4 - 2 \cdot 0{,}2 \log 0{,}2 - 0{,}1 \log 0{,}1 - 2 \cdot 0{,}05 \log 0{,}5 \approx 2{,}22.$$

Аналогично этому для рассмотренного примера 18-буквенного алфавита наилучший равномерный код состоит из пятизначных кодовых обозначений (так как $2^4 < 18 < 2^5$); в случае же кода Шеннона — Фано имеются буквы, кодируемые даже семью двоичными сигналами, но зато среднее число элементарных сигналов, приходящихся на одну букву, здесь равно

$$2 \cdot 0{,}5 + 3 \cdot 0{,}1 + 4 \cdot 0{,}15 + 5 \cdot 0{,}15 + 6 \cdot 0{,}06 + 7 \cdot 0{,}04 = 3{,}29.$$

Последнее значение заметно меньше, чем 5, — и уже не намного отличается от величины

$$H = -0{,}3 \log 0{,}3 - 0{,}2 \log 0{,}2 - \ldots - 6 \cdot 0{,}0{,}01 \log 0{,}01 \approx \approx 3{,}25.$$

Особенно выгодно бывает кодировать по методу Шеннона — Фано не отдельные буквы, а сразу целые блоки из нескольких букв. Правда, при этом все равно невозможно превзойти предельное значение H двоичных знаков на одну букву сообщения (ибо для случая независимости отдельных букв энтропия N-буквенного блока будет равна NH и, следовательно, при любом методе кодирования на один блок никак не может прийтись в среднем меньше, чем NH двоичных знаков); однако даже в сравнительно неблагоприятных случаях кодирование целыми блоками позволяет весьма быстро приблизиться к этому минимальному значению. Рассмотрим, например, случай, когда имеются лишь две различные буквы A и $Б$, имеющие вероятности $p(A) = 0{,}7$ и $p(Б) = 0{,}3$; тогда

$$H = -0{,}7 \log 0{,}7 - 0{,}3 \log 0{,}3 = 0{,}881\ldots$$

Применение метода Шеннона — Фано к исходному двухбуквенному алфавиту здесь оказывается бесцельным: оно приводит нас лишь к простейшему равномерному коду

буква	вероятность	кодовое обозначение
A	0,7	1
$Б$	0,3	0

требующему для передачи каждой буквы одного двоичного знака — на 12% больше минимального достижимого зна-

чения 0,881 дв. зн./букву. Применяя же метод Шеннона — Фано к кодированию всевозможных двухбуквенных комбинаций (вероятности которых определяются правилом умножения вероятностей для независимых событий; см. стр. 29), мы придем к следующему коду:

комбинация букв	вероятность	кодовое обозначение
АА	0,49	1
АБ	0,21	01
БА	0,21	001
ББ	0,09	000

Среднее значение длины кодового обозначения здесь равно

$$1 \cdot 0{,}49 + 2 \cdot 0{,}21 + 3 \cdot 0{,}30 = 1{,}81,$$

так что на одну букву алфавита здесь приходится в среднем $\frac{1{,}81}{2} = 0{,}905$ двоичных знаков — лишь на 3% больше значения 0,881 дв. зн./букву. Еще лучшие результаты мы получим, применив метод Шеннона — Фано к кодированию трехбуквенных комбинаций; при этом мы придем к следующему коду:

комбинация букв	вероятность	кодовое обозначение
ААА	0,343	11
ААБ	0,147	10
АБА	0,147	011
БАА	0,147	010
АББ	0,063	0010
БАБ	0,063	0011
ББА	0,063	0001
ББББ	0,027	0000

Среднее значение длины кодового обозначения здесь равно 2,686, т. е. на одну букву текста приходится в среднем 0,895 двоичных знаков, что всего на 1,5% больше значения $H \approx 0{,}881$ дв. зн./букву.

В случае еще большей разницы в вероятностях букв A и $Б$ приближение к минимально возможному значению H дв. зн./букву может быть несколько менее быстрым, но оно проявляется не менее наглядно. Так, при $p(A) = 0{,}89$ и $p(Б) = 0{,}11$ это значение равно $-0{,}89 \log 0{,}89-$

$-0{,}11\log 0{,}11 \approx 0{,}5$ дв. зн./букву, а равномерный код $A \to 1$, $Б \to 0$ (равносильный применению кода Шеннона — Фано к совокупности двух имеющихся букв) требует затраты одного двоичного знака на каждую букву — в два раза больше. Нетрудно проверить, однако, что применение кода Шеннона — Фано к всевозможным двухбуквенным комбинациям здесь приводит к коду, в котором на каждую букву приходится в среднем 0,66 двоичных знаков; применение того же кода к блокам из трех букв позволяет понизить среднее число двоичных знаков, приходящихся на одну букву, до 0,55; наконец, кодирование по методу Шеннона — Фано всевозможных четырехбуквенных блоков требует затраты на каждую букву в среднем 0,52 двоичных знаков — всего на 4% больше минимального значения 0,50 дв. зн./букву.

Близок к коду Шеннона — Фано, но еще выгодней, чем этот последний, так называемый код Хаффмана (см. [63]), к описанию которого мы сейчас и перейдем. Построение этого кода опирается на простое преобразование того алфавита, на котором записываются передаваемые по линии связи сообщения, называемое *сжатием* алфавита. Пусть мы имеем алфавит A, содержащий буквы $a_1, a_2, \ldots, a_n$, вероятности появления которых в сообщении соответственно равны $p_1, p_2, \ldots, p_n$; при этом мы считаем буквы расположенными в порядке убывания их вероятностей (или частот), т. е. полагаем, что

$$p_1 \geqslant p_2 \geqslant p_3 \geqslant \ldots \geqslant p_{n-1} \geqslant p_n.$$

Условимся теперь *не различать между собой две наименее вероятные буквы нашего алфавита*, т. е. будем считать, что a_{n-1} и a_n — это одна и та же буква b нового алфавита A_1, содержащего, очевидно, буквы $a_1, a_2, \ldots, a_{n-2}$ и b (т. е. a_{n-1} или a_n), вероятности появления которых в сообщении соответственно равны $p_1, p_2, \ldots, p_{n-2}$ и $p_{n-1} + p_n$. Алфавит A_1 и называется *полученным из алфавита A с помощью сжатия* (или *однократного сжатия*).

Прилагательное «однократное» в скобках в конце последней фразы имеет следующий смысл. Расположим буквы нового алфавита A_1 в порядке убывания их вероятностей и подвергнем сжатию алфавит A_1; при этом мы придем к алфавиту A_2, про который естественно сказать, что он

получается из первоначального алфавита A с помощью *двукратного сжатия* (а из алфавита A_1 — с помощью простого или однократного сжатия). Ясно, что алфавит A_2 будет содержать уже всего $n - 2$ буквы. Продолжая эту процедуру, мы будем приходить ко все более коротким алфавитам; после $(n-2)$-кратного сжатия мы придем к алфавиту A_{n-2}, содержащему уже всего д в е буквы.

Вот, например, как преобразуется с помощью последовательных сжатий рассмотренный выше алфавит, содержащий 6 букв, вероятности которых равны 0,4, 0,2, 0,2, 0,1, 0,05 и 0,05:

№ буквы	Вероятности				
	исходный алфавит A	сжатые алфавиты			
		A_1	A_2	A_3	A_4
1	0,4	0,4	0,4	0,4	→0,6
2	0,2	0,2	0,2	→0,4	0,4
3	0,2	0,2	0,2	0,2	
4	0,1	0,1	→0,2		
5	0,05	→0,1			
6	0,05				

Условимся теперь приписывать двум буквам последнего алфавита A_{n-2} кодовые обозначения 1 и 0. Далее, если кодовые обозначения уже приписаны всем буквам алфавита A_j, то буквам «предыдущего» алфавита A_{j-1} (где, разумеется, $A_{l-1} = A_0$ — это исходный алфавит A), сохранившимся и в алфавите A_j, мы припишем т е ж е кодовые обозначения, которые они имели в алфавите A_{j-1}; двум же буквам a' и a'' алфавита A_j, «слившимся» в одну букву b алфавита A_{j-1}, мы припишем обозначения, получающиеся из кодового обозначения буквы b добавлением цифр 1 и 0 в конце (см. таблицу на следующей странице).

Легко видеть, что из самого построения получаемого таким образом *кода Хафмана* вытекает, что он удовлетворяет указанному на стр. 188 общему условию: никакое кодовое обозначение не является здесь началом другого, более длинного кодового обозначения. Заметим еще, что кодирование некоторого алфавита по методу Хафмана (так же, впрочем, как и по методу Шеннона — Фано) не

№ буквы	вероятности и кодовые обозначения				
	исходный алфавит A	сжатые алфавиты			
		A_1	A_2	A_3	A_4
1	0,4 0	0,4 0	0,4 0	0,4 0	0,6 1
2	0,2 10	0,2 10	0,2 10	0,4 11	0,4 0
3	0,2 111	0,2 111	0,2 111	0,2 10	
4	0,1 1101	0,1 1101	0,2 110		
5	0,05 11001	0,1 1100			
6	0,05 11000				

является однозначно определенной процедурой. Так, например, на любом этапе построения кода можно, разумеется, заменить цифру 1 на цифру 0 и наоборот; при этом мы получим два разных кода (отличающихся, правда, весьма несущественно друг от друга и имеющих то же длины всех кодовых обозначений). Но помимо того в некоторых случаях можно построить и несколько существенно различающихся кодов Хафмана; так, например, в разобранном выше примере можно строить код и в соответствии со следующей таблицей:

№ буквы	вероятности и кодовые обозначения				
	исходный алфавит A	сжатые алфавиты			
		A_1	A_2	A_3	A_4
1	0,4 11	0,4 11	0,4 11	0,4 0	0,6 1
2	0,2 01	0,2 01	0,2 10	0,4 11	0,4 0
3	0,2 00	0,2 00	0,2 01	0,2 10	
4	0,1 100	0,1 101	0,2 00		
5	0,05 1011	0,1 100			
6	0,05 1010				

Получаемый при этом новый код также является кодом Хафмана, но длины имеющихся кодовых обозначений теперь уже оказываются совсем другими. Отметим, однако, что среднее число элементарных сигналов, приходящихся на одну букву, для обоих построенных кодов Хафмана оказывается точно одинаковым: в первом случае оно равно

$$1 \cdot 0{,}4 + 2 \cdot 0{,}2 + 3 \cdot 0{,}2 + 4 \cdot 0{,}1 + 5 \cdot (0{,}05 + 0{,}05) = 2{,}3,$$

а во втором — равно

$$2\cdot(0{,}4 + 0{,}2 + 0{,}2) + 3\cdot 0{,}1 + 4\cdot(0{,}05 + 0{,}05) = 2{,}3.$$

Далее, оба кода явно относятся к числу весьма экономных (в данном конкретном случае средняя длина кодового обозначения здесь совпадает с той, которая получилась выше при использовании кода Шеннона — Фано). Более того, можно показать, что код Хафмана всегда является самым экономным из всех возможных в том смысле, что *ни для какого другого метода кодирования букв некоторого алфавита среднее число элементарных сигналов, приходящихся на одну букву, не может быть меньше того, какое получается при кодировании по методу Хафмана* (отсюда, разумеется, сразу вытекает и то, что для любых двух кодов Хафмана средняя длина кодового обозначения должна быть точно одинаковой — ведь оба они являются наиболее экономными).

Доказательство этого свойства оптимальности кодов Хафмана совсем несложно. Рассмотрим снова какой-то n-буквенный алфавит (обозначим его, например, через B), содержащий буквы $b_1, b_2, \ldots, b_{n-1}, b_n$, вероятности которых равны $q_1, q_2, \ldots, q_{n-1}, q_n$, где

$$q_1 \geqslant q_2 \geqslant \ldots \geqslant q_{n-1} \geqslant q_n, \qquad (*)$$

и получающийся из него сжатием (n — 1)-буквенный алфавит (алфавит B_1), содержащий буквы $b_1, b_2, \ldots, b_{n-2}, c$, вероятности появления которых соответственно равны $q_1, q_2, \ldots, q_{n-2}, q_{n-1} + q_n = q$. Предположим теперь, что мы имеем какую-то систему кодовых обозначений для букв алфавита B_1; эту систему кодовых обозначений мы перенесем затем и в алфавит B, сохранив обозначения всех букв, входящих одновременно и оба алфавита, а буквам b_{n-1} и b_n приписав обозначения, получающиеся из обозначения буквы c прибавлением в конце соответственно цифр 1 и 0. Покажем теперь, что *если код для алфавита* B_1 *был оптимальным, то и полученный таким путем код для алфавита* B *будет оптимальным.*

Выделенное курсивом утверждение мы будем доказывать от противного. А именно, мы предположим, что полученный код для B *не является оптимальным* и исходный *код для* B_1. В самом деле, обозначим среднюю длину кодового обозначения буквы (т. е. среднее число приходящихся на одну букву элементарных сигналов) для рассматриваемых кодов, отвечающих алфавитам B_1 и B, через L_1 и L; при этом, очевидно,

$$L = L_1 + q_2 \qquad (**)$$

Действительно, алфавиты B_1 и B отличаются лишь тем, что имеющая вероятность q буква c алфавита B_1 заменяется в алфавите B двумя буквами b_{n-1} и b_n с той же самой общей вероятностью

появления q ($= q_{n-1} + q_n$); отвечающие же этим алфавитам длины кодовых обозначений отличаются лишь увеличением на единицу длин, отвечающих буквам b_{n-1} и b_n, по сравнению с длиной, отвечающей букве c алфавита B_1. Отсюда и из определения средней длины кодового обозначения сразу следует соотношение (**).

Мы предположили, что отвечающий алфавиту B код и е о п т и м а л е н; другими словами — что существует отличный от рассматриваемого код, сопоставляющий буквам $b_1, b_2, \ldots, b_{n-1}, b_n$ кодовые обозначения длин (в элементарных сигналах) $k_1, k_2, \ldots, k_{n-1}, k_n$, такой, что для него средняя длина кодового обозначения одной буквы

$$L' = k_1 \cdot q_1 + k_2 \cdot q_2 + \ldots + k_{n-1} \cdot q_{n-1} + k_n \cdot q_n$$

меньше L. При этом мы можем считать, что

$$k_1 \leqslant k_2 \leqslant \ldots \leqslant k_{n-1} \leqslant k_n. \qquad (***)$$

В самом деле, если b_i и b_j (где i и j — какие-то два из номеров $1, 2, \ldots, n$) — такие буквы, что $q_i < q_j$ (откуда в силу (*) следует неравенство $i > j$), а $k_i < k_j$, то мы просто поменяем кодовые обозначения букв b_i и b_j, после чего средняя длина кодового обозначения буквы еще уменьшится; поэтому если $q_i > q_j$, то обязательно $k_i \leqslant k_j$. Ну а в пределах группы букв $b_u, b_{u+1}, \ldots, b_v$ (где $1 < u < v \leqslant n$) такой, что $q_u = q_{u+1} = \ldots = q_v$, мы всегда можем расположить буквы в таком порядке, что $k_u \leqslant k_{u+1} \leqslant \ldots \leqslant k_v$.

Из неравенств (***), в частности, следует, что букве b_n отвечает кодовое обозначение, имеющее с р е д н ю ю б о л ь ш у ю длину k_n. Далее, мы можем быть уверены в существовании такой буквы b_l алфавита B, кодовое обозначение которой *получается из кодового обозначения буквы* b_n *заменой последнего элементарного сигнала* — 1 на 0 или 0 на 1. В самом деле, если бы такое кодовое обозначение вовсе отсутствовало, то мы могли бы просто откинуть последний элементарный сигнал в кодовом обозначении буквы b_n, не придя при этом в противоречие с основным условием, определяющим коды без разделительного знака (напомним, что букв, имеющих более длинные, чем b_n, кодовые обозначения, у нас нет). Но при этом мы снова уменьшили бы среднюю длину кодового обозначения одной буквы, что противоречит предположению об оптимальности рассматриваемого кода.

Но из неравенств (***) и равенства $k_l = k_n$ следует, что неизбежно $k_l = k_{n-1}$ (но при этом не обязательно $l = n - 1$). Поменяем теперь кодовые обозначения букв b_l и b_{n-1}, если $l \neq n - 1$ (если $l = n - 1$, то этот этап рассуждения является лишним); при этом величина L', очевидно, не изменится. А теперь перейдем от рассматриваемого кода для алфавита B к коду для алфавита B_1, сохранив кодовые обозначения всех букв $b_1, b_2, \ldots, b_{n-2}$, а букве c приписав кодовое обозначение, получающееся из кодовых обозначений букв b_{n-1} и b_n отбрасыванием последней цифры (которой эти кодовые обозначения лишь и отличаются). Очевидно, что средняя длина L_1' полученного таким путем кода для алфавита B_1 связана

со средней длиной L' кода для B аналогичным (**) соотношением

$$L' = L'_1 + q,$$

откуда, в силу неравенства $L' < L$, следует, что

$$L'_1 < L_1.$$

Но это и доказывает, что исходный код для B_1 *не был оптимальным*.

Мы, по существу, уже завершили доказательство оптимальности кодов Хафмана. Действительно, ясно, что принятый нами код для последнего алфавита A_{n-2}, приписывающий двум буквам, из которых этот алфавит состоит, кодовые обозначения 1 и 0, является оптимальным: отвечающая ему средняя длина 1 кодового обозначения буквы никак не может быть уменьшена. Но отсюда, в силу только что доказанного, следует, что и код для алфавита A_{n-3} является оптимальным, откуда, в свою очередь, вытекает оптимальность кода для алфавита A_{n-4} и т. д.— и так до последнего кода (кода Хафмана), отвечающего исходному алфавиту $A_{1-1} = A_0$, т. е. алфавиту A.

Достигнутая в рассмотренных выше примерах степень близости среднего числа двоичных знаков, приходящихся на одну букву сообщения, к значению H может быть еще сколь угодно увеличена при помощи перехода к кодированию все более и более длинных блоков. Это вытекает из следующего общего утверждения, которое мы будем в дальнейшем называть основной теоремой о кодировании[1]): *при кодировании сообщения, разбитого на N-буквенные блоки, можно, выбрав N достаточно большим, добиться того, чтобы среднее число двоичных элементарных сигналов, приходящихся на одну букву исходного сообщения, было сколь угодно близко к H* (другими словами — сколь угодно близко к отношению количества H информации, содержащейся в одной букве сообщения, к 1 биту, т. е. к наибольшему количеству информации, могущему содержаться в одном элементарном сигнале). Иначе это можно сформулировать еще так: *очень длинное сообщение из M букв может быть закодировано при помощи сколь угодно близкого к MH* (но, разумеется, ни в каком случае не меньшего!) *числа элементарных сигналов, если только предварительно разбить это сообщение на*

[1]) Точнее следовало бы сказать: основной теоремой о кодировании при отсутствии помех. Обобщение этого результата на случай наиболее выгодного кодирования, учитывающего влияние помех, будет рассмотрено в § 4.

достаточно длинные блоки из N букв и сопоставлять отдельные кодовые обозначения сразу целым блокам. Отметим еще, что мы не случайно ничего не говорим здесь о том, как именно следует производить кодирование N-буквенных блоков: как будет видно из дальнейшего, методы кодирования блоков могут быть весьма различными (так, например, здесь можно — хоть это, разумеется, далеко не единственные имеющиеся здесь возможности — следовать методу кодирования Шеннона — Фано или методу Хафмана). Таким образом, основную роль при получении наиболее экономного кода играет именно разбиение сообщения на весьма длинные блоки. В § 4 мы увидим, что кодирование сразу длинных блоков имеет значительные преимущества и при наличии помех, препятствующих работе линий связи (хотя сами методы кодирования при этом требуют существенного изменения).

Ввиду большой важности сформулированной здесь основной теоремы о кодировании мы приведем ниже д в а совершенно разных ее доказательства (оба они фактически принадлежат К. Шеннону). Первое из них, по существу, опирается на использование метода кодирования Шеннона—Фано, хотя, как мы увидим ниже, прямой апелляции к этому методу доказательство не содержит. Предположим сначала, что при составляющем основу метода Шеннона — Фано последовательном делении совокупности кодируемых букв (под которыми могут пониматься также и целые «блоки») на все меньшие и меньшие группы нам каждый раз удается добиться того, чтобы вероятности двух получаемых групп были точно р а в н ы между собой. В таком случае после первого деления мы придем к группам, имеющим суммарную вероятность 1/2, после второго — к группам суммарной вероятности 1/4, ..., после l-го деления — к группам суммарной вероятности $1/2^l$. При этом l-значное кодовое обозначение будут иметь те буквы, которые оказываются выделенными в группу из одного элемента ровно после l делений, т. е. буквы, вероятность которых равна $1/2^l$; иначе говоря, при выполнении этого условия *длина l_i кодового обозначения будет связана с вероятностью p_i соответствующей буквы формулой*

$$p_i = \frac{1}{2^{l_i}}, \quad l_i = \log \frac{1}{p_i} = -\log p_i.$$

На самом деле, наше условие может быть точно выполнено лишь в некоторых исключительных случаях: из последних формул сразу следует, что вероятности p_i всех букв алфавита здесь должны равняться единице, деленной на целую степень числа 2. В общем же случае величина $-\log p_i$, где p_i — вероятность i-й буквы алфавита, как правило, целым числом не будет; поэтому длина l_i кодового обозначения i-й буквы не сможет быть равна $-\log p_i$. Поскольку, однако, при кодировании по методу Шеннона — Фано мы последовательно делим наш алфавит на группы по возможности близкой суммарной вероятности, то длина кодового обозначения i-й буквы при таком кодировании будет *близка к* $-\log p_i$. Обозначим, в этой связи, через l_i первое целое число, не меньшее чем $-\log p_i$, т. е. такое, что

$$-\log p_i \leqslant l_i < -\log p_i + 1. \qquad (А)$$

Последнее неравенство можно переписать еще так:

$$-l_i \leqslant \log p_i < -(l_i - 1),$$

или

$$\frac{1}{2^{l_i}} \leqslant p_i < \frac{1}{2^{l_i - 1}}. \qquad (Б)$$

Докажем теперь, что *существует такой метод кодирования, при котором длина кодового обозначения i-й буквы точно равна этому числу l_i*; только этот факт (а не описание соответствующего метода кодирования [1])) понадобится нам для доказательства основной теоремы.

Покажем прежде всего, что *в случае любых n чисел $l_1, l_2, \ldots, l_n$, удовлетворяющих неравенству*

$$\frac{1}{2^{l_1}} + \frac{1}{2^{l_2}} + \cdots + \frac{1}{2^{l_n}} \leqslant 1, \qquad (1)$$

существует двоичный код, для которого эти числа являются длинами кодовых обозначений, сопоставляемых n буквам некоторого алфавита. В самом деле, пусть n_1 — число тех из чисел $l_1, l_2, \ldots, l_n$, которые равны 1; n_2 — число тех

[1]) О нем см. текст, напечатанный мелким шрифтом на стр. 231 и след.

из них, которые равны 2; ... ; наконец, n_k — число тех из этих чисел, которые равны k (где $n_1 + n_2 + \ldots + n_k = n$, так что k — это значение наибольших из чисел $l_1, l_2, \ldots, l_n$). В таком случае неравенство (1) можно переписать в виде

$$\frac{n_1}{2} + \frac{n_2}{4} + \frac{n_3}{8} + \cdots + \frac{n_k}{2^k} \leqslant 1,$$

откуда сразу следует, что

$$\frac{n_1}{2} \leqslant 1 \qquad \text{или } n_1 \leqslant 2;$$

$$\frac{n_2}{4} \leqslant 1 - \frac{n_1}{2} \qquad \text{или } n_2 \leqslant 2(2 - n_1);$$

$$\frac{n_3}{8} \leqslant 1 - \frac{n_1}{2} - \frac{n_2}{4} \quad \text{или } n_3 \leqslant 2[4 - (2n_1 + n_2)];$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\frac{n_k}{2^k} \leqslant 1 - \frac{n_1}{2} - \frac{n_2}{4} - \frac{n_3}{8} - \cdots - \frac{n_{k-1}}{2^{k-1}} \quad \text{или}$$

$$n_k \leqslant 2[2^{k-1} - (2^{k-2}n_1 + 2^{k-3}n_2 + \cdots + n_{k-1})]$$

(ср. выше, стр. 181—182). Но ясно, что условие $n_1 \leqslant 2$ гарантирует возможность выбрать n_1 различных кодовых обозначений длины 1. Аналогично этому неравенство $n_2 \leqslant 2(2 - n_1)$ указывает на возможность выбрать дополнительно n_2 кодовых обозначений длины 2, начинающихся с двоичных цифр, отличных от тех, которые уже «заняты» кодовыми обозначениями длины 1: ведь число таких «свободных» начальных цифр равно $2 - n_1$ и к каждой из них можно приписать в конце или цифру 0, или цифру 1. Точно так же неравенство $n_3 \leqslant 2[4 - (2n_1 + n_2)]$ гарантирует возможность выбирать n_3 кодовых обозначений длины 3, первая цифра которых отлична от n_1 цифр, «занятых» кодовыми обозначениями длины 1, а первые две цифры — от n_2 двузначных чисел, «занятых» кодовыми обозначениями длины 2 (так как $2n_1 + n_2$ — это число двузначных двоичных чисел, или начинающихся одной из n_1 цифр, являющихся кодовыми обозначениями длины 1, или же совпадающих с одним из n_2 обозначений длины 2, а 4 — это число всех возможных двузначных двоичных чисел, с которых в принципе может начинаться кодовое

обозначение длины 3). Последнее рассуждение может быть продолжено и дальше; при этом неравенство

$$n_k \leqslant 2\,[2^{k-1} - (2^{k-2}n_1 + 2^{k-3}n_2 + \cdots + n_{k-1})],$$

относящееся к числу n_k, обеспечивает возможность выбора n_k кодовых обозначений длины k, первая цифра которых не совпадает ни с одним из n_1 кодовых обозначений длины 1, первые две цифры не совпадают ни с одним из n_2 кодовых обозначений длины 2, первые три цифры не совпадают ни с одним из n_3 кодовых обозначений длины 3 и т. д., так что из 2^{k-1} в принципе возможных начальных комбинаций $k-1$ двоичных цифр $2^{k-2}n_1 + 2^{k-3}n_2 + \ldots$ $\ldots + n_{k-1}$ комбинаций являются «занятыми» (ср. выше, стр. 181—182). В результате как раз и получается результат, согласно которому выполнение неравенства (1) гарантирует возможность выбора n кодовых обозначений длин $l_1, l_2, \ldots, l_n$, удовлетворяющих условию, напечатанному на стр. 188 курсивом; эти-то обозначения мы и можем сопоставить имеющимся n буквам алфавита.

Для завершения доказательства существования требуемого кода нам остается только заметить, что в силу неравенства (Б), определяющего длины l_i кодовых обозначений, $\frac{1}{2^{l_i}} \leqslant p_i$ при всех $i = 1, 2, \ldots, n$, где p_i — вероятность i-й буквы алфавита. Таким образом,

$$\frac{1}{2^{l_1}} + \frac{1}{2^{l_2}} + \cdots + \frac{1}{2^{l_n}} \leqslant p_1 + p_2 + \cdots + p_n = 1;$$

поэтому числа l_i действительно удовлетворяют неравенству (1), нужному для того, чтобы они могли быть длинами кодовых обозначений двоичного кода.

Теперь уже совсем легко доказать основную теорему о кодировании. В самом деле, среднее число l двоичных сигналов, приходящихся на одну букву исходного сообщения (иначе говоря, средняя длина кодового обозначения), по определению, дается суммой

$$l = p_1 l_1 + p_2 l_2 + \ldots + p_n l_n.$$

Умножим теперь задающее величину l_i неравенство (А) на p_i, сложим все полученные таким образом неравенства,

отвечающие значениям $i = 1, 2, \ldots, n$, и учтем, что

$$H = -p_1 \log p_1 - p_2 \log p_2 - \ldots - p_n \log p_n,$$

где $H = H(\alpha)$ — энтропия опыта α, состоящего в определении одной буквы сообщения, и что $p_1 + p_2 + \ldots + p_n = 1$. В результате получаем, что

$$H \leqslant l < H + 1.$$

Применим это неравенство к случаю, когда описанный выше метод используется для кодирования всевозможных N-буквенных блоков (которые можно считать «буквами» нового алфавита). В силу предположения о независимости последовательных букв сообщения энтропия опыта $\alpha_1 \alpha_2 \ldots \alpha_N$, состоящего в определении всех букв блока, равна

$$H(\alpha_1 \alpha_2 \ldots \alpha_N) = H(\alpha_1) + H(\alpha_2) + \ldots + H(\alpha_N) = NH(\alpha) = NH.$$

Следовательно, средняя длина l_N кодового обозначения N-буквенного блока удовлетворяет неравенствам

$$NH \leqslant l_N < NH + 1.$$

Но при кодировании сразу N-буквенных блоков среднее число l двоичных элементарных сигналов, приходящихся на одну букву сообщения, будет равно средней длине l_N кодового обозначения одного блока, деленной на число N букв в блоке:

$$l = \frac{l_N}{N}.$$

Поэтому при таком кодировании

$$H \leqslant l < H + \frac{1}{N},$$

т. е. здесь среднее число элементарных сигналов, приходящихся на одну букву, отличается от минимального значения H не больше, чем на $\frac{1}{N}$. Полагая $N \to \infty$, мы сразу приходим к основной теореме о кодировании.

Прежде чем идти дальше, отметим, что приведенное здесь доказательство может быть применено также и к

более общему случаю, когда последовательные буквы текста являются взаимно зависимыми. При этом придется только неравенство для величины l_N писать в виде

$$H^{(N)} \leqslant l_N < H^{(N)} + 1,$$

где

$$H^{(N)} = H(\alpha_1\alpha_2\alpha_3 \dots \alpha_N) = H(\alpha_1) + H_{\alpha_1}(\alpha_2) + H_{\alpha_1\alpha_2}(\alpha_3) + \dots + H_{\alpha_1\alpha_2\dots\alpha_{N-1}}(\alpha_N)$$

— энтропия N-буквенного блока, которая в случае зависимости букв сообщения друг от друга всегда будет меньше чем NH (ибо $H(\alpha_1) = H$ и $H(\alpha_1) \geqslant H_{\alpha_1}(\alpha_2) \geqslant H_{\alpha_1\alpha_2}(\alpha_3) \geqslant \dots \geqslant H_{\alpha_1\alpha_2\dots\alpha_{N-1}}(\alpha_N)$). Отсюда следует, что

$$\frac{H^{(N)}}{N} \leqslant l \leqslant \frac{H^{(N)}}{N} + \frac{1}{N},$$

где l есть среднее число элементарных сигналов, приходящихся на одну букву сообщения, и, значит, в этом более общем случае *при* $N \to \infty$ (при безграничном увеличении длины блоков) *среднее число элементарных сигналов, затрачиваемых на передачу одной буквы, неограниченно приближается к величине* H_∞, *где*

$$H_\infty = \lim_{N\to\infty} \frac{H^{(N)}}{N}$$

есть «удельная энтропия», приходящаяся на одну букву многобуквенного текста (об этой последней величине мы еще будем подробнее говорить в следующем параграфе) [1]).

Перейдем теперь ко второму доказательству нашей основной теоремы о кодировании; последовательные буквы сообщения мы при этом снова будем считать взаимно независимыми. Доказательство, которое будет дано ниже, является несколько более длинным, чем первое, но

---

[1]) Существование предела H_∞ сразу следует из неравенств $H(\alpha_1) \geqslant H_{\alpha_1}(\alpha_2) \geqslant H_{\alpha_1\alpha_2}(\alpha_3) \geqslant \dots$, показывающих, что последовательность $H(\alpha) = H^{(1)}, \frac{H^{(2)}}{2}, \frac{H^{(3)}}{3}, \dots, \frac{H^{(N)}}{N}, \dots$ является монотонно невозрастающей последовательностью положительных (т. е. больших нуля) чисел.

зато оно более поучительно, так как хорошо поясняет смысл самого понятия энтропии (см. выше, стр. 82—83). Кроме того, это новое доказательство покажет нам, что даже и в случае резко отличающихся вероятностей различных букв при кодировании очень длинных блоков все равно можно пользоваться «почти равномерным» кодом, сопоставляющим всем блокам, кроме некоторой части их, имеющей ничтожно малую суммарную вероятность, кодовые обозначения одинаковой длины. Что же касается последних «маловероятных» блоков, то легко понять, что их мы можем кодировать уже почти «как попало»: так как вероятность появления какого-либо из них очень мала, то метод кодирования этих блоков не будет играть существенной роли.

Для большей наглядности мы начнем наше доказательство с подробного рассмотрения простейшего случая, когда весь «алфавит» состоит всего из двух букв a и b, имеющих вероятности $p_1 = p$ и $p_2 = 1 - p = q$. Будем кодировать всевозможные цепочки («блоки»), состоящие из N последовательных букв a и b. Общее число различных таких N-членных цепочек будет равно 2^N (ср. стр. 82). Однако большинство из этих N-членных цепочек будут иметь ничтожную вероятность: так как относительная частота появления двух букв рассматриваемого «алфавита» равна p и q, то при достаточно большом N заметную вероятность будет иметь лишь совокупность тех цепочек, в которых из общего числа N букв примерно Np раз встречается буква a, а остальные примерно $N - Np = Nq$ раз — буква b. Выражаясь точнее, можно сказать, что при очень большом N все цепочки, в которых относительная частота появления буквы a не заключена в пределах между $p - \varepsilon$ и $p + \varepsilon$, где ε — произвольно выбранное очень маленькое число (например, 0,001 или 0,0001, или 0,000001; за ε можно принять любое из этих чисел и даже любое еще меньшее число, если только N будет достаточно велико), имеют крайне малую суммарную вероятность, так что их можно вообще не принимать в расчет. Что же касается цепочек, в которых буква a встречается от $N(p - \varepsilon)$ до $N(p + \varepsilon)$ раз, то каждая такая цепочка в отдельности также, разумеется, будет очень маловероятной (при большом N общее число возможных цепочек очень велико, а вероятность каждой из них в от-

дельности очень мала), но суммарная вероятность всех этих цепочек будет весьма близка к 1.

Заметим теперь, что число N-буквенных цепочек, в которых буква a встречается ровно Np раз [1]), равно числу C_N^{Np} сочетаний из N элементов по Np. Поэтому нам надлежит теперь оценить величину C_N^K (см. сноску [1])) в зависимости от N и K.

Для того чтобы идея приведенного ниже рассуждения стала более ясной, напомним сначала (не нужный нам для дальнейшего!) вывод формулы для числа C_N^K. Предположим, что *мы имеем* N (бумажных) *жетонов и* N *различных красок, которыми мы хотим окрасить эти жетоны* — каждый своей краской. Так как первый жетон мы можем окрасить любой из имеющихся N красок, второй — любой из $N-1$ оставшихся, третий — любой из неиспользованных ранее $N-2$ красок и т. д., наконец, последний жетон — единственной оставшейся в нашем распоряжении краской, то *общее число возможных окрасок жетонов равно*

$$N\cdot(N-1)\cdot(N-2)\cdot(N-3)\cdot\ldots\cdot 1 = N!.$$

Назовем теперь какие-то K красок «первыми», а оставшие $N-K$ красок — «вторыми»; далее выберем какие-то K жетонов, которые мы будем считать «первыми» (а прочие $N-K$ жетонов — «вторыми»). В таком случае мы будем иметь $K!$ способов окраски K «первых» жетонов K «первыми» красками и $(N-K)!$ способов окраски оставшихся $N-K$ жетонов $N-K$ «вторыми» красками. Комбинируя любой из $K!$ способов окраски K «первых» жетонов с любым из $(N-K)!$ способов окраски оставшихся жетонов, мы получим всего

$$K!\cdot(N-K)!$$

способов окраски N жетонов, при которых «первыми» K красками окрашены выбранные K «первых» жетонов. А так как, кроме того, K «первых» жетонов можно выбрать из общего числа N жетонов C_N^K способами, то *общее число*

[1]) Если Np не целое, то это число мы заменим ближайшим целым числом K: при большом N различие между Np и K будет весьма малоощутимым. Аналогичное замечание можно сделать и относительно числа $N\varepsilon$.

различных окрасок должно быть равным

$$C_N^K K!\,(N-K)!.$$

Следовательно,

$$N! = C_N^K K!\,(N-K)!,$$

откуда и следует искомая формула

$$C_N^K = \frac{N!}{K!\,(N-K)!}\,. \qquad (*)$$

Хорошо известная формула (*) дает *точное* выражение для числа C_N^K через числа N и K; однако при б о л ь ш о м N (а ведь нас в дальнейшем только и будет интересовать случай большого N) она малоудобна. В самом деле, $N!$ есть произведение N р а з л и ч н ы х множителей; оценить его величину при большом N совсем не просто. Поэтому в дальнейшем мы будем пользоваться не этой формулой, а *приближенной оценкой* для величины C_N^K, отличающейся от правой части формулы (*) главным образом тем, что в ней будут фигурировать с т е п е н и чисел N, K и $N-K$, которые легко оценить с помощью логарифмирования. Вот как получается эта оценка числа C_N^K.

Рассмотрим ту же задачу об окрашивании N жетонов N красками, которая использовалась для вывода формулы (*); только теперь мы откажемся от условия, чтобы *каждый жетон был* обязательно *окрашен своей краской*. В таком случае первый жетон мы по-прежнему сможем окрасить любой из N красок; однако также и второй, и третий, ..., и последний жетон мы также сможем покрасить любой из N красок, так что *общее число возможных раскрасок* будет в этом случае *равно*

$$\underbrace{N \cdot N \cdot \ldots N}_{N \text{ сомножителей}} = N^N.$$

Если теперь мы снова выберем какие-то K «первых» красок и K «первых» жетонов, то эти K жетонов K красками можно будет окрасить K^K способами; оставшиеся же $N-K$ жетонов можно будет $(N-K)^{N-K}$ способами окрасить $N-K$ «вторыми» красками. Комбинируя каждую из возможных K^K окрасок «первых» жетонов с каждой из $(N-K)^{N-K}$ окрасок оставшихся жетонов, мы получим

всего

$$K^K \cdot (N-K)^{N-K}$$

различных способов окраски всех N жетонов. Это последнее число надо еще умножить на C_N^K поскольку C_N^K, равно числу способов, каким можно выбрать K «первых» жетонов из общего числа N жетонов.

Заметим теперь, что получившееся число

$$C_N^K K^K (N-K)^{N-K}$$

н е р а в н о общему числу N^N возможных окрасок N жетонов, а *меньше* этого числа: в самом деле, $C_N^K \cdot K^K \times \times (N-K)^{N-K}$ — это число таких окрасок, в которых K *«первых» красок используются ровно K раз* (а ведь существуют и окраски, в которых эти K красок, например, используются N раз или не используются вовсе!). Таким образом, окончательно получаем

$$C_N^K K^K (N-K)^{N-K} < N^N,$$

откуда и следует нужная нам оценка величины C_N^K:

$$C_N^K < \frac{N^N}{K^K (N-K)^{N-K}}. \qquad (**)$$

Заменим теперь в неравенстве (**) K на Np; при этом $N-K$ обратится в $N-Np = N(1-p) = Nq$. Поэтому для числа C_N^{Np} «наиболее вероятных» N буквенных цепочек, т. е. таких, в которых буква a встречается ровно Np раз (а буква $б$ — оставшиеся Nq раз), мы получаем оценку

$$C_N^{Np} < \frac{N^N}{(Np)^{Np}(Nq)^{Nq}} = \frac{N^N}{N^{Np+Nq} p^{Np} q^{Nq}} =$$

$$= \frac{N^N}{N^N p^{Np} q^{Nq}} = \frac{1}{p^{Np} q^{Nq}}.$$

Примерно, столько же будет цепочек, в которых буква a встречается $Np+1$, $Np+2$, ..., $Np+N\varepsilon$ раз или же $Np-1$, $Np-2$, ..., $Np-N\varepsilon$ раз (так как во всех этих случаях отклонение частоты появления буквы a от той,

для которой мы производили наш расчет, будет очень небольшим). Поэтому, не делая большой ошибки, можно считать, что общее число «вероятных» цепочек (т. е. таких цепочек, что все остальные цепочки вместе взятые имеют ничтожно малую вероятность, которой можно пренебречь) не превосходит значения

$$M_1 = 2N\varepsilon \cdot \frac{1}{p^{Np}q^{Nq}} = \frac{2N\varepsilon}{p^{Np}q^{Nq}},$$

где ε — некоторое малое число.

Воспользуемся теперь для кодирования наших M_1 (или менее M_1) «вероятных» цепочек наилучшим равномерным кодом [1]). Так как число таких цепочек очень велико, то длина кодового обозначения при этом будет практически совпадать с двоичным логарифмом числа цепочек (ср. стр. 143), т. е. будет не больше чем

$$\log M_1 = \log 2\varepsilon + \log N - N\,(p \log p + q \log q).$$

Поэтому среднее число двоичных знаков, приходящихся на одну букву сообщения, здесь не превосходит величины

$$H + \frac{\log N}{N} + \frac{\log 2\varepsilon}{N},$$

где

$$H = -p \log p - q \log q.$$

При $N \to \infty$ оба члена в предпоследней формуле, следующие за H, стремятся к нулю (напомним, что отношение $\frac{\log N}{N} = -\frac{1}{N}\log\frac{1}{N}$ неограниченно убывает при возрастании N; см. стр. 72), откуда и вытекает, что ограничившись одними лишь «вероятными» цепочками, можно сделать среднее число двоичных знаков, приходящихся на одну букву сообщения, сколь угодно близким к H [2]).

---

[1]) Отметим, что применение к этим «вероятным» цепочкам неравномерного кода не может дать существенной выгоды, так как вероятности всех таких цепочек сравнительно мало отличаются друг от друга (поскольку относительная частота отдельных букв во всех их примерно одна и та же).

[2]) Меньше H это число быть не может (см. выше, стр. 200).

Что же касается оставшихся «маловероятных» цепочек, то если даже мы потратим на кодирование каждой буквы этих цепочек в несколько раз больше чем H двоичных знаков, то все равно среднее значение числа таких знаков, приходящихся на одну букву сообщения, при этом почти не изменится (так как суммарная вероятность всех таких цепочек ничтожно мала). Поэтому при кодировании фактически надо лишь позаботиться, чтобы ни одно из соответствующих кодовых обозначений не совпало с продолжением какого-либо из прочих используемых обозначений. Для этого можно, например, с самого начала добавить единицу к общему числу «вероятных» цепочек (замена M_1 на $M_1 + 1$, разумеется, не повлияет на дальнейшие оценки) и воспользоваться тем, что в таком случае мы, наверное, будем иметь по крайней мере одно «незанятое» кодовое обозначение той же длины, что и все обозначения для «вероятных» цепочек. Если теперь мы поместим в начале всех обозначений для «маловероятных» цепочек это «незанятое» кодовое обозначение, то тем самым уже будет гарантировано, что ни одно из новых обозначений не будет продолжением одного из старых. Вслед за этим обозначением мы можем добавить, например, результат применения к «маловероятным» цепочкам какого-либо наиболее экономного равномерного кода, после чего для всех «маловероятных» цепочек окончательно получаются обозначения одной и той же длины, удовлетворяющие требуемым условиям.

Общий случай n-буквенного алфавита, в котором отдельные буквы имеют вероятности p_1, p_2, ..., p_n, где $p_1 + p_2 + \ldots + p_n = 1$, разбирается почти так же. В случае длинных цепочек из N букв наибольшую вероятность будут иметь цепочки, в которых первая буква встречается около Np_1 раз, вторая — около Np_2 раз, ..., n-я — около Np_n раз. Число цепочек, в которых первая буква встречается ровно Np_1 раз, вторая — ровно Np_2 раз, ..., n-я — ровно Np_n раз, равно числу разбиений N элементов на n групп, содержащих соответственно Np_1, Np_2, ..., Np_n элементов.

Рассмотрев теперь *задачу об окрашивании N жетонов N красками так, чтобы каждая краска была использована ровно один раз*, и разбив краски на n групп, содержащих, соответственно, Np_1, Np_2, ..., Np_n красок, мы сможем

совершенно аналогично выводу формулы (*) доказать, что *число таких разбиений N элементов на n групп равно*

$$\frac{N!}{(Np_1)!\,(Np_2)!\,\ldots\,(Np_n)!}$$

(эта формула обобщает обычную формулу для числа сочетаний [1])). Рассмотрев далее задачу об окрашивании N жетонов N красками (по-прежнему разбитыми на n групп, первая из которых содержит Np_1 красок, вторая — Np_2 красок, ..., последняя — Np_n красок), в которой уже *не требуется, чтобы каждая краска была использована ровно один раз*, мы аналогично выводу неравенства (**) убедимся, что *интересующее нас число разбиений N элементов на n групп меньше величины*

$$\frac{1}{p_1^{Np_1}p_2^{Np_2}\ldots p_n^{Np_n}}\,.$$

Применив этот результат к «вероятным» цепочкам, у которых частота появления первой буквы заключена между $p_1-\varepsilon$ и $p_1+\varepsilon$, частота появления второй — между $p_2-\varepsilon$ и $p_2+\varepsilon$, ..., частота появления n-й буквы — между $p_n-\varepsilon$ и $p_n+\varepsilon$, получим, что общее число таких цепочек, наверное, не превосходит числа

$$(2N\varepsilon)^n\cdot\frac{1}{p_1^{Np_1}p_2^{Np_2}\ldots p_n^{Np_n}}=\frac{2^n\varepsilon^nN^n}{p_1^{Np_1}p_2^{Np_2}\ldots p_n^{Np_n}}\,.$$

Что касается остальных цепочек, у которых частота появления хоть одной из букв не укладывается в указанные пределы, то суммарная вероятность всех этих цепочек будет ничтожно мала, так что их можно вовсе не принимать во внимание.

Теперь уже совсем легко показать, что закодировав все наши «вероятные» цепочки с помощью наиболее экономного равномерного кода, мы придем к кодовым обозначениям, длина которых не больше чем

$$NH+n\log N+n\log 2\varepsilon,$$

где

$$H=-p_1\log p_1-p_2\log p_2-\ldots-p_n\log p_n.$$

[1]) Вывод этой формулы имеется также в книге [26], стр. 75.

Следовательно, среднее число двоичных знаков, требующихся для записи одной буквы, здесь не превосходит

$$H + n\frac{\log N}{N} + \frac{n\log 2\varepsilon}{N}.$$

При $N \to \infty$ это число стремится к H, что и дает предельное среднее число двоичных знаков, приходящихся на одну букву сообщения при кодировании по такому методу. Это и есть тот результат, который мы стремились доказать.

В заключение стоит еще раз подчеркнуть принципиальную основу приведенного доказательства. Если мы будем рассматривать все цепочки из N букв n-буквенного «алфавита» (или, что то же самое, все цепочки из N последовательных исходов многократно повторяющегося опыта, могущего иметь n различных исходов), то общее число различных таких цепочек будет равно

$$n^N = 2^{N\log n}.$$

Однако вероятность каждой отдельной такой цепочки и даже некоторых значительных совокупностей таких цепочек при большом N будет совершенно ничтожной. Нами было доказано, что если мы разрешим исключить из рассмотрения часть наименее вероятных цепочек, но только так, чтобы суммарная вероятность всех отброшенных цепочек была достаточно мала (скажем, не превосходила некоторого заранее выбранного очень малого числа δ), то при л ю б о м (сколь угодно малом!) δ в случае достаточно большого N можно добиться того, чтобы оставшееся число цепочек имело порядок

$$\left(\frac{1}{p_1}\right)^{Np_1}\left(\frac{1}{p_2}\right)^{Np_2}\cdots\left(\frac{1}{p_n}\right)^{Np_n} = 2^{NH},$$

где H — энтропия [1]). Отметим тут же, что поскольку H меньше, чем $\log n$ (за исключением случая равной

[1]) Выражение «имело порядок» здесь означает, что на самом деле перед 2^{NH} может стоять еще некоторый множитель, пропорциональный конечной степени N (т. е. пропорциональный $2^{A\log N}$, где A — фиксированное число); ясно, что при очень большом N этот множитель будет во много раз меньше основного члена 2^{NH} и не будет играть существенной роли. Отметим в этой связи, что в приведенном выше выводе мы доказали только, что число

вероятности всех букв или всех исходов), то число наших «вероятных» цепочек при очень большом N будет несравненно меньшим общего числа цепочек (отношение $2^{NH} : 2^{N\log n} = 2^{-N(\log n - H)}$ числа «вероятных» цепочек к числу всех цепочек быстро стремится к нулю при $N \to \infty$). Кроме того, мы доказали, что при большом N можно добиться того, чтобы относительные частоты появления отдельных букв в наших «вероятных» цепочках сколь угодно мало отличались от наиболее вероятных частот $p_1, p_2, \ldots, p_n$. Так как вероятность той или иной цепочки зависит лишь от количества встречающихся в ней отдельных букв (вероятность цепочки, в которой первая буква встречается N_1 раз, вторая — N_2 раз, ..., n-я — N_n раз, равна $p_1^{N_1} p_2^{N_2} \ldots p_n^{N_n}$), то отсюда видно, что при большом N можно добиться, чтобы все «вероятные» цепочки очень мало отличались по своим вероятностям. Иначе говоря, мы доказали здесь утверждение, выделенное курсивом на стр. 82—83; именно это утверждение и определяет основную роль понятия энтропии в теории кодирования.

Ввиду особой важности указанного утверждения имеет смысл задержаться на нем немного больше и привести еще одно простое его доказательство. Выше мы исходили из подсчета общего числа N-буквенных цепочек, в которых частоты отдельных букв алфавита мало отличаются от соответствующих вероятностей $p_1, p_2, \ldots, p_n$. При этом было также отмечено, что вероятности всех таких цепочек близки между собой и практически не отличаются от вероятности $p_1^{Np_1} p_2^{Np_2} \ldots p_n^{Np_n}$ цепочки, в которой $N_1 =$

«вероятных» цепочек не превосходит значения $(2\varepsilon)^n N^n \cdot 2^{NH}$. Легко понять, однако, что оно во всяком случае не меньше, чем число цепочек, в которых первая буква встречается ровно Np_1 раз, вторая — ровно Np_2 раз, ..., n-я — ровно Np_n раз, а последнее число, как было показано, всегда больше, чем

$$\frac{1}{p_1^{Np_1} p_2^{Np_2} \ldots p_n^{Np_n}} = 2^{NH}.$$

Таким образом, с точностью до множителя порядка конечной степени N число «вероятных» цепочек действительно совпадает с 2^{NH}.

$= Np_1$, $N_2 = Np_2$, ..., $N_n = Np_n$, т. е. частоты появления каждой из n букв алфавита точно совпадают с вероятностями p_1, p_2, ..., p_n. Последнюю вероятность, очевидно, можно переписать в виде

$$(2^{\log p_1})^{Np_1} (2^{\log p_2})^{Np_2} \dots (2^{\log p_n})^{Np_n} = \\ = 2^{N(p_1 \log p_1 + p_2 \log p_2 + \dots + p_n \log p_n)} = 2^{-HN}$$

(так как $H = -p_1 \log p_1 - p_2 \log p_2 - \dots - p_n \log p_n$ — это фиксированное конечное число, а N очень велико, то ясно, что 2^{-HN} — это очень малая вероятность). Заметим теперь, что полученная формула сразу влечет за собой также и нужную нам оценку общего числа различных «вероятных» цепочек. В самом деле, ведь суммарная вероятность всех таких цепочек очень близка к единице (она отличается от единицы лишь на какое-то очень малое число); поскольку вероятность суммы несовместимых событий равна сумме соответствующих вероятностей, ясно, что общее число рассматриваемых цепочек должно быть близко к единице, деленной на вероятность отдельной цепочки, т. е. к числу 2^{HN}. Таким образом, интересующее нас утверждение будет доказано, если только мы сможем показать, что из совокупности всех n^N цепочек из N букв можно выбросить какую-то совокупность «редких» цепочек (суммарная вероятность которых при достаточно большом N может быть сделана сколь угодно малой) так, чтобы все оставшиеся цепочки имели уже практически одинаковую вероятность, равную 2^{-HN}.

Заметим теперь, что вероятность любой цепочки из N букв n-буквенного алфавита (характеризуемого вероятностями 1-й, 2-й, ..., n-й букв, равными соответственно p_1, p_2, ..., p_n) такой, что эти N букв выбираются последовательно одна за другой независимо от ранее выбранных букв, равна произведению $p_{i_1} p_{i_2} \dots p_{i_N}$, где i_1, i_2, ..., i_N — номера последовательных букв нашей цепочки. Следовательно, логарифм этой вероятности равен

$$\log p_{i_1} + \log p_{i_2} + \dots + \log p_{i_N} = \\ = \frac{\log p_{i_1} + \log p_{i_2} + \dots + \log p_{i_N}}{N} \cdot N.$$

Но величины p_{i_1}, p_{i_2}, ..., p_{i_N} все определяются результатами опытов, состоящих в выборе одной из букв среди n

букв алфавита; поэтому все они представляют собой случайные величины, могущие принимать n значений $\log p_1$, $\log p_2$, ..., $\log p_n$ с вероятностями, равными, соответственно, p_1, p_2, ..., p_n. Применив к такой случайной величине доказанный на стр. 57—59 закон больших чисел, мы найдем, что с вероятностью, которую при достаточно большом N можно считать сколь угодно близкой к единице, среднее арифметическое

$$\frac{\log p_{i_1} + \log p_{i_2} + \cdots + \log p_{i_N}}{N}$$

будет отличаться от

$$\text{ср.зн.}\log p = p_1 \log p_1 + p_2 \log p_2 + \ldots + p_n \log p_n = -H$$

не больше, чем на заданное очень малое число ε. Но это и значит, что *из числа всех N-буквенных цепочек можно выбросить какую-то совокупность «редких» цепочек очень малой суммарной вероятности так, чтобы вероятность всех оставшихся цепочек была уже примерно одинаковой и весьма близкой к 2^{-HN}*. Последнее утверждение и есть то, которое мы стремились доказать.

Остановимся еще вкратце на вопросе о роли предположения, согласно которому последовательные буквы сообщения выбираются каждый раз независимо от всех предшествующих букв. На стр. 216—217 мы уже указывали, что первое из рассматривавшихся доказательств основной теоремы о кодировании на самом деле не зависит от выполнения этого условия, причем в общем случае *взаимно зависимых* букв значение энтропии H одной буквы должно быть заменено приходящейся на одну букву удельной энтропией $H_\infty = \lim\limits_{N\to\infty} \frac{H^{(N)}}{N}$ (где $H^{(N)}$ — энтропия блока из N букв). Исходя отсюда кажется естественным предположить, что и второе доказательство, в ходе которого существенно использовалось предположение о независимости букв сообщения, на самом деле должно быть применимо и к общему случаю сообщений со взаимно зависимыми буквами. Иначе говоря, можно думать, что и в случае сообщений, буквы которых зависят друг от друга, *среди всех цепочек из N букв, где N достаточно велико, также можно выделить совокупность «вероятных» цепочек, суммарная вероятность которых будет*

лишь очень мало отличаться от единицы, причем число этих вероятных цепочек будет иметь порядок $2^{H_\infty N} \approx 2^{H^{(N)}}$, *а вероятность каждый из них будет близка к* $2^{-H_\infty N} \approx 2^{-H^{(N)}}$. Выделенное курсивом утверждение играет в теории информации очень важную роль; однако его доказательство не очень просто, и, кроме того, оно вообще не может быть получено для всех без исключения случаев, а требует, чтобы распределения вероятностей для последовательных букв сообщения удовлетворяли бы некоторым дополнительным условиям (весьма общим и на практике всегда выполняющимся, но не могущим даже быть сформулированными без привлечения ряда совсем новых теоретико-вероятностных понятий). Отметим еще, что сами эти условия могут выбираться по-разному; так, для одних таких условий сделанное выше утверждение было доказано еще Шенноном [1], теорема 3), в то время как позже совсем другие очень общие условия его справедливости были указаны Б. Макмилланом [64]. Мы здесь не будем более подробно задерживаться на этом вопросе, а ограничимся лишь ссылкой на книги [5] — [7] и [22], в которых он разбирается во всех деталях.

Все предыдущее содержание этого параграфа легко переносится также и на случай m-ичных кодов, использующих m элементарных сигналов. Так, например, для построения m-ичных кодов Шеннона — Фано надо лишь разбивать группы символов не на две, а на m частей по возможности близкой вероятности, а для построения m-ичного кода Хафмана надо использовать операцию сжатия алфавита, при которой каждый раз сливаются не две, а m букв исходного алфавита, имеющих наименьшие вероятности. Ввиду важности кодов Хафмана, остановимся на последнем вопросе чуть подробнее. Сжатие алфавита, при котором m букв заменяются на одну, приводит к уменьшению числа букв на $m - 1$; так как для построения m-ичного кода, очевидно, требуется, чтобы последовательность «сжатий» в конце концов привела нас к алфавиту из m букв (сопоставляемых m сигналам кода), то необходимо, чтобы число n букв первоначального алфавита было представимо в виде $n = m + k(m - 1)$, где k — целое число. Этого, однако, всегда можно добиться, добавив, если нужно, к первоначальному алфавиту еще

несколько «фиктивных букв», вероятности которых считаются равными нулю. После этого построение m-ичного кода Хафмана и доказательство его оптимальности (среди всех m-ичных кодов) проводятся уже точно так же, как и в случае двоичного кода. Так, например, в случае уже рассматривавшегося выше алфавита из 6 букв, имеющих вероятности 0,4, 0,2, 0,2, 0,1, 0,05 и 0,05 для построения троичного кода Хафмана, надо присоединить к нашему алфавиту еще одну фиктивную букву нулевой вероятности и далее поступать так, как указано ниже:

№ буквы	вероятности и кодовые обозначения: исходный алфавит		сжатые алфавиты		
1	0,4	0	0,4 0		0,4 0
2	0,2	2	0,2 2		→0,4 1
3	0,2	10	0,2 10		0,2 2
4	0,1	11	0,1 11	←	
5	0,05	120	→0,1 12		
6	0,05	121	←		
7	0	—			

Столь же просто переносятся на случай m-ичных кодов и оба приведенных выше доказательства основной теоремы о кодировании. В частности, соответствующее видоизменение первого доказательства основывается на том факте, что *любые n чисел $l_1, l_2, \ldots, l_n$, удовлетворяющих неравенству*

$$\frac{1}{m^{l_1}} + \frac{1}{m^{l_2}} + \cdots + \frac{1}{m^{l_n}} \leqslant 1, \qquad (2)$$

являются длинами кодовых обозначений некоторого m-ичного кода для n-буквенного алфавита. Доказательство этого факта точно повторяет рассуждения, приведенные на стр. 213—215 для случая $m = 2$; поэтому на нем здесь можно не задерживаться. Используя неравенство (2) так же, как на стр. 215—216 использовалось неравенство (1), легко получить следующий результат (называемый основной теоремой о кодировании для m-ичных кодов): *при любом методе кодирования,*

использующем m-ичный код, среднее число элементарных сигналов, приходящихся на одну букву сообщения, никогда не может быть меньше отношения $\frac{H}{\log m}$ (*где H — энтропия одной буквы сообщения*); *однако оно всегда может быть сделано сколь угодно близким к этой величине, если кодировать сразу достаточно длинные «блоки» из N букв.* Отсюда ясно, что если по линии связи за единицу времени можно передать L элементарных сигналов (принимающих m различных значений), то *скорость передачи сообщений по такой линии не может быть большей, чем*

$$v = \frac{L \log m}{H} \quad \textit{букв/ед. времени};$$

однако передача со скоростью, сколь угодно близкой к v (но меньшей v!), уже является возможной. Величина

$$C = L \log m,$$

стоящая в числителе выражения для v, зависит лишь от самой линии связи (в то время как знаменатель H характеризует передаваемое сообщение). Эта величина указывает наибольшее количество единиц информации, которое можно передать по нашей линии за единицу времени (ибо один элементарный сигнал, как мы знаем, может содержать самое большее $\log m$ единиц информации); она называется пропускной способностью линии связи. Понятие пропускной способности играет важную роль в теории связи; мы к нему еще вернемся в дальнейшем (см. стр. 312—320 и § 4 этой главы).

Сделаем еще одно замечание по поводу приведенного на стр. 212 и след. первого доказательства основной теоремы о кодировании. Центральную роль в этом доказательстве играл факт существования двоичного кода, в котором длина l_i кодового обозначения i-й буквы удовлетворяет неравенствам

$$-\log p_i \leqslant l_i < -\log p_i + 1 \qquad \text{(А)}$$

или, что то же самое,

$$\frac{1}{2^{l_i}} \leqslant p_i < \frac{1}{2^{l_i - 1}}. \qquad \text{(Б)}$$

В случае произвольного m-ичного кода эти неравенства принимают вид

$$-\frac{\log p_i}{\log m} \leqslant l_i < -\frac{\log p_i}{\log m} + 1 \qquad \text{(А')}$$

или, что то же самое,

$$\frac{1}{m^{l_i}} \leqslant p_i < \frac{1}{m^{l_i-1}}. \qquad (Б')$$

Выше мы доказали существование такого двоичного кода, опираясь на неравенство (1) на стр. 213, но не выписывая явным образом сами кодовые обозначения; в случае m-ичного кода совершенно также может быть использовано неравенство (2) на стр. 230. Сейчас мы опишем один метод построения самого соответствующего кода; при этом, для простоты, мы ограничимся случаем *десятичного* кода, сопоставляющего каждой из n б у к в алфавита какую-то последовательность ц и ф р 0, 1, . . ., 9 [1]). Для такого десятичного кода неравенства (А′) и (Б′), очевидно, принимают вид

$$-\lg p_i \leqslant l_i < -\lg p_i + 1 \qquad (А'')$$

(логарифмы — десятичные!) и

$$\frac{1}{10^{l_i}} \leqslant p_i < \frac{1}{10^{l_i-1}}. \qquad (Б'')$$

Расположим все буквы «алфавита» в порядке убывания их вероятностей: $p_1 \geqslant p_2 \geqslant p_3 \geqslant \ldots \geqslant p_n$. Среди этих вероятностей, разумеется, могут встречаться и одинаковые; поэтому вероятность сама по себе не может однозначно характеризовать соответствующей буквы. Если, однако, мы составим суммы:

$$P_1 = 0, \quad P_2 = p_1, \quad P_3 = p_1 + p_2, \quad P_4 = p_1 + p_2 + p_3, \ldots$$
$$\ldots, \quad P_n = p_1 + p_2 + \ldots + p_{n-1},$$

то эти суммы будут уже все различны; таким образом, n ч и с е л $P_1, P_2, \ldots, P_n$ можно рассматривать как своеобразный «алфавит», однозначно соответствующий исходному алфавиту из n б у к в. Нам надо теперь только закодировать этот новый «алфавит», т. е. сопоставить каждому из n чисел P_i определенную последовательность элементарных сигналов (или цифр) — этим одновременно будет решена и задача кодирования первоначального алфавита.

Нетрудно указать путь решения этой задачи. Представим каждое из (меньших единицы!) чисел P_i в виде (вообще говоря,

---

[1]) Общий случай будет отличаться от этого, в основном, тем, что там нам придется разлагать фигурирующие ниже числа P_i в (бесконечные) *m-ичные дроби*, т. е. представлять каждое число P_i в виде суммы

$$P_i = \frac{a_1}{m} + \frac{a_2}{m^2} + \frac{a_3}{m^3} + \ldots + \frac{a_k}{m^k} + \ldots,$$

где все «цифры» $a_1, a_2, \ldots, a_k, \ldots$ в записи этой дроби принимают какие-то из значений $0, 1, \ldots, m-1$. Мы рекомендуем читателю самостоятельно продумать соответствующее построение.

бесконечной) *десятичной дроби*:

$$P_i = 0,a_1a_2a_3 \ldots a_k \ldots,$$

где a_1, a_2, a_3, ... — какие-то цифры (если P_i выражается в виде конечной десятичной дроби, то все цифры a_k, начиная с некоторой, будут равны 0). Тем самым каждому P_i сопоставляется бесконечная последовательность $a_1a_2a_3 \ldots$ цифр (т. е. элементарных сигналов); при этом полученные таким образом n последовательностей цифр будут, разумеется, все различны, ибо никакие два числа P_i не равны между собой.

Заметим теперь, что различие между последовательностями $a_1a_2a_3 \ldots$ не может проявляться лишь в очень далеких от начала цифрах. В самом деле, поскольку

$$P_{i+1} - P_i = p_i, \quad P_{i+2} - P_i = p_i + p_{i+1} \quad \text{и т. д.},$$

то в силу неравенств (Б″) все числа P_{i+1}, P_{i+2}, ..., P_n отличаются от числа P_i не меньше, чем на $\frac{1}{10^{l_i}}$; поэтому разложения всех этих чисел в десятичную дробь отличаются от десятичной дроби, выражающей число P_i, не позже чем в l_i-й цифре: *все десятичные дроби P_{i+1}, P_{i+2}, ..., P_n отличаются от десятичной дроби P_i хоть в одной из первых l_i цифр*. Поэтому, если мы оставим в отвечающей P_i (где $i = 1, 2, \ldots, n$) десятичной дроби лишь первые l_i цифр, то мы получим n (конечных!) десятичных дробей, которые все различны и ни одна из которых не является началом другой. Соответствующие n последовательностей $a_1a_2a_3 \ldots a_{l_i}$ цифр (сопоставляемые n буквам первоначального алфавита) и образуют требуемый десятичный код.

Выше уже указывалось, что *любые n чисел l_1, l_2, ... l_n, удовлетворяющие неравенству*

$$\frac{1}{m^{l_1}} + \frac{1}{m^{l_2}} + \ldots + \frac{1}{m^{l_n}} \leqslant 1, \tag{2}$$

являются длинами кодовых обозначений для некоторого m-ичного кода, сопоставляющего n буквам алфавита n последовательностей элементарных сигналов, принимающих m возможных значений. Проведя соответствующие рассуждения в обратном порядке, нетрудно доказать и что *длины l_1, l_2, ..., l_n кодовых обозначений любого m-ичного кода для n-буквенного алфавита обязательно удовлетворяют неравенству* (2) — фактически это было уже установлено в конце предыдущей главы (см. стр. 181—182), правда, без использования принятых в этой главе терминов. Таким образом, *выполнение неравенства* (2) *необходимо и достаточно для того, чтобы числа l_1, l_2, ..., l_n могли являться длинами кодовых обозначений некоторого m-ичного кода*. Это утверждение впервые было доказано в 1949 г. американским ученым Л. Крафтом в его неопубликованной диссертации (см., например, книги [6] и [20]), а позже

было еще обобщено Б. Макмилланом [65]; поэтому неравенство (2) часто называют неравенством Крафта или неравенством Макмиллана. Обобщение Макмиллана связано с тем обстоятельством, что до сих пор мы рассматривали только коды, удовлетворяющие общему условию, выделенному курсивом на стр. 188 (и названные мгновенными или мгновенно декодируемыми в подстрочном примечании на той же странице); только к этим кодам и относились все приведенные выше рассуждения. Макмиллан доказал, однако, что *условие* (2) *необходимо и достаточно также и для существования однозначно декодируемого* (но не обязательно мгновенного!) *m-ичного кода с длинами кодовых обозначений, равными* $l_1, l_2, \ldots, l_n$. Так как любой мгновенный код одновременно является и однозначно декодируемым, то доказательства, очевидно, требует только необходимость указанного неравенства для однозначно декодируемых кодов, т. е. тот факт, что в случае любого однозначно декодируемого m-ичного кода для n-буквенного алфавита *длины* $l_1, l_2, \ldots, l_n$ *кодовых обозначений обязательно удовлетворяют неравенству* (2). Это последнее утверждение наиболее просто было доказано Дж. Карушем [66], которому мы и будем следовать в нашем изложении.

Обозначим сумму $\frac{1}{m^{l_1}}+\frac{1}{m^{l_2}}+\ldots+\frac{1}{m^{l_n}}$, где $l_1, l_2, \ldots, l_n$ — длины кодовых обозначений некоторого однозначно декодируемого m-ичного кода, сопоставляемых n буквам алфавита, через A и образуем выражение

$$A^t=\left(\frac{1}{m^{l_1}}+\frac{1}{m^{l_2}}+\ldots+\frac{1}{m^{l_n}}\right)^t=$$

$$=\underbrace{\left(\frac{1}{m^{l_1}}+\frac{1}{m^{l_2}}+\ldots+\frac{1}{m^{l_n}}\right)\left(\frac{1}{m^{l_1}}+\frac{1}{m^{l_2}}+\ldots+\frac{1}{m^{l_n}}\right)\cdots\left(\frac{1}{m^{l_1}}+\frac{1}{m^{l_2}}+\ldots+\frac{1}{m^{l_n}}\right)}_{t\text{ раз}}.$$

Раскрыв в последнем произведении скобки, мы получим сумму n^t слагаемых вида $\frac{1}{m^N}$, где каждый показатель степени N равен какой-то сумме вида $l_{i_1}+l_{i_2}+\ldots+l_{i_t}$ (номера $i_1, i_2, \ldots, i_t$ здесь принимают значения $1, 2, \ldots, n$, причем они, разумеется, не должны быть все различными). Если предположить, что длины n кодовых обозначений для однозначно декодируемого m-ичного кода упорядочены так, что $1 \leqslant l_1 \leqslant l_2 \leqslant \ldots \leqslant l_n$, то для каждой суммы

$$N=l_{i_1}+l_{i_2}+\ldots+l_{i_t}$$

будет справедливо двойное неравенство

$$t \leqslant N \leqslant tl_n$$

(ясно, что $N = t$, если $l_{i_1} = l_{i_2} = \ldots = l_{i_t} = 1$, и $N = tl_n$, если $l_{i_1} = l_{i_2} = \ldots = l_{i_t} = l_n$). Обозначим теперь число различных сумм $l_{i_1} + l_{i_2} + \ldots + l_{i_t}$, принимающих значение N, через K_N. Тогда легко видеть, что, раскрыв скобки в выражении A^t, мы получим

$$A^t = \left(\frac{1}{m^{l_1}} + \frac{1}{m^{l_2}} + \ldots + \frac{1}{m^{l_n}}\right)^t = K_t \frac{1}{m^t} + K_{t+1}\frac{1}{m^{t+1}} + \ldots + K_{tl_n}\frac{1}{m^{tl_n}}$$

(где, вообще говоря, некоторые из коэффициентов K_t, K_{t+1}, ..., K_{tl_n} будут равны нулю). Заметим теперь, что число K_N различных сумм $l_{i_1} + l_{i_2} + \ldots + l_{i_t}$, принимающих значение N, равно числу разных t-буквенных слов $b_{i_1}b_{i_2}\ldots b_{i_t}$ (где $b_1, b_2, \ldots, b_n$ — буквы нашего алфавита), кодируемых последовательностями из N элементарных сигналов. Так как рассматриваемый код является однозначно декодируемым, то ясно, что

$$K_N \leqslant m^N;$$

действительно, m^N — это общее число различных последовательностей из N сигналов, каждый из которых может принимать одно из m значений, а если бы какие-то два разных слова кодировались бы одинаковыми последовательностями элементарных сигналов, то это значило бы, что код не является однозначно декодируемым. Поэтому при любом (натуральном) t

$$A^t = K_t\frac{1}{m^t} + K_{t+1}\frac{1}{m^{t+1}} + \cdots + K_{tl_n}\frac{1}{m^{tl_n}} \leqslant m^t\frac{1}{m^t} +$$

$$+ m^{t+1}\frac{1}{m^{t+1}} + \cdots + m^{tl_n}\frac{1}{m^{tl_n}} = tl_n - (t-1) \leqslant tl_n.$$

А отсюда уже и следует, что

$$A \leqslant 1$$

(т. е. что справедливо неравенство (2)!): в самом деле, при любом $A > 1$ величина A^t росла бы при возрастании t быстрее, чем ct, где c — какое угодно фиксированное число [1]) (например, число l_n), и поэтому при достаточно большом t обязательно выполнилось бы неравенство $A^t > l_n t$.

---

[1]) Обозначим $1/A$ через p; тогда $\log(A^t) = t\log A = \dfrac{\log A}{p}$, а $\log(ct) = \log c + \log t = \log c - \log p$. Ясно, что при малом p (т. е. при большом t) первое из этих чисел гораздо больше второго, ибо $\log c$ — постоянное (не зависящее от p) число, $\log A > 0$ (так как $A > 1$), а отношение $(-\log p)$: $\dfrac{\log A}{p} = \dfrac{1}{\log A}(-p\log p)$ стремится к нулю при $p \to 0$ (ср. стр. 72).

Из того, что и для мгновенных и для любых однозначно декодируемых кодов необходимое и достаточное условие существования кода с кодовыми обозначениями данных длин $l_1, l_2, \ldots, l_n$ имеет один и тот же вид (2), следует, что *для любого однозначно декодируемого m-ичного кода существует мгновенный код с кодовыми обозначениями букв, имеющими те же длины, что и в случае исходного однозначно декодируемого кода.* А отсюда, в свою очередь, в частности вытекает, что *коды Хафмана являются оптимальными* (т. е. имеющими наименьшую среднюю длину кодового обозначения одной буквы) не только среди всех мгновенных кодов (а именно это и было нами доказано на стр. 209—211; ср. также стр. 230), но и *среди всех* вообще *однозначно декодируемых кодов.*

§ 3. Энтропия и информация конкретных типов сообщений

В предыдущих двух параграфах мы занимались вопросом о кодировании и передаче абстрактных «сообщений», записанных на некотором «языке», «алфавит» которого содержит n «букв». Здесь мы скажем о тех выводах, которые можно сделать отсюда в применении к конкретным типам сообщений — в первую очередь к сообщениям на русском языке или на каком-либо из иностранных языков (см. также [67], [68]).

Письменная речь

Основной результат § 1 этой главы состоял в том, что для передачи M-буквенного сообщения (где M считается достаточно большим) по линии связи, допускающей m различных элементарных сигналов, требуется затратить не меньше чем $\frac{M \log n}{\log m}$ сигналов, где n — число букв «алфавита», с помощью которого записано сообщение; при этом существуют методы кодирования, позволяющие сколь угодно близко подойти к границе $\frac{M \log n}{\log m}$. Так как русский «телеграфный» алфавит содержит 32 буквы (мы здесь не различаем букв *е* и *ё*, *ь* и *ъ*, которые в большинстве телеграфных кодов передаются одной и той же комбинацией элементарных сигналов, но причисляем к числу букв и «нулевую букву» — пустой промежуток между словами), то согласно этому результату на передачу M-буквенного сообщения надо затратить $M \frac{\log 32}{\log m} = M \frac{H_0}{\log m}$

элементарных сигналов. Здесь

$$H_0 = \log 32 = 5 \text{ бит}$$

— энтропия опыта, заключающегося в приеме одной буквы русского текста (информация, содержащаяся в одной букве), при условии, что все буквы считаются о д и н а к о в о в е р о я т н ы м и.

На самом деле, однако, появление в сообщении на русском языке разных букв совсем не одинаково вероятно. Так, например, в любом тексте буквы *о* и *е* встречаются много чаще, чем буквы *ф* или *щ*; поскольку средняя длина слова в русском языке значительно меньше 31 буквы, то вероятность появления пробела («нулевой буквы») на много превосходит значение 1/32, которое мы имели бы, если бы все 32 буквы были равновероятны. Поэтому информация, содержащаяся в одной букве любого осмысленного русского текста, всегда меньше, чем $\log 32 = 5$ бит. Отсюда ясно, что для получения текста, в котором каждая буква содержит 5 бит информации, нельзя просто взять отрывок из какой-либо русской книги; для этого требуется выписать 32 буквы на отдельных билетиках, сложить все эти билетики в урну и затем вытаскивать их по одному, каждый раз записывая вытянутую букву, а билетик возвращая обратно в урну и снова перемешивая ее содержимое. Произведя такой опыт, мы придем к «фразе» вроде следующей:

СУХЕРРОБЬДЩ ЯЫХВЩИЮЛЙЖТЛФВНЗАГФОЕНВШТЦР ШХГБКУЧТЭЮРЯПЧЬКЙХРЫС

Разумеется, этот текст, хоть он и составлен из русских букв, имеет очень мало общего с русским языком!

Для более точного вычисления информации, содержащейся в одной букве русского текста, надо знать в е р о я т н о с т и появления различных букв. Эти вероятности можно приближенно определить, взяв достаточно большой отрывок, написанный по-русски, и рассчитав для него относительные частоты отдельных букв. Строго говоря, эти частоты могут несколько зависеть от характера текста (например, в учебнике по высшей математике частота обычно очень редкой буквы *ф* будет заметно выше средней из-за частого повторения слов «функция», «дифференциал»,

«коэффициент» и некоторых других; еще большие отклонения от нормы в частоте употребления отдельных букв можно наблюдать в некоторых художественных произведениях, особенно в стихах[1]); поэтому для надежного определения «средней частоты» буквы желательно иметь набор различных текстов, заимствованных из различных источников. Как правило, однако, подобные отклонения будут все же сравнительно небольшими и в первом приближении ими можно пренебречь. Ориентировочные значения частот отдельных букв русского языка собраны в следующей таблице (ср. А. А. Харкевич [69], Д. С. Лебедев и В. А. Гармаш [70]; тире здесь означает пробел между словами):

буква	—	*о*	*е, ё*	*а*	*и*	*т*	*н*	*с*
относ. частота	0,175	0,090	0,072	0,062	0,062	0,053	0,053	0,045
буква	*р*	*в*	*л*	*к*	*м*	*д*	*п*	*у*
относ. частота	0,040	0,038	0,035	0,028	0,026	0,025	0,023	0,021
буква	*я*	*ы*	*з*	*ь, ъ*	*б*	*г*	*ч*	*й*
относ. частота	0,018	0,016	0,016	0,014	0,014	0,013	0,012	0,010
буква	*х*	*ж*	*ю*	*ш*	*ц*	*щ*	*э*	*ф*
относ. частота	0,009	0,007	0,006	0,006	0,004	0,003	0,003	0,002

Приравняв эти частоты вероятностям появления соответствующих букв, получим для энтропии одной буквы русского текста приближенное значение[2]):

$$H_1 = H(\alpha_1) = -0{,}175 \cdot \log 0{,}175 - 0{,}090 \cdot \log 0{,}90 - \\ -0{,}072 \cdot \log 0{,}072 - \ldots - 0{,}002 \cdot \log 0{,}002 \approx 4{,}35 \text{ бит.}$$

[1]) В качестве примера здесь можно назвать, скажем, некогда знаменитое стихотворение К. Д. Бальмонта «Камыши» (Полночной порою в болотной глуши/Чуть слышно, бесшумно шуршат камыши...), все построенное на обыгрывании шипящих звуков *ч* и *ш*. Другие, еще гораздо более выразительные примеры того же рода, заимствованные из немецкой, английской и португальской художественной литературы, читатель может найти в гл. 3 книги [3].

[2]) Так как значения частот отдельных букв в отрывке, содержащем конечное число N букв, не совпадают точно с соответствующими вероятностями, то ясно, что значение энтропии, получающееся при замене вероятностей частотами, не будет точным. Вопросу о степени точности получаемых таким образом значений энтропии и о поправках, которые целесообразно вводить в них при недостаточно большом N, посвящена статья Г. П. Башарина [71]; те же поправки рассматриваются и в статье Дж. Миллера (G. A. Miller) в сборнике [46], стр. 95—100.

Из сравнения этого значения с величиной $H_0 = \log 32 = 5$ бит видно, что неравномерность появления различных букв алфавита приводит к уменьшению информации, содержащейся в одной букве русского текста, примерно на 0,65 бит.

Воспользовавшись этим обстоятельством, можно уменьшить число элементарных сигналов, необходимых для передачи M-буквенного сообщения, до значения $M\frac{H_1}{\log m}$ (т. е. в случае двоичного кода — до значения $H_1 M \approx \approx 4{,}35\,M$; для сравнения укажем, что $H_0 M = 5M$ — это значение, достигаемое при кодировании по методу Бодо, сопоставляющему M-буквенному сообщению цепочку из $5M$ элементарных сигналов). Сокращение числа требующихся элементарных сигналов может быть достигнуто, например, кодированием отдельных букв русского алфавита по методу Шеннона — Фано (см. выше, стр. 201 и след.). Нетрудно проверить, что применение этого метода к русскому алфавиту приводит к следующей таблице кодовых обозначений:

буква	код. обозн.	буква	код. обозн.	буква	код. обозн.
—	111	*к*	01000	*х*	0000100
а	1010	*л*	01001	*ц*	00000010
б	000101	*м*	00111	*ч*	000011
в	01010	*н*	0111	*ш*	00000011
г	000100	*о*	110	*щ*	00000001
д	001101	*п*	001100	*ы*	001000
е, ё	1011	*р*	01011	*ь, ъ*	000110
ж	0000011	*с*	0110	*э*	000000001
з	000111	*т*	1000	*ю*	00000010
и	1001	*у*	00101	*я*	001001
й	0000101	*ф*	000000000		

Среднее количество элементарных сигналов, требующихся для передачи одной буквы сообщения при таком методе кодирования, будет равно

$$0{,}265\cdot 3 + 0{,}347\cdot 4 + 0{,}188\cdot 5 + 0{,}150\cdot 6 + 0{,}032\cdot 7 + \\ + 0{,}013\cdot 8 + 0{,}005\cdot 9 \approx 4{,}4,$$

т. е. будет весьма близко к значению $H_1 \approx 4{,}35$ [1]).

[1]) Впрочем, значительная сложность расшифровки закодированного таким способом сообщения делает этот код практически

Но и равное $\frac{H_1}{\log m}$ значение среднего числа элементарных сигналов, приходящихся на одну букву передаваемого сообщения, также не является наилучшим. В самом деле, при определении энтропии $H_1 = H(\alpha_1)$ опыта α_1, состоящего в определении одной буквы русского текста, мы считали все буквы *независимыми*. Это значит, что для составления «текста», в котором каждая буква содержит $H_1 \approx 4{,}35$ бит информации, мы должны прибегнуть к помощи урны, в которой лежат тщательно перемешанные 1000 бумажек, на 175 из которых не написано ничего, на 90 — написана буква *о*, на 72 — буква *е*, ..., наконец, на 2 бумажках — буква *ф* (см. таблицу частот русских букв на стр. 238). Извлекая из такой урны бумажки по одной, мы придем к «фразе» вроде следующей [1]):

ЕЫНТ ЦИЯЬА ОЕРВ ОДНГ ЬУЕМЛОЛЙК
ЗБЯ ЕНВТША.

Эта «фраза» несколько более похожа на осмысленную русскую речь, чем предыдущая (здесь все же наблюдается сравнительно правдоподобное распределение числа гласных и согласных и близкая к обычной средняя длина «слова»), — но и она, разумеется, еще очень далека от разумного текста.

Несходство нашей фразы с осмысленным текстом естественно объясняется тем, что на самом деле последовательные буквы русского текста вовсе не независимы друг от друга. Так, например, если мы знаем, что очередной буквой явилась гласная, то значительно возрастает вероятность появления на следующем месте согласной буквы;

мало удобным. Это можно проверить, например, попытавшись декодировать, скажем, следующую «фразу»: 0101001011001001001101111010011001111010001100011011110000000000000000010111111011001010011011100001010111101111011111010001100011101101011100010111000110111 (декодирование заметно облегчается, если предварительно выписать все кодовые обозначения в порядке убывания вероятностей соответствующих букв).

[1]) Этот и последующие примеры «искусственных фраз» заимствованы из статьи Р. Л. Добрушина [72]. (Как объяснено в этой статье, вместо вытаскивания бумажек из урны с 1000 бумажками можно поступить значительно проще: воспользоваться любой русской книгой и выбрать из нее ряд букв *наудачу*.)

буква «ь» никак не может следовать ни за пробелом, ни за гласной буквой (в явном противоречии с тем, как обстояло дело во втором и пятом «словах» нашей «фразы»); за буквой «ч» никак не могут появиться буквы «ы», «я» или «ю», а скорее всего будет стоять одна из гласных «и» и «е» или согласная «т» (слово «что») и т. д.

Наличие в русском языке дополнительных закономерностей, не учтенных в нашей «фразе», приводит к дальнейшему уменьшению степени неопределенности (энтропии) одной буквы русского текста. Поэтому при передаче такого текста по линии связи можно еще уменьшить среднее число элементарных сигналов, затрачиваемых на передачу одной буквы. Нетрудно понять, как можно охарактеризовать количественно это уменьшение. Для этого надо лишь подсчитать условную энтропию $H_2 = H_{\alpha_1}(\alpha_2)$ опыта α_2, состоящего в определении одной буквы русского текста, при условии, что нам известен исход опыта α_1, состоящего в определении предшествующей буквы того же текста (заметим, что при приеме очередной буквы сообщения мы всегда знаем уже предшествующую букву). Согласно сказанному на стр. 91, условная энтропия H_2 определяется следующей формулой:

$$\begin{aligned} H_2 = H_{\alpha_1}(\alpha_2) &= H(\alpha_1\alpha_2) - H(\alpha_1) = \\ &= -p(--)\log p(--) - p(-а)\log p(-а) - \\ &\quad - p(-б)\log p(-б) - \ldots - p(яя)\log p(яя) + \\ &\quad + p(-)\log p(-) + p(а)\log p(а) + \\ &\quad + p(б)\log p(б) + \ldots + p(я)\log p(я), \end{aligned}$$

где через $p(-)$, $p(а)$, $p(б)$, ..., $p(я)$ обозначены вероятности (частоты) отдельных букв русского языка (их значения были нами указаны на стр. 238), а через $p(--)$, $p(-а)$, $p(-б)$, ..., $p(яя)$ — вероятности (частоты) всевозможных двухбуквенных сочетаний. Для приближенного определения таких «двухбуквенных вероятностей» надо лишь подсчитать частоты появления различных комбинаций двух соседних букв в каком-либо длинном русском отрывке; при этом, разумеется заранее можно сказать, что вероятности $p(--)$, $p(яь)$ и многие другие (например, $p(ьь)$, $p(-ь)$, $p(чя)$ и т. д.) будут равны нулю. Существенно подчеркнуть, что в силу результатов § 2 гл. II

мы можем быть уверены, что условная энтропия $H_2 = H_{\alpha_1}(\alpha_2)$ окажется м е н ь ш е безусловной энтропии H_1.

Величину H_2 можно конкретизировать как «среднюю информацию», содержащуюся в определении исхода следующего опыта. Имеется 32 урны, обозначенные 32 буквами русского алфавита; в каждой из урн лежат бумажки, на которых выписаны двухбуквенные сочетания, начинающиеся с обозначенной на урне буквы, причем количества бумажек с разными парами букв пропорциональны частотам (вероятностям) соответствующих двухбуквенных сочетаний. Опыт состоит в многократном извлечении бумажек из урн и выписывании с них последней буквы. При этом каждый раз (начиная со второго) бумажка извлекается из той урны, которая содержит сочетания, начинающиеся с последней выписанной буквы; после того как буква выписана, бумажка возвращается в урну, содержимое которой снова тщательно перемешивается. [Можно также (что заметно удобнее практически) вместо урны воспользоваться какой-либо русской книгой, в которой надо лишь, начиная каждый раз с выбранного наудачу места, отыскивать первое появление последней уже выписанной нами буквы и следующую за ней букву книги дописывать к уже имеющемуся тексту.] Опыт такого рода приводит к «фразе» вроде следующей:

УМАРОНО КАЧ ВСВАННЫЙ РОСЯ НЫХ КОВКРОВ НЕДАРЕ.

По звучанию эта «фраза» заметно ближе к русскому языку, чем «фраза», выписанная на стр. 240 (например, здесь мы имеем не только правдоподобное соотношение числа гласных и согласных букв, но и близкое к привычному чередование их, благодаря чему фразу уже можно «произнести»).

Разумеется, и величина $\frac{H_2}{\log m}$ не дает еще окончательной оценки наименьшего значения среднего числа элементарных сигналов, требующихся для передачи одной буквы русского текста. Дело в том, что в русском языке (как и любом другом) каждая буква зависит не только от непосредственно предшествующей ей, но и от р я д а предыдущих букв. Например, известно, что сочетание *ее* является

довольно частым, так что после буквы *е* мы свободно можем ожидать появления еще одного *е*; однако если также и предпоследней буквой является *е*, то появление еще одного *е* становится уже почти невероятным (ибо сочетание *еее* встречается крайне редко); после сочетания — *и* (буква *и* после пробела) весьма часто следует еще один пробел (союз «*и*»), а после сочетания *тс* естественно ожидать букву *я* (глагольное окончание «*тся*») и т. д. Поэтому знание д в у х предшествующих букв еще более уменьшает неопределенность опыта, состоящего в определении следующей буквы, что находит отражение в положительности разности $H_2 - H_3$, где H_3 — «условная энтропия второго порядка»:

$$H_3 = H_{\alpha_1\alpha_2}(\alpha_3) = H(\alpha_1\alpha_2\alpha_3) - H(\alpha_1\alpha_2) =$$
$$= -p\,(---)\log p\,(---) - p\,(--a)\log p\,(--a) - \ldots$$
$$\ldots - p\,(яяя)\log p\,(яяя) +$$
$$+ p\,(--)\log p\,(--) + p\,(-a)\log p\,(-a) + \ldots$$
$$\ldots + p\,(яя)\log p\,(яя).$$

Наглядным подтверждением сказанного является то обстоятельство, что опыт, состоящий в вытаскивании бумажек с трехбуквенными сочетаниями из 32^2 урн, в каждой из которых лежат бумажки, начинающиеся на одни и те же две буквы (или, что то же самое, опыт с русской книгой, в которой много раз наудачу отыскивается первое повторение последнего уже выписанного двухбуквенного сочетания и выписывается следующая за ним буква), приводит к «фразе» вроде следующей:

ПОКАК ПОТ ДУРНОСКАКА НАКОНЕПНО ЗНЕ СТВОЛОВИЛ СЕ ТВОЙ ОБНИЛЬ,

еще более близкой к русской речи, чем предыдущая.

Аналогично этому можно определить и энтропию

$$H_4 = H_{\alpha_1\alpha_2\alpha_3}(\alpha_4) = H(\alpha_1\alpha_2\alpha_3\alpha_4) - H(\alpha_1\alpha_2\alpha_3) =$$
$$= -p\,(----)\log p\,(----) -$$
$$- p\,(---a)\log p(---a) - \ldots$$
$$\ldots - p\,(яяяя)\log p\,(яяяя) +$$
$$+ p\,(---)\log p\,(---) + p\,(--a)\log p\,(--a) + \ldots$$
$$\ldots + p\,(яя)\log p\,(яя),$$

отвечающую опыту по определению буквы русского текста при условии знания т р е х предшествующих букв. Соответствующий этой величине опыт, состоящий в извлечении бумажек из 32^3 урн с четырехбуквенными сочетаниями (или — аналогичный описанному выше эксперимент с русской книгой), приводит к «фразе» вроде следующей:

ВЕСЕЛ ВРАТЬСЯ НЕ СУХОМ И НЕПО И КОРКО,

составленной уже из «почти русских» слов. Еще лучшее приближение к энтропии буквы осмысленного русского текста дают величины

$$H_N = H_{\alpha_1\alpha_2 \dots \alpha_{N-1}}(\alpha_N) = H(\alpha_1\alpha_2 \dots \alpha_N) - H(\alpha_1\alpha_2 \dots \alpha_{N-1})$$

при $N = 5, 6, \dots$. Нетрудно видеть, что с ростом N энтропия H_N может только убывать (ср. выше, стр. 125). Если еще учесть, что все величины H_N положительны, то отсюда можно будет вывести, что величина $H_{\alpha_1\alpha_2 \dots \alpha_{N-1}}(\alpha_N) = H_N$ при $N \to \infty$ стремится к определенному пределу H_∞, очевидно, совпадающему с пределом H_∞, о котором шла речь в предыдущем параграфе (см. стр. 217) [1]).

[1]) Равенство рассмотренного в § 2 предела

$$\lim_{N\to\infty} \frac{H^{(N)}}{N} = \lim_{N\to\infty} \frac{H(\alpha_1) + H_{\alpha_1}(\alpha_2) + \dots + H_{\alpha_1 \dots \alpha_{N-1}}(\alpha_N)}{N}$$

введенной здесь величине H_∞ следует из того, что при большом N почти все слагаемые в числителе дроби $\frac{H^{(N)}}{N}$ будут близки к

$$H_\infty = \lim_{N\to\infty} H_{\alpha_1\alpha_2 \dots \alpha_{N-1}}(\alpha_N);$$

исключение составят лишь первые слагаемые, вклад которых в общую сумму при очень большом N будет незначителен.

Таким образом, и последовательность «удельных энтропий» $h_N = \frac{H^{(N)}}{N}$ и последовательность «условных энтропий»

$$H_N = H_{\alpha_1\alpha_2 \dots \alpha_{N-1}}(\alpha_N)$$

сходятся при $N \to \infty$ к одному и тому же пределу H_∞. При этом $h_1 = H_1 = H(\alpha_1)$, но $H_N < h_N$ при $N > 1$ (так как h_N равно среднему арифметическому N чисел, лишь последнее из которых равно H_N, а все остальные больше); поэтому величины H_N, $N = 1, 2, 3 \dots$ будут заметно быстрее приближаться к предельному значению H_∞, чем величины h_N (ср. сноску на стр. 303).

Из результатов § 2 следует, что *среднее число элементарных сигналов, необходимое для передачи одной буквы русского текста, не может быть меньшим* $\frac{H_\infty}{\log m}$; с другой стороны, *возможно кодирование, при котором это среднее число будет сколь угодно близко к величине* $\frac{H_\infty}{\log m}$ (ср. выше, стр. 217). Разность $R = 1 - \frac{H_\infty}{H_0}$, показывающую, насколько меньше единицы отношение «предельной энтропии» H_∞ к величине $H_0 = \log n$, характеризующей наибольшую информацию, которая может содержаться в одной букве алфавита с данным числом букв, Шеннон назвал и з б ы т о ч н о с т ь ю языка (в рассматриваемом случае — русского). Данные, о которых мы будем говорить ниже, заставляют предполагать, что избыточность русского языка (как и избыточность других европейских языков) заметно превышает 50%. Говоря не совсем точно, мы можем сказать, что выбор следующей буквы осмысленного текста более, чем на 50% определяется самой структурой языка и, следовательно, случаен лишь в сравнительно небольшой степени. Именно избыточность языка позволяет сокращать телеграфный текст за счет отбрасывания некоторых легко отгадываемых слов (предлогов и союзов); она же позволяет легко восстановить истинный текст даже при наличии значительного числа ошибок в телеграмме или описок в книге.

Для того чтобы яснее представить себе смысл величины R, предположим, что русский текст кодируется при помощи 32-ичного кода, в котором элементарными сигналами служат те же русские буквы. Такой «код» будет представлять собой некоторый способ сокращенной записи русской речи при помощи обычных букв. В случае наиболее экономного кодирования для записи M-буквенного сообщения нам понадобится в среднем $\frac{H_\infty}{H_0} M = \frac{H_\infty}{5} M = (1-R) M$ элементарных сигналов(букв), т. е. по сравнению с обычной записью текст удается сократить на RM букв. Этот результат, разумеется, не означает, что мы можем произвольным образом отбросить RM букв и по оставшимся безошибочно восстановить исходное сообщение: для сокращения сообщения на RM букв необходимо

воспользоваться специальным «самым лучшим» методом кодирования, после применения которого все буквы сообщения становятся взаимно независимыми и равновероятными. Отсюда ясно, что закодированный текст при этом будет иметь тот же характер, что и «фраза» на стр. 237, т. е. будет казаться совершенно бессмысленным; «прочесть» такой текст будет много труднее, чем прочесть «фразу», приведенную в подстрочном примечании на стр. 239—240 (так как теперь кодовые обозначения сопоставляются уже не отдельным буквам, а сразу длинным «блокам»). Отметим еще, что при таком кодировании любая описка будет «роковой»: при декодировании она приведет к новому осмысленному тексту и мы ее не заметим, а если и заметим, то не сможем понять, что же было написано на самом деле. Что же касается до сокращения текста при помощи непосредственного пропуска части букв, выбранных наудачу, то заранее можно лишь утверждать, что при отбрасывании более чем RM букв мы заведомо не сможем безошибочно восстановить первоначальный текст. Специальные опыты (относящиеся к английскому языку) показывают, что обычно такое восстановление удается лишь тогда, когда число отброшенных букв не превосходит 25% от общего их числа.

Избыточность R является весьма важной статистической характеристикой языка; однако ее численное значение пока ни для одного языка не определено с удовлетворительной точностью. В отношении русского языка, в частности, как будто имеются лишь данные о значениях величин H_2 и H_3, полученные в Институте проблем передачи информации Академии наук СССР (см. Д. С. Лебедев, В. А. Гармаш [70]). В этой работе для нахождения относительных частот (т. е. приближенных значений вероятностей) всевозможных двухбуквенных и трехбуквенных сочетаний был использован отрывок из романа «Война и мир» Л. Н. Толстого, содержащий около 30 000 букв; подсчет числа повторений различных двухбуквенных и трехбуквенных комбинаций в этом отрывке осуществлялся с помощью счетно-аналитических машин. В результате были получены следующие значения (в битах):

H_0	H_1	H_2	H_3
$\log 32 = 5$	4,35	3,52	3,01

(для полноты мы здесь привели также и значения энтропий H_0 и H_1, указанные ранее на стр. 237 и 238). Строго говоря, отсюда можно только вывести, что для русского языка $R \geqslant 1 - \frac{H_3}{H_0} \approx 0{,}4$; естественно думать, однако, что на самом деле величина R значительно больше этого числа (энтропия H_3 равна средней информации, содержащейся в букве «фразы», приведенной на стр. 243, а эта «фраза» является заметно менее «упорядоченной», чем осмысленный русский текст). Последнее заключение подтверждается также и имеющимися в настоящее время (весьма неполными) данными об избыточности других языков.

Ясно, что для всех языков, использующих *латинский* алфавит, максимальная информация H_0, которая могла бы приходиться на одну букву текста, имеет одно и то же значение:

$$H_0 = \log 27 \approx 4{,}76 \text{ бит}$$

(латинский алфавит содержит 26 различных букв, к которым мы добавляем 27-ю «букву» — пустой промежуток между словами). Дальнейшие подсчеты, однако, должны производиться отдельно для каждого языка, так как частоты появления тех или иных букв или многобуквенных сочетаний не одинаковы в различных языках. Так, например, расположив все буквы в порядке убывания вероятностей (начиная с самой частой из них), мы придем к последовательности букв, начинающейся с — *ETAONRI*... в случае английского языка, с — *ENISTRAD*... в случае немецкого языка и с — *ESANITUR*... в случае французского (см. [73]; «—» во всех случаях обозначает пробел между словами); средняя длина слова, определяющая вероятность «пробела» в немецком языке заметно больше, чем в английском или французском; буквы *W* и *K* сравнительно часто встречаются в немецком и английском языках, но имеют практически нулевую вероятность во французском; сочетание *TH* очень распространено в английском языке, а сочетание *SCH* — в немецком, но в других языках эти сочетания весьма редки; за буквой *C* почти всегда следует буква *H* в немецком языке, но не в английском или во французском и т. д. Использовав таблицы относительных

частот различных букв в *английском*, *немецком*, *французском* и *испанском* языках, можно показать, что энтропия H_1 для этих языков равна (в битах):

язык	англ.	немецк.	франц.	испанск.
H_1	4,03	4,10	3,96	3,98

(ср. Б а р н а р д [74]). Мы видим, что во всех случаях величина H_1 заметно меньше, чем $H_0 = \log 27 \approx 4{,}76$ бит, причем ее значения для различных языков не очень сильно разнятся между собой.

Что же касается «условных энтропий» H_N (где $N > 1$), то они основательнее всего изучены для *английского* языка, которым мы, в основном, и ограничимся в дальнейшем. Величины H_2 и H_3 для этого языка были еще в 1951 г. подсчитаны Ш е н н о н о м [75]; при этом он использовал имеющиеся таблицы частот в английском языке различных двухбуквенных и трехбуквенных сочетаний. Учтя также и статистические данные о частотах появления различных с л о в в английском языке, Шеннон сумел приближенно оценить и значения величин H_8 и H_∞[1]).

---

[1]) Зная частоты (вероятности) $p_1, p_2, \ldots, p_K$ отдельных слов (здесь K — общее число слов, встречающихся в рассматриваемых текстах), можно определить «энтропию первого порядка» $H_1^{(\text{слова})} = -p_1 \log p_1 - p_2 \log p_2 - \ldots - p_K \log p_K$. Разделив полученную величину на среднее число w букв в слове, мы получим оценку для условной энтропии H_w порядка w. А именно, нетрудно понять, что $\frac{H_1^{(\text{слова})}}{w} < H_w$, ибо связи между w буквами одного слова заметно сильней связей между w произвольными последовательными буквами осмысленного текста. С другой стороны, отношение $\frac{H_1^{(\text{слова})}}{w}$, наверное, больше средней информации $H = H_\infty$, содержащейся в одной букве текста, поскольку величина $H_1^{(\text{слова})}$ совсем не учитывает зависимостей, существующих между словами (ср. ниже, стр. 263 и след.). [Ср., впрочем, работу В. Ю. У р б а х а [76], в которой критикуется методика Шеннона и приводятся несколько иные чем в [75] значения энтропий H_N (в работе [76] пробел между словами не включался в число букв, что, однако, само по себе учитывается очень просто — см. ниже, стр. 260 и след.).]

В результате он получил следующий ряд чисел:

H_0	H_1	H_2	H_3	H_5	H_8
4,76	4,03	3,32	3,10	$\approx$2,1	$\approx$1,9

Отсюда можно заключить, что для английского языка избыточность R во всяком случае не меньше, чем $1 - \frac{1,9}{4,76} \approx$ $\approx 0,6$, т. е., наверное, превосходит 60%.

Для более точной оценки величины R надо еще выяснить, насколько отличается величина H_8 — средняя информация, содержащаяся в букве текста при условии, что нам уже известны предыдущие 7 букв, от предельного значения H_∞. Иначе говоря, нас интересует вопрос о том, насколько существенно ограничивает произвол в выборе очередной буквы английского текста знание той части предшествующего текста, которая удалена от этой буквы более чем на 7 букв (при условии, что и последующие 7 букв нам также известны). Поскольку средняя длина английского слова равна всего лишь 4—5 буквам, т. е. заметно меньше 7 букв, то речь здесь может идти лишь о влиянии статистических закономерностей, определяющих порядок следования отдельных с л о в друг за другом (или даже еще более общих закономерностей, касающихся целых фраз). Непосредственное решение интересующего нас вопроса при помощи подсчета величин H_9, H_{10} и т. д. по приведенной на стр. 244 формуле невозможно, так как уже для вычисления H_8 требуется знание вероятностей всех 9-буквенных комбинаций, число которых выражается 13-значным числом (триллионы!). Поэтому для оценки величин H_N при больших значениях N приходится ограничиваться косвенными методами. На одном остроумном методе такого рода, предложенном Ш е н н о н о м [75], мы здесь вкратце остановимся.

«Условная энтропия» H_N представляет собой меру степени неопределенности опыта α_N, состоящего в определении N-й буквы текста, при условии, что предшествующие $N - 1$ букв нам известны. Естественно, что эта величина определяет степень трудности отгадывания N-й буквы по $N - 1$ предыдущим. Но эксперимент по отгадыванию N-й буквы легко может быть поставлен: для этого достаточно выбрать $(N - 1)$-буквенный отрывок осмысленного текста и предложить кому-либо отгадать

следующую букву [1]). Подобный опыт может быть повторен многократно; при этом трудность отгадывания N-й буквы может быть оценена с помощью среднего значения Q_N числа попыток, требующихся для нахождения правильного ответа. Ясно, что величины Q_N, определенные для разных значений N, являются определенными характеристиками статистической структуры языка, в частности, его и з б ы т о ч н о с т и: в случае нулевой избыточности знание сколь угодно длинного отрывка текста не увеличит вероятность правильно угадать следующую букву (эта вероятность во всех случаях будет равна $\frac{1}{n}$, где n — число букв алфавита); равенство же избыточности величине $\frac{1}{m}$ можно весьма огрубленно описать как утверждение о том, что каждая m-я буква текста является «лишней», однозначно восстанавливаемой по $m-1$ предыдущим.

Очевидно, что среднее число попыток Q_N с возрастанием N может только уменьшаться; прекращение этого уменьшения будет свидетельствовать о том, что соответствующие опыты имеют одинаковую степень неопределенности, т. е. что отвечающая им «условная энтропия» H_N практически уже достигла предельного значения H_∞. Исходя из этих соображений, Шеннон провел ряд подобных экспериментов, в которых N принимало значения 1, 2, 3, ..., 14, 15 и 100. При этом он обнаружил, что отгадывание 100-й буквы по 99 предшествующим является заметно более простой задачей, чем отгадывание 15-й буквы по 14 предыдущим. Отсюда можно сделать вывод, что H_{15} ощутимо больше, чем H_{100}, т. е. что H_{15} никак еще нельзя отождествить с предельным значением H_∞. Впоследствии такие же опыты были проведены на несколько большем материале Н. Б е р т о н о м и Дж. Л и к л а й д е р о м [77] для $N = 1, 2, 4, 8, 16, 32, 64, 128$ и $N \approx \approx 10\,000$; из их данных можно заключить, что величина H_{32} (так же как и H_{64} и H_{128}) практически не отличается от H_{10000}, в то время как «условная энтропия» H_{16} еще

[1]) Шеннон предлагает задавать вопросы ряду лиц и остановиться на том из них, ответы которого окажутся наиболее удачными, поскольку здесь считается, что отгадывание происходит н а и б о л е е р а ц и о н а л ь н ы м о б р а з о м, т. е. с полным знанием всех присущих языку статистических закономерностей.

заметно больше этой величины. Таким образом, можно предположить, что при возрастании N величина H_N убывает вплоть до значений N, имеющих порядок 30, но при дальнейшем росте N она уже практически не меняется; поэтому вместо «предельной энтропии» H_∞ можно говорить, например, об условной энтропии H_{30} или H_{40}.

Эксперименты по отгадыванию букв не только позволяют судить о сравнительной величине условных энтропий H_N при разных N, но дают также возможность оценить и сами значения H_N. Эта возможность связана с тем, что по данным таких экспериментов можно определить не только среднее число Q_N попыток, требующихся для отгадывания N-й буквы текста по $N-1$ предшествующим, но и вероятности (частоты) $q_N^1, q_N^2, \ldots, q_N^n$ того, что буква будет правильно угадана с 1-й, 2-й, 3-й, ..., n-й попытки (где $n = 27$ — число букв алфавита; очевидно, что $Q_N = q_N^1 \cdot 1 + q_N^2 \cdot 2 + \ldots + q_N^n \cdot n$). Нетрудно понять, что вероятности $q_1^1, q_1^2, \ldots, q_1^n$ равны вероятностям $p(a_1), p(a_2), \ldots, p(a_n)$ букв $a_1, a_2, \ldots, a_n$ алфавита, расположенных в порядке убывания частот. В самом деле, если ни одна из букв, предшествующих отгадываемой букве x, нам не известна, то естественно прежде всего предположить, что x совпадает с самой распространенной буквой a_1 (причем вероятность правильно угадать здесь будет равна $p(a_1)$); затем следует предположить, что x совпадает с a_2 (вероятность правильного ответа здесь будет равна $p(a_2)$) и т. д. Отсюда следует, что энтропия H_1 равна сумме

$$-q_1^1 \log q_1^1 - q_1^2 \log q_1^2 - \ldots - q_1^n \log q_1^n.$$

Если же $N > 1$, то можно показать, что сумма

$$-q_N^1 \log q_N^1 - q_N^2 \log q_N^2 - \ldots - q_N^n \log q_N^n \qquad (*)$$

не будет превосходить условную энтропию H_N (это связано с тем, что величины $q_N^1, q_N^2, \ldots, q_N^n$ представляют собой определенным образом усредненные вероятности исходов опыта α_N). С другой стороны, несколько более сложные соображения, на которых мы здесь не будем останавливаться, позволяют доказать, что сумма

$$(q_N^1 - q_N^2) \cdot \log 1 + 2(q_N^2 - q_N^3) \log 2 + \ldots$$
$$\ldots + (n-1)(q_N^{n-1} - q_N^n) \log (n-1) + n q_N^n \log n \qquad (**)$$

при всяком N будет не больше условной энтропии H_N. Таким образом, выражения (*) и (**) (составленные из вероятностей q_N^1, q_N^2, ..., q_N^n, которые можно оценить по данным эксперимента) определяют границы, между которыми должна заключаться величина H_N.

Надо только еще иметь в виду, что обе оценки (*) и (**) получаются в предположении, что q_N^1, q_N^2, ..., q_N^n — это те вероятности угадывания буквы по $N-1$ предыдущим буквам с первой, второй, третьей и т. д. попыток, которые получаются в предположении, что отгадывающий всегда называет очередную букву наиболее целесообразно — с полным учетом всех статистических закономерностей данного языка (ср. сноску на стр. 250). В случае же реальных опытов любые ошибки в стратегии отгадывающего (т. е. отличия называемых им букв от тех, которые следовало бы назвать, исходя из точной статистики языка) будут неизбежно приводить к завышению обеих сумм (*) и (**); именно поэтому целесообразно учитывать лишь данные «наиболее успешного отгадывающего», так как для него это завышение будет наименьшим [1]). Поскольку, однако, каждый отгадывающий иногда ошибается, то оценку (**) на практике нельзя считать вполне надежной оценкой снизу истинной энтропии (в отличие от оценки сверху (*), которая из-за ошибок отгадывающего может только стать еще больше).

Кроме того, значения сумм (*) и (**), к сожалению, не сближаются неограниченно при увеличении N (начиная с $N \approx 30$ эти суммы вообще перестают зависеть от N); поэтому полученные на этом пути оценки избыточности

---

[1]) Ясно, что бо́льшая или меньшая удачливость отгадывающего характеризует степень (обычно — интуитивного) понимания им статистических законов языка, т. е. присущее данному лицу «чувство языка» (или «чувство стиля» данного автора, тексты которого используются для предсказания букв,— ср. замечание в первой из работ [15] о «телепатической связи с авторами» одного из принимавших участие в подобных опытах лиц, видимо, обладавшего особо развитым литературным чутьем). В соответствии с этим предпринимались попытки использования результатов опытов по предсказанию букв для объективной характеристики степени владения иностранным языком ([78]; ср. также [79]) или родным языком (см. [80], где описаны опыты по отгадыванию букв сугубо специального текста несколькими группами лиц, имеющими весьма разную практику в чтении текстов подобного содержания).

языка не будут особенно точными [1]). В частности, опыты Шеннона [75] показали лишь, что величина H_{100}, по-видимому, заключается между 0,6 и 1,3 бит. Отсюда можно заключить, что избыточность

$$R = 1 - \frac{H_\infty}{H_0} \approx 1 - \frac{H_{100}}{\log 27}$$

для английского языка по порядку величины должна быть близка к 80%. Эксперименты Н. Бертона и Дж. Миклайдера [77] привели к близким результатам: по их данным истинное значение избыточности английского языка лежит где-то между 2/3 (т. е. 67%) и 4/5 (т. е. 80%).

Аналогичное (но несколько менее полное) исследование избыточности *немецкого* языка было выполнено известным немецким специалистом в области электросвязи К. Кюпфмюллером [82]. Использовав имеющиеся данные о частотах появления различных слогов и слов в немецком языке и произведя некоторые опыты по отгадыванию последующих слогов или слов немецкого текста по известному предшествующему отрывку, Кюпфмюллер пришел к выводу, что для немецкого языка $H_\infty \approx 1{,}3$ бит. Отсюда вытекает, что избыточность R этого языка близка к $1 - \frac{1{,}3}{4{,}76} \approx 0{,}7$ — значение, которое имеет тот же порядок величины, что и приведенные выше оценки избыточности для английского языка. Значение H_2 для немецкого языка может быть найдено, в частности, в работе [115], о которой подробнее мы будем говорить в разделе, посвященном устной речи.

Для *французского* языка наиболее полное исследование его энтропии и избыточности было выполнено Н. В. Петровой [83]. Для нахождения величин H_N, где N невелико, Петрова обработала совокупность текстов общим объемом в 30 000 букв, исходя из которой вычислялись вероятности (частоты) отдельных букв, а также их двухбуквенных и трехбуквенных сочетаний. При этом она пришла к следующим результатам:

H_0	H_1	H_2	H_3
$\log 27 \approx 4{,}76$	3,95	3,17	2,83

(ср. выше, стр. 249). Для определения величин H_N при больших N использовались опыты по отгадыванию букв, причем частично

---

[1]) Ср. работу А. П. Савчук [81], в которой сконструированы совершенно искусственные «языки», для которых шенноновские оценки (*) или, соответственно, (**) энтропии являются точными.

применялась предложенная А. Н. Колмогоровым методика, о которой мы еще скажем ниже; при этом было получено, что $H_\infty \approx 1{,}40$ бит, а следовательно, $R \approx 71\%$. Сходные результаты были получены также для *итальянского*, *шведского* и ряда других языков (см., например, [67], [84], [85]).

Разумеется, тот факт, что оценки избыточности нескольких европейских языков, использующих одинаковый алфавит, приводят к заключению, что эти величины имеют примерно одинаковые значения, не позволяет еще распространить этот вывод также и на очень далекие по лингвистической структуре языки или на языки, резко различающиеся своими алфавитами. В этой связи представляет интерес исследование Е. Ньюмана и Н. Во [86], попытавшихся сравнить энтропии H_N и избыточности R для трех языков с заметно различающимся числом букв в алфавите: полинезийского *языка Самоа*, алфавит которого содержит всего 16 букв (около 60% которых являются гласными), *английского* языка и *русского* языка, причем в последнем случае специально выбирались тексты, напечатанные по старой орфографии (принятой в России до 1917 г.), использующей 35-буквенный алфавит: кроме знакомых нам букв он содержал еще буквы ѣ («ять»; читалась как буква *е*), і («и десятеричное»; читалась как *и*), ѳ («фита», читалась как *ф*) и очень редко употребляемую букву $\vee$ («ижица», также читалась как *и*). Естественно, что величины H_0 для этих трех языков имеют очень разное значение (см. таблицу ниже). Еще более сильно различаются приведенные в той же таблице значения H_1 для трех языков, подсчитанные Ньюманом и Во на основе анализа одного и того же отрывка (длиной около 10 000 знаков) из трех переводов Библии. Грубо говоря, это означает, что распределения вероятностей отдельных букв в русском языке являются наиболее равномерными, а в языке Самоа — самым неравномерным (и значительной степени это объясняется тем, что в языке Самоа средняя длина слова очень мала: она составляет всего около 3,2 буквы против 4,1 буквы для английского языка и 5,3—для русского языка; поэтому пробел в языке Самоа имеет очень большую вероятность, в английском языке — меньшую и в русском — еще меньшую). Однако уже величины H_2 для трех языков оказываются более близкими, чем значения H_1: двухбуквенные связи в русском языке являются более жесткими, чем в английском и тем более — чем в языке Самоа.

К сожалению, оценки последующих значений H_N, указываемые Ньюманом и Во, являются менее надежными (они были получены авторами с помощью разработанной Е. Ньюманом и Л. Герстманом [87] методики, вызывающей известные сомнения); однако их выводы, касающиеся сравнительных значений H_N для трех языков, являются довольно правдоподобными:

	Самоа	Английский	Русский (старая орфография)
H_0	$\log 17 \approx 4{,}08$	$\log 27 \approx 4{,}76$	$\log 36 \approx 5{,}17$
H_1	3,40	4,08	4,55
H_2	2,68	3,23	3,44

Согласно этим выводам величины H_N убывают в русском языке быстрее всего, а в языке Самоа — медленнее всего; в результате, начиная примерно с $N = 10$, величины H_N (а следовательно,—и величины H_∞) для трех языков оказываются довольно близкими друг к другу. Это означает, что среднее количество информации, приходящейся на одну букву текста, для трех языков с заметно различающимся количеством букв алфавита оказывается примерно одинаковым. Если этот вывод является справедливым, то из него, разумеется, следует, что избыточность для языков с большим числом различных букв будет заметно большей, чем для языков с более бедным алфавитом.

Заметим также, что во всех европейских языках гласные буквы являются гораздо более частыми, чем согласные; это обстоятельство создает значительные различия в частотах отдельных букв, заметно отражающиеся уже на значении «энтропии 1-го порядка» H_1 (а также на «предельной энтропии» $H = H_\infty$ и на избыточности R) языка. Иным будет положение в ряде восточных языков, например, в *арабском* и *древнееврейском* (*иврит*): в этих языках гласные отсутствуют — они опускаются в письменном тексте и восстанавливаются читателем «по смыслу» (что возможно в силу избыточности языка). Ясно, что статистическая структура записанного на этих языках текста будет резко отличаться от той, с какой мы сталкиваемся в случае европейских языков, в силу чего и значения всех теоретико-информационных характеристик языка здесь могут принимать совсем другие значения (в частности, избыточность будет заметно меньше). В качестве иллюстрации к этому замечанию можно сослаться на работу немецкого лингвиста Г. Блюме [88], который сравнил статистические характеристики совокупностей трехбуквенных слов иврита и английского языка и нашел, что для этой совокупности

$$H_3^{(\text{ивр})} \approx 3{,}73 \text{ (бит/букву) и } R_3^{(\text{ивр})} = 1 - \frac{H_3}{H_0} \approx 0{,}16,$$

в то время как

$$H_3^{(\text{англ})} \approx 0{,}83 \text{ (бит/букву) и } R_3^{(\text{англ})} \approx 0{,}82.$$

Обстоятельно исследовался в 60-годах и вопрос об энтропии отдельных *индийских* языков, в первую очередь — распространенных в южной Индии *дравидских* языков, принадлежащих к числу древнейших на земле [89]; в этих работах, исходя из данных статистики языка (и с учетом введенной в [71] поправки), находились значения энтропий невысокого порядка, а также использовался «метод отгадывания» К. Шеннона, позволяющий оценить значения H_N, где N сравнительно велико. При этом новые — по сравнению с относящимися к европейским языкам работами — затруднения возникали здесь в силу некоторой неопределенности алфавитов большинства из рассматриваемых языков (ср. со сказанным ниже, стр. 265 и 278). Так, например, в языке *тамили* (исследованию которого посвящена работа Г. Сиромони — см. [89]) имеется старинный алфавит и современный алфавит; в современном

алфавите (близком к алфавитам ряда других индийских языков) имеется 12 гласных, 18 согласных, 216 слитных согласных-гласных и еще один не произносимый знак (Aitham) специального назначения. В работе Сиромони Aitham игнорировался вовсе, а «согласные-гласные» рассматривались как пары букв; однако такой подход к языку тамили не является единственно возможным. Кое что о конкретных результатах исследований, посвященных индийским языкам, будет еще сказано ниже (см. стр. 271).

Наконец, укажем, что различия в имеющихся оценках значения энтропии $H = H_\infty$ (или даже величин H_N, где N сравнительно невелико), найденных для разных европейских языков с помощью «метода отгадывания», являются, как правило, заметно меньшими, чем точность соответствующих оценок, определяемая различием между выражениями (*) и (**) для энтропии N-го порядка.

Таким образом, метод Шеннона оказывается явно недостаточным для определения различий в удельной энтропии (приходящейся на одну букву) для различных языков, хотя существование различия в средней длине слов для разных языков и различия в длине параллельных текстов на разных языках, имеющих одно и то же содержание (ср. Б. Рамакришна и Р. Субраманиан [90], а также последнюю из работ [89] [1]), создают впечатление, что эти различия в удельной энтропии вполне могут иметь порядок 10—20%. То же самое можно сказать и о различиях в энтропии текстов различного характера (в частности, принадлежащих различным авторам), написанных на одном и том же языке: представляется довольно очевидным, что различия эти могут быть довольно большими, — но и они могут быть обнаружены с помощью метода Шеннона только в самых крайних исключительных случаях (вроде того, к которому относятся работы Фрика и Самби или Фрица и Грайера, указанные на стр. 268).

---

[1]) Впрочем, указанные две работы на самом деле представляют интерес лишь с точки зрения постановки вопроса, но не с точки зрения полученных здесь конкретных результатов, так как для оценки «эффективности» различных языков здесь используется только сравнение относящихся к этим языкам «энтропий первого порядка» H_1, совершенно не учитывающих крайне важные для структуры языка статистические связи между последовательными буквами текста.

В этой связи представляются крайне желательным иметь более точные методы определения энтропии языка. А. Н. Колмогоров указал не так давно, что такие более точные методы могут быть сравнительно просто получены с помощью дальнейшего развития метода отгадывания. Прежде всего Колмогоровым было отмечено, что в принципе метод отгадывания (в предположении, что отгадывающий всегда будет следовать «оптимальной стратегии», вытекающей из полного учета всех присущих данному языку статистических закономерностей) позволяет получить не только оценки энтропии сверху и снизу, но и т о ч н у ю о ц е н к у значения этой величины. В самом деле, предположим, что отгадывающему предлагается каждый раз не перечислять по порядку те буквы, которые, как он думает, должны появиться, а сразу назвать все условные вероятности p_1^N, p_2^N, ..., p_n^N того, что появится 1-я, 2-я, ..., n-я из букв алфавита (при условии, что предшествующие $N-1$ букв текста ему известны). Пусть теперь этот опыт повторяется много раз и каждый раз подсчитывается величина $-\log p_k^N$, где k — порядковый номер той буквы, которая появилась на самом деле (таким образом, в каждом отдельном опыте из названных «отгадывающим» n чисел p_1^N, ..., p_n^N, где n — число букв алфавита, на самом деле учитывается лишь одно, но какое именно — заранее неизвестно). Тогда нетрудно показать, что если условные вероятности всегда будут указываться точно, то среднее значение подсчитываемой величины — $\log p_k^N$ (т. е. сумма всех таких величин, определенных в большом числе M опытов, деленная на M) при неограниченном увеличении M будет неограниченно приближаться к истинной энтропии H_N одной буквы текста.

Разумеется, этот метод является совершенно непрактичным: немыслимо требовать от отгадывающего, чтобы он каждый раз указывал весь набор условных вероятностей всевозможных букв — и при этом никогда не ошибался. Существенно, однако, что любые ошибки в названных значениях условных вероятностей приведут лишь к в о з р а с т а н и ю соответствующей суммы значений $-\log p_k^N$ (это обстоятельство, как нетрудно показать, следует из неравенства (*) на стр. 251). Поэтому вполне допустимо заранее ограничить множество распределений

вероятностей, которые может называть отгадывающий, и тем существенно облегчить его работу; при этом сумма полученных таким образом значений $-\log p_k^N$, разделенная на число M опытов, все равно будет оценкой сверху истинной энтропии H_N.

В реальных опытах, проводившихся под руководством Колмогорова над русскими литературными текстами, отгадывающему позволялось делать следующие предсказания (см. [91]):

1) следующей буквой наверное будет одна определенная (скажем, k-я) буква алфавита;

2) следующей буквой наверное будет одна из указываемых отгадывающим двух или трех букв алфавита;

3) следующей буквой вероятно (но не наверное!) будет одна определенная (скажем, k-я) буква алфавита;

4) следующей буквой вероятно будет одна из указываемых отгадывающим двух или трех букв;

5) кроме того, отгадывающему позволялось сказать, что он не знает, какой будет следующая буква.

При этом считалось, что каждое из этих утверждений равносильно выбору следующего условного распределения вероятностей для последующей буквы текста:

1) k-я буква имеет некоторую заранее фиксированную большую вероятность P; для i-й же буквы, где $i \neq k$, вероятность появиться принимается равной $p_i' = p_i \cdot \frac{1-P}{1-p_k}$, где p_i и p_k — безусловные вероятности i-й и k-й букв русского языка, указанные в таблице на стр. 238;

2) выбранные две или три буквы имеют одинаковую условную вероятность $P/2$ или $P/3$; остальные буквы по-прежнему имеют вероятности p_i', пропорциональные их безусловным вероятностям p_i;

3) k-я буква имеет некоторую фиксированную вероятность Q (меньшую, чем P!), а i-я буква при $i \neq k$ имеет вероятность $p_i' = p_i \cdot \frac{1-Q}{1-p_k}$;

4) выбранные две или три буквы имеют одинаковую вероятность $Q/2$ или $Q/3$, а остальные буквы — вероятности, пропорциональные их безусловным вероятностям;

5) условная вероятность появления i-й буквы алфавита при всех i принимается равной ее безусловной вероятности p_i.

Вероятности P и Q пока остаются неопределенными; так как, однако, любая неточность в предсказываемых условных распределениях вероятностей может лишь увеличить получаемую оценку для H_N, то вполне допустимо подобрать эти две вероятности по известным результатам опытов так, чтобы сумма всех величин $-\log p_k^N$ (где p_k^N — предсказанная условная вероятность реально появившейся буквы) была возможно меньшей.

Нетрудно подсчитать, что при таком определении вероятностей P и Q окончательная оценка энтропии H_N будет даваться формулой

$$H_N \approx \frac{1}{M}[M_1 h_1 + M_2 h_2 + M_1' + M_2' \log 3 + S],$$

где M — общее число опытов; M_1 — число предсказаний типа 1) или 2); M_2 — число предсказаний типа 3) или 4); M_1' — число предсказаний типа 2) или 4), в которых предсказывается одна из двух букв; M_2' — число предсказаний типа 2) или 4), в которых предсказывается одна из трех букв; $h_1 = q_1 \log q_1 - (1 - q_1) \log (1 - q_1)$, где $q_1 = \frac{m_1}{M_1}$, а m_1 — число ошибок в предсказаниях типов 1) и 2); $h_2 = -q_2 \log q_2 - (1 - q_2) \log (1 - q_2)$, где $q_2 = \frac{m_2}{M_2}$ — средняя доля ошибок в предсказаниях типов 3) и 4); наконец, S — распространенная по всем случаям ошибок в предсказаниях типов 1) — 4) и всем «отказам» (предсказаниям типа 5)) сумма выражений $-\log p_i'$, где p_i' — или «безусловная вероятность» p_i реально появившейся буквы (в случае предсказаний типа 5)), или же «предсказанная вероятность» p_i', разделенная на $1 - P$ (в случае предсказаний типов 1 и 2)), или, наконец, она же, разделенная на $1 - Q$ (в случае предсказаний типов 3) и 4)).

Выписанная здесь формула на первый взгляд кажется сравнительно сложной, но на практике она оказывается довольно удобной и приводит к не слишком громоздким расчетам. Опыты подобного рода, проводившиеся в статистической лаборатории Московского государственного университета, позволили получить для классической русской прозы С. Т. Аксакова («Детские годы Багрова-внука») и

И. А. Гончарова («Литературный вечер»), оценку удельной энтропии H_∞ (не отличающейся, например, от H_{50}) порядка 1—1,2 бит, являющуюся, по-видимому, довольно точной (вероятно, превышающую истинное значение H_∞ не больше, чем на 10—15%). Соответственно этому для избыточности литературного языка русской классической прозы отсюда получается значение, по порядку величины близкое к 80%.

Напомним, что во всем предыдущем к числу «букв» мы причисляли и пустой *промежуток между словами*, что совершенно естественно с точки зрения телеграфии. Иногда, однако, представляет интерес также рассмотрение обычного алфавита, не учитывающего пробелов; так, например, может встать вопрос об информации, содержащейся в одной напечатанной букве текста. Естественно, что при этом приведенные выше результаты претерпят некоторые изменения. Так, например, русский алфавит теперь придется считать 31-буквенным (буквы ь и ъ мы отождествляем по-прежнему), так что $H_0 = \log 31 \approx 4{,}95$ бит; частоты отдельных букв также изменят свои значения (см. таблицу этих частот в книге А. А. Харкевича [69]), что приводит к новому значению энтропии H_1, а именно $H_1 \approx 4{,}46$ бит. Латинский алфавит при таком рассмотрении надо будет считать 26-буквенным, так что для всех языков, использующих этот алфавит, $H_0 = \log 26 \approx 4{,}7$ бит. Значения (в битах) энтропий H_1, H_2 и H_3, а также приближенные значения энтропий H_5 и H_8 для английского языка, полученные в пренебрежении пробелами между словами, приведены в следующей таблице (ср. Шеннон [75]):

H_0	H_1	H_2	H_3	H_5	H_8
4,70	4,14	3,56	3,3	$\approx 2{,}6$	$\approx 2{,}3$

Сравнив эту таблицу с приведенной на стр. 249, мы убедимся, что учет пробелов между словами в английском языке приводит к увеличению энтропии H_0 и уменьшению всех последующих энтропий H_N. То, что для всех языков $H_0^{(\text{с проб})} > H_0^{(\text{без проб})}$ совершенно очевидно: ведь всегда $\log n > \log (n-1)$. Далее, учет пробела приводит к появлению дополнительной «буквы», имеющей

сравнительно с другими очень большую вероятность, что облегчает предсказание исхода опыта α_1, а следовательно, уменьшает степень его неопределенности H_1. Аналогично объясняется уменьшение H_N при учете пробела и для других значений N. В частности, при достаточно большом N (превышающем среднюю длину слова) исход опыта, состоящего в определении N-й буквы текста по известным $N-1$ предшествующим буквам во всех тех случаях, когда этой N-й буквой оказывается «пробел», будет практически однозначно определяться самой структурой языка (легко понять, что при большом N ошибки при отгадывании исхода этого опыта обычно будут иметь место лишь тогда, когда N-я буква оказывается первой или, в крайнем случае, второй буквой нового слова). Отсюда вытекает, что учет пробела заметно уменьшает неопределенность этого опыта, и, значит, $H_N^{(\text{с проб})} < H_N^{(\text{без проб})}$.

Можно получить даже точную зависимость, связывающую два значения избыточности R — вычисленное при условии пренебрежения пробелами между словами и при учете этих пробелов. В самом деле, рассмотрим два одинаковых достаточно длинных текста, отличающихся лишь тем, что в одном из них мы не отмечаем промежутков между словами. Каждый из текстов однозначно восстанавливается по другому: разумеется, мы можем отбросить все промежутки между словами в обычном тексте и почти столь же просто восстановить пробелы в написанном «вплотную» (без интервалов между словами) тексте на знакомом языке. Отсюда можно заключить, что «полная информация» (произведение «удельной информации» или «информации, приходящейся на одну букву текста» H_∞, на число букв), содержащаяся в том и другом тексте, должна быть одной и той же. А так как число «букв» в тексте с пробелами превосходит число букв написанного «вплотную» текста в $\frac{s+1}{s}$ раз, где s — средняя длина слова (ибо в среднем один пробел приходится на s букв текста), то

$$H_\infty^{(\text{с проб})} = H_\infty^{(\text{без проб})} : \frac{s+1}{s}.$$

Учитывая еще, что вероятность p_0 пробела равна $\frac{1}{s+1}$ (один пробел приходится на $s+1$ «букв» текста с

пробелами) и, следовательно, $s = \frac{1}{p_0} - 1$, мы можем переписать эту формулу так [1]):

$$H_\infty^{(\text{с проб})} = H_\infty^{(\text{без проб})} : \frac{\frac{1}{p_0}}{\frac{1}{p_0} - 1}$$

или

$$H_\infty^{(\text{с проб})} = (1 - p_0)\, H_\infty^{(\text{без проб})}.$$

Но если общее число букв алфавита (включая пробел) равно n, то $H_0^{(\text{с проб})} = \log n$, $H_0^{(\text{без проб})} = \log (n - 1)$ и

$$\frac{H_\infty^{(\text{с проб})}}{H_0^{(\text{с проб})}} = \frac{H_\infty^{(\text{без проб})}}{H_0^{(\text{без проб})}} \cdot (1 - p_0) : \frac{\log n}{\log (n-1)}$$

или

$$(1 - R^{(\text{с проб})}) = (1 - R^{(\text{без проб})}) \cdot (1 - p_0) \frac{\log (n-1)}{\log n}.$$

Это и есть формула, связывающая значения избыточности языка, полученные при пренебрежении пробелами и при учете пробелов.

---

[1]) Последний результат можно весьма просто доказать и не ссылаясь на постоянство «полной информации». В самом деле, пусть α_N — опыт, состоящий в отгадывании N-й «буквы» текста с пробелами между словами по $N - 1$ предшествующим буквам. Выяснение исхода α_N мы будем производить в два этапа: прежде всего проверим, не является ли N-й «буквой» пробел (опыт β); если же это не так, то мы дополнительно выясним, какая именно эта буква (опыт α'_N). Если p_0 — вероятность пробела, то второй опыт α'_N нам, очевидно, придется производить лишь в $(1 - p_0)$-й части всех случаев. Отсюда вытекает, что

$$H(\alpha_N) = H(\beta) + (1 - p_0)\, H(\alpha'_N),$$

где $H(\alpha_N)$, $H(\alpha'_N)$, $H(\beta)$ — средние у с л о в н ы е энтропии соответствующих опытов при условии, что нам известны $N - 1$ предшествующих букв (ср. с § 4 гл. II). А так как при большом N можно считать, что $H(\beta) = 0$ (пробел восстанавливается по предшествующим $N - 1$ буквам однозначно) и $H(\alpha_N) = H_\infty^{(\text{с проб})}$, $H(\alpha'_N) = H_\infty^{(\text{без проб})}$, то мы получаем

$$H_\infty^{(\text{с проб})} = (1 - p_0)\, H_\infty^{(\text{без проб})}.$$

Сходные соображения могут быть использованы и для определения среднего количества информации $H_\infty^{(\text{слова})}$, содержащейся в одном *слове* текста. Энтропию нулевого порядка одного слова $H_0^{(\text{слова})} = \log K$ можно оценить, подсчитав число K слов в каком-либо достаточно полном словаре данного языка; энтропию $H_1^{(\text{слова})} = -p_1 \log p_1 - p_2 \log p_2 - \ldots - p_K \log p_K$ можно подсчитать с помощью «частотного словаря», указывающего частоты (вероятности) $p_1, p_2, \ldots, p_K$ отдельных слов [1]). Однако непосредственное вычисление «условной энтропии первого порядка» $H_2^{(\text{слова})}$ требует уже знания частот всевозможных сочетаний из двух слов, определить которые практически невозможно, так как общее число таких сочетаний громадно. Еще менее перспективна задача вычисления последующих «условных энтропий» $H_3^{(\text{слова})}$, $H_4^{(\text{слова})}$ и т. д. При этом надо иметь в виду, что статистические связи между отдельными словами зачастую являются заметно более жесткими, чем связи между буквами (появление в тексте слова «дифференциальный» сильнее ограничивает вероятности следующих за ним слов, чем, скажем, появление буквы «г» — вероятности последующих букв) и что связи эти заметно более «дальнодействующие» (появление в начале сколь угодно толстой книги слова «лемма» резко уменьшает вероятность встретить слово «любовь» в ее конце). Все это делает вопрос об определении «предельной энтропии» («удельной информации») $H_\infty^{(\text{слова})}$ как будто бы чрезвычайно трудным.

Сопоставим теперь друг другу два текста — написанный обычным образом с помощью букв и «иероглифический», в котором за единую «букву» принимается целое слово (иероглифическая письменность как раз и характеризуется тем, что в ней отдельные знаки обозначают целые слова). При этом каждый из двух текстов, разумеется, однозначно, восстанавливается по другому — зная все буквы какого-либо текста, мы знаем тем самым и все входящие в него слова, а знание всех слов равносильно знанию буквенной записи. Поэтому и здесь «полная информация»,

---

[1]) Вроде известного словаря Торндайка, о котором мы говорили на стр. 87 (см. также [92] и другие статьи на тему о частотных словарях, напечатанные в том же сборнике).

содержащаяся в двух текстах, будет одна и та же, т. е. $H_\infty^{(\text{слова})}\cdot$число слов текста $= H_\infty^{(\text{буквы})}\cdot$число букв текста. А так как отношение числа букв к числу слов равно средней длине слова, то, следовательно,

$$H_\infty^{(\text{слова})} = H_\infty^{(\text{без проб})}\cdot s \quad \text{или} \quad H_\infty^{(\text{слова})} = H_\infty^{(\text{с проб})}\cdot(s+1),$$

где s — средняя длина слова (и, значит, $s+1$ — среднее число приходящихся на одно слово «букв», к числу которых причисляется также и пробел между словами).

Из последней формулы вытекает соотношение

$$\frac{H_\infty^{(\text{слова})}}{H_0^{(\text{слова})}} = \frac{H_\infty^{(\text{буквы})}}{H_0^{(\text{буквы})}}\cdot(s+1):\frac{\log K}{\log n}$$

или

$$(1-R^{(\text{слова})}) = (1-R^{(\text{буквы})})\cdot(s+1)\ \frac{\log n}{\log K},$$

где, как и выше, s — средняя длина слова, K — общее число слов, встречающихся в рассматриваемых текстах, n — число «букв» алфавита, к которым причисляется и пробел между словами; под $H^{(\text{буквы})}$ и $R^{(\text{буквы})}$ здесь, как и почти всюду выше, понимается $H^{(\text{с проб})}$ и $R^{(\text{с проб})}$. В частности, для русского языка мы имеем $n = 32$ и $s+1 = \frac{1}{p_0} = \frac{1}{0{,}175} \approx 5{,}7$; положив $K = 50\,000$ (таково примерное число слов в довольно полных словарях) [1]), мы получим

$$(1-R^{(\text{слова})}) = (1-R^{(\text{буквы})})\cdot 5{,}7\,\frac{\log 32}{\log 50\,000} \approx$$
$$\approx 1{,}85\,(1-R^{(\text{буквы})}).$$

Таким образом, мы видим, что избыточность для слов заметно меньше избыточности для букв, т. е. что «иероглифическая» письменность в известном смысле является

---

[1]) Так как число слов K входит в предыдущую формулу под знаком логарифма, то неточность определения этого числа лишь незначительно отражается на результате (если положить $K = 100\,000$, то множитель 1,85 в нижеследующей формуле заменится на 1,74).

более «выгодной», чем буквенная. Это обстоятельство тесно связано с выгодой кодирования сразу длинных блоков из большого числа «букв», о которой много говорится в этой главе; слова как раз и являются подобными «блоками» (причем «блоками», вероятности появления которых сравнительно высоки).

Ясно, что сходные соображения позволяют также связать отнесенные к одной букве текста значения энтропии (информации) $H = H_\infty$ и избыточности R с теми же величинами, определенными для какой-либо другой лингвистической единицы (*слога*, *фразы*, *морфемы* и т. д.; ср. со сказанным на стр. 280 относительно *фонем*). Это обстоятельство поясняет причины, в силу которых подавляющее большинство теоретико-информационных исследований языка исходит из буквенного его алфавита: связь отнесенных к одной букве, слогу, слову и т. д. значений энтропии позволяет ограничиться рассмотрением какой-либо одной из этих величин; с другой стороны, буквенный алфавит обладает преимуществами привычности, однозначной определенности (ибо для большинства других лингвистических единиц вроде слога, морфемы или даже слова не существует точных определений, не допускающих разных толкований самого определяемого понятия) и ограниченности (поскольку «алфавит» слов или, тем более, фраз языка является практически необъятным).

Укажем еще, что связь между значениями $H^{(\text{буквы})}$ и $H^{(\text{слова})}$ может быть использована двояким образом: она позволяет свести определение величины $H^{(\text{слова})}$ к (предполагаемой известной) величине $H^{(\text{буквы})}$; с другой стороны, эти же соображения позволяют оценить энтропию $H^{(\text{буквы})}$, опираясь на полученные тем или иным способом приближенные значения $H^{(\text{слова})}$. Приближенное значение $H^{(\text{слова})}$ (точнее говоря, значение энтропии первого порядка $H_1^{(\text{слова})}$) можно вычислить, например, воспользовавшись так называемым законом Ципфа, утверждающим, что *при упорядочивании слов языка в порядке их частот* (т. е. вероятностей) *частота n-го по порядку слова для всех не слишком больших значений n оказывается примерно пропорциональной* $1/n$. Этот закон был сформулирован и проверен на большом лингвистическом материале в книге Дж. Ципфа [93]; в дальнейшем

он многократно обсуждался и уточнялся целым рядом авторов [1]). Широко обсуждается, в частности, закон Ципфа в гл. 5 и 12 книги [3], в ч. I книги [94] и в статьях [95] — [96], где, в частности, воспроизведены заимствованные из книги [93] графики, демонстрирующие приложимость закона Ципфа к текстам, написанным на разных языках и имеющих разный характер (скажем — к тексту романа «Улисс» Дж. Джойса и к «среднестатистическому» американскому газетному тексту). Первые применения закона Ципфа к определению энтропии слова (и оценке, исходя отсюда, также и энтропии одной буквы) указал еще Шеннон [75]; дальнейшие относящиеся сюда данные могут быть найдены в статьях Е. Ньюмана и Л. Герстмана [87], Дж. Миллера [95] и М. Григнетти [97].

Приближенная оценка энтропии первого порядка $H_1^{(\text{слова})}$ по формуле $H_1^{(\text{слова})} = -p_1 \log p_1 - p_2 \log p_2 - \ldots - p_K \log p_K$ была получена (в применении к *румынскому* языку) И. Воинеску, А. Фрадисом и Л. Михайлеску (см. третью из работ [118]). Фактически, однако, эта работа посвящена энтропии не письменной, а устной речи (частоты $p_1, p_2, \ldots, p_K$ здесь определялись из анализа магнитофонной записи ответов на длинную серию стандартных вопросов десяти разных испытуемых); поэтому более уместно о ней говорить в следующем разделе нашей книги (см. ниже, стр. 279). Заметим, кроме того, что основная цель исследования Воинеску и др. заключалась вовсе не в определении величины $H_1^{(\text{слова})}$ для обычного румынского языка, а в сравнении значений $H_1^{(\text{слова})}$, отвечающих речи здоровых людей, с соответствующими значениями, отвечающими речи других десяти испытуемых, больных афазией (т. е. расстройством речи вследствие частичного поражения головного мозга). Поэтому оно примыкает и к исследованиям статистических характеристик «специальных языков», к рассмотрению которых мы теперь и перейдем.

---

[1]) Так, еще сам Ципф заметил, что в некоторых случаях более точно считать, что частота n-го слова на самом деле пропорциональна $1/n^a$, где постоянная a близка к единице, но все же не равна точно единице (см. по этому поводу также работы [94], [96]).

Данные об энтропии одной буквы текста, о которых речь шла выше, относились, как правило, к «среднему литературному языку», поскольку материалом для экспериментов по определению энтропии служили чаще всего литературные тексты: так А. Н. Колмогоров и его сотрудники использовали произведения С. Т. Аксакова и И. А. Гончарова (см. стр. 259—260), а К. Шеннон [75], работавший в сотрудничестве со своей женой Бетти Шеннон, анализировал отрывки из книги Дюма Малона «Вирджинец Джефферсон». Но на стр. 237—238 уже указывалось, что частоты появления различных букв могут зависеть от характера рассматриваемого текста; точно так же и значения энтропий H_N или избыточности R будут различными для текстов, заимствованных из разных источников. При этом любой «специальный язык» (например, научный или технический текст по определенной специальности, деловая переписка, какой-либо жаргон) будет, как правило, иметь избыточность выше средней из-за меньшего количества употребляемых слов и наличия часто повторяющихся специальных терминов и оборотов — весьма благоприятное обстоятельство, очень облегчающее просмотр научной литературы по определенной специальности или чтение такой литературы на недостаточно знакомом языке. Исключение в этом отношении могут представлять жаргоны, специально преследующие своей целью уменьшение избыточности языка, например, воровской жаргон, на котором весьма краткие выражения могут иногда заменять длинные и содержательные фразы, или некоторые научные жаргоны с широко разработанной терминологией, вроде того, который употребляется в математике французской школой Бурбаки [1]); еще более яркий пример в этом направлении доставляет символический язык современной математической логики, характеризующийся исключительной смысловой насыщенностью.

Вопрос о влиянии характера текста на значения энтропии и избыточности, приходящиеся на одну букву текста, исследовался ленинградским лингвистом Р. Г. Пиотровским и его] учениками, в частности — Н. В. Петровой, сравнившими теоретико-информационные характеристики разных типов русской

---

[1]) Более общедоступный пример разобран в статье [80], о которой мы уже говорили выше (стр. 252).

и французской речи (см. [67], [83], [98]) [1]). При этом в соответствии со сказанным выше избыточность «деловых» текстов оказалась заметно больше «средней» избыточности языка и избыточности литературных текстов. В противоположность этому избыточность разговорной речи, найденная в этих работах, оказалась немного ниже средней — видимо, в первую очередь в силу большей «вольности» разговорной речи, меньшей стесненности ее правилами стилистики и даже просто грамматики. Полученные Р. Г. Пиотровским и его группой результаты собраны в следующей таблице:

	$H = H_\infty$ (в бит/букву)		R (в процентах)	
	русск. яз.	франц. яз.	русск. яз.	франц. яз.
Язык в целом	1,37	1,40	72,6	70,6
Разговорная речь	1,40	1,50	72,0	68,4
Литературные тексты	1,19	1,38	76,2	71,0
Деловые тексты	0,83	1,22	83,4	74,4

Более частный характер имеет исследование ленинградцев О. Л. Смирнова и А. В. Екимова [99], изучавших характеристики случайно взятой выборки телеграфных текстов объемом в 15 000 букв; при этом использовался метод угадывания Шеннона (и идущее от А. Н. Колмогорова его уточнение; ср. стр. 256 и след.). Полученный Смирновым и Екимовым результат: $H^{(\text{телегр. русск. языка})} \approx 1{,}4 \cdot H^{(\text{литературн. русск. языка})}$ естественно связан с намеренным уменьшением избыточности телеграфных текстов (например, за счет пропуска союзов).

Другой «специальный язык» — язык переговоров по радио между дежурным на аэродроме и пилотами находящихся в воздухе самолетов, был изучен американскими учеными Ф. Фриком и У. Самби [100], а также Э. Фрицем и Дж. Грайером [101]. Естественно, что рассматриваемые в этих работах переговоры очень стандартны по своей форме и ограничиваются несколькими постоянно повторяющимися узкими темами. Неудивительно поэтому, что избыточность соответствующей речи (оцениваемая или с помощью «опытов по отгадыванию», или же с помощью непосредственного изучения статистики небольшого числа стандартных обо-

[1]) Р. Г. Пиотровский и Н. В. Петрова использовали метод угадывания Шеннона в уточненном А. Н. Колмогоровым его варианте (ср. стр. 257—259). Указываемые этими авторами значения избыточности R ниже приведены в соответствие с взятыми из тех же работ значениями H.

ротов, из которых складываются эти переговоры) оказалась заметно превышающей избыточность «литературного языка». В частности, ограничившись еще более узким классом сообщений, передаваемых дежурным пилоту приземляющегося в определенных условиях самолета, Фрик и Самби получили для избыточности значение, близкое к 96% (почти то же значение избыточности, близкое к 93%, можно получить, исходя из результатов Фрица и Грайера). Столь большая избыточность здесь имеет вполне ясные основания — из-за наличия значительных помех (связанных с создаваемым самолетами шумом) меньшая избыточность могла бы привести к ошибкам при приеме, могущим в рассматриваемом случае иметь самые тяжелые (даже трагические) последствия.

Высокая избыточность, характерная для любого «специального языка», учитывается, например, при составлении кодов для деловой переписки больших американских фирм. В настоящее время такие коды разрабатываются с непременным участием специалистов по теории информации, и наличие в ведущихся фирмой переговорах частых повторений отдельных слов и целых оборотов позволяет весьма значительно повысить экономность кодов.

Со сказанным связан также интересный, но пока мало изученный вопрос о различиях в избыточности языка разных литературных текстов. Можно предполагать, что разные литературные жанры отличаются разной избыточностью, связанной с присущей именно этому типу произведений манерой изложения; можно думать также, что и внутри одного литературного произведения в разных отрывках (диалог, описание и т. д.) избыточность будет разной. Высокая избыточность может характеризовать избитый, шаблонный язык литературного произведения, но может также служить лишь свидетельством неторопливой манеры автора (так, высокая избыточность была обнаружена в упомянутых на стр. 259—260 экспериментах по определению энтропии одной буквы текста в «Литературном вечере» И. А. Гончарова, написанном спокойным, плавным языком, характеризующимся большим числом достаточно естественных подробностей). Низкая избыточность может служить свидетельством богатства и яркости (неожиданности, нестандартности) литературной речи (возможно, здесь примером может служить язык У. Фолкнера) — однако слишком низкая избыточность языка литературного произведения неизбежно будет восприниматься как нарочитая усложненность речи. Еще более низкую избыточность будет иметь «заумь» типа той, которую употреблял русский поэт В. Хлебников (напомним, что нулевая избыточность характеризует

приведенную на стр. 237 «фразу», которую вряд ли можно считать отличающейся «хорошей» литературной формой).

Примыкает сюда и широко обсуждавшийся в 60-х годах вопрос о сравнении избыточности прозаической и поэтической речи (см. [102] — [104] и ряд статей в сборнике [105]; ср. также включенные в ссылку [84] статьи Л. Долежеля и Е. Николау, К. Сала, А. Рочерик). Ясно, что поэтическая форма (ритм, рифма) накладывает на язык некоторые дополнительные ограничения, т. е. повышает его избыточность. Можно даже пытаться оценить численно, скажем, влияние ритма стиха, определив количество словосочетаний, удовлетворяющих заданной ритмической схеме, и сравнив его со всем богатством словосочетаний; удобно при этом исходить из словаря, определенного по прозаическим произведениям того же автора [1]). Несколько сложнее учесть влияние рифмы, но и здесь вполне возможны грубые оценки. Ориентировочные оценки, проведенные А. Н. Колмогоровым для классического русского четырехстопного ямба (этим стихом написан, например, «Евгений Онегин» А. С. Пушкина [2])), показали, что выполнение требований, накладываемых поэтической формой, снижает «неопределенность» H_1 одной буквы текста на довольно значительную величину, порядок которой сравним с половиной величины H_1, подсчитанной для «среднелитературного» текста. И в самом деле, проведенный А. Н. Колмогоровым опыт по угадыванию последующих букв показал, что для «плохого» стиха, в котором уменьшение содержащейся в одной букве информации не компенсируется свойственными «хорошим» стихам повышенной эмоциональностью, яркостью речи и богатством словаря, «предельная информация» H_∞, приходящаяся на одну букву текста, существенно (примерно вдвое) меньше величины H_∞, определенной для класси-

[1]) См., например, работу А. М. Кондратова [103], в которой подсчитывается энтропия невысокого порядка, определяемая ритмической схемой русского стихотворного и прозаического (научного, делового, художественного и разговорного) текста (в битах/слог); ср. также статью Г. Людтке (H. Lüdtke) «Сравнение метрических схем в отношении их избыточности» в сборнике [105].

[2]) Четырехстопный ямб характеризуется строфой, теоретически состоящей из восьми правильно чередующихся ударных и безударных слогов (на практике некоторые ударения иногда выпадают).

ческой русской прозы [1]). Однако в произведениях больших поэтов уменьшение информационной насыщенности одной буквы текста, связанное с соблюдением известных формальных правил, по-видимому, в очень большой степени компенсируется повышенной яркостью и нестандартностью речи, так что вполне можно ожидать, что здесь избыточность языка имеет тот же порядок, что и избыточность прозаических литературных текстов.

Обсуждению влияния различных связанных с литературным стилем факторов на значение энтропии и избыточности речи посвящена работа У. Пейсли [106] (в которой, к сожалению, использовалась не особенно надежная методика Е. Ньюмана и Л. Герстмана [87] и Е. Ньюмана и Н. Во [86]). Пейсли проанализировал 39 разных отрывков английского текста и сравнивал между собой энтропии: а) двух стихотворных переводов «Илиады», принадлежащих разным авторам; б) четырех переводов двух различных отрывков из той же «Илиады», а также четырех (современных) переводов двух отрывков из одной главы евангелия от Матвея (в обоих случаях выбирались заметно отличающиеся по содержанию отрывки); в) четырех прозаических и четырех стихотворных переводов «Илиады», г) девяти разных переводов евангелия от Матвея, относящихся к разным эпохам. В ряде случаев обнаруженная У. Пейсли разница между значениями энтропии, приходящейся на одну букву текста, оказалась ощутимой, причем здесь можно было заметить даже некоторые общие закономерности (вроде уменьшения избыточности литературных текстов с приближением времени написания к современному); впрочем, все эти выводы еще нуждаются в дополнительной проверке.

Близкий характер имеют упоминавшиеся выше исследования [89], посвященные ряду индийских языков; в этих работах также сопоставляются значения энтропии, вычисленные для текстов разного характера (например, прозаических и поэтических) и разного времени написания. Некоторые из полученных в работах [89] результатов определенным образом перекликаются с полученными У. Пейсли на материале английского языка; впрочем, сопоставление здесь затрудняется существенно разными алфавитами английской и индийских письменностей (ср. со сказанным на стр. 255—256).

Из работ, более непосредственно связанных с сопоставлением прозаической и поэтической речи (вопрос, не обойденный вниманием также в статьях [106] и [89]), в первую очередь назовем исследования Л. Долежела и Е. Николау, К. Сала, А. Рочерик (см. [84]), подсчитавших энтропии разных порядков для прозаической и поэтической чешской и румынской речи и даже для отдельных прозаиков и поэтов; впрочем, полученные этими авторами предварительные оценки явно еще нуждаются в

---

[1]) Сопоставлялись «Поединок» А. И. Куприна и напечатанное на обороте одного из листков отрывного календаря стихотворение весьма скромного литературного достоинства.

уточнении. С. Маркус [104] предпринял рискованную попытку перенести в поэтику связи между понятиями «энтропии» и «энергии»; на этой базе он рассмотрел некоторые содержащиеся в работе Е. Николау, К. Сала и А. Рочерик результаты, касающиеся подсчета энтропии для произведений Эминеску, относящихся к различным периодам творчества поэта. Более частный характер имеет работа Т. Тарноци [102], в которой подсчитан ряд теоретико-информационных характеристик венгерской прозы и поэзии.

Укажем в заключение, что само применение к (уникальным по самому определению!) литературным текстам стандартных теоретико-информационных представлений, возникших в связи с чисто прикладными задачами техники связи и игнорирующих вопрос о смысловом содержании передаваемого сообщения, а базирующихся лишь на чисто статистических понятиях (типа частот букв в «статистическом ансамбле» некоторого «среднего текста»; однако какое содержание можно вложить в понятие «статистического ансамбля» стихов А. С. Пушкина?), вызывало и вызывает известные сомнения. Для А. Н. Колмогорова (см. [15]) эти соображения] послужили поводом для широкой постановки вопроса о возможности разных подходов к самому понятию «количества информации» и для пропаганды «чисто комбинаторного» подхода к этому понятию, в частности, в применении к изучению энтропии языка и, особенно, литературных текстов.

Сущность комбинаторного подхода к определению энтропии заключается в следующем. *Комбинаторную энтропию* H приходящуюся на одну букву текста, можно определить условием, что для n-буквенного алфавита число N-буквенных текстов (где N достаточно велико), удовлетворяющих заданным статистическим ограничениям, равно не $n^N = 2^{\log n \cdot N}$ $(= 2^{H_0 N})$, как было бы, если бы мы имели право брать любые наборы из N последовательных букв, а всего лишь $M = 2^{HN}$ (ср. стр. 82—83 и 225). В соответствии с этим, владея понятием «осмысленного» текста, мы можем определить энтропию H как

$$H_{\text{комб}} = \lim_{N \to \infty} \left(\frac{1}{N} \log M(N) \right),$$

где $M(N)$ есть число всевозможных осмысленных текстов длины N; это последнее определение уже не зависит ни от каких теоретико-вероятностных представлений.

Пытаясь численно оценить значение «комбинаторной энтропии» $H_{\text{комб}}$, число $M(N)$ можно оценивать с помощью подсчета числа возможных продолжений текста. А именно, пусть $*$ — «пустое» слово, вовсе не содержащее букв; далее через $l(*a_1a_2 \ldots a_k)$ (или через $l(a_1a_2 \ldots a_k)$, где $a_1, a_2, \ldots, a_k$ — некоторые буквы рассматриваемого языка) обозначим число всевозможных «осмысленных продолжений» последовательности букв $a_1a_2 \ldots a_k$, т. е. число таких букв x, что отрывок $a_1a_2 \ldots a_kx$ может быть продолжен до осмысленного текста. В таком случае значение

$$\bar{M}(n) = l(*)l(*a_1)l(*a_1a_2) \ldots l(*a_1a_2 \ldots a_{N-1}),$$

осредненное по ряду цепочек букв, можно рассматривать как оценку интересующей нас величины M (N).

Сказанное намечает путь к чисто комбинаторным расчетам энтропии и избыточности «грамматически правильного» текста. Первые попытки такого рода были выполнены А. Н. Колмогоровым и его сотрудниками (см. первую из работ [15]); при этом число возможных продолжений текста здесь определялось по списку слов, включенных в «Словарь русского языка» С. И. Ожегова. Полученная при этом оценка $H = (1{,}9 \pm 0{,}1)$ бит/букву, естественно, заметно превышает указанные на стр. 260 оценки энтропии «литературных текстов» (так как «степень неопределенности» буквы литературного текста ограничена отнюдь не одними лишь требованиями грамматической правильности). К сожалению, более подробное описание этих исследований, а также результатов аналогичных исследований, начатых в Ленинграде Р. А. Зайдманом, пока не опубликовано.

Устная речь

Перейдем теперь к затронутому уже на стр. 268—269 вопросу об энтропии и информации устной речи. Естественно думать, что все статистические характеристики такой речи будут еще более зависеть от выбора разговаривающих лиц и от характера их разговора, чем это наблюдалось в случае речи письменной — ведь письменная речь, как правило, является более «сглаженной», чем устная. И хотя по данным Р. Г. Пиотровского и его сотрудников «в среднем» энтропия устной речи несколько выше энтропии письменных текстов, для некоторых типов устной речи (см., скажем, пример в конце стр. 268) это будет безусловно не так. Пониженное значение энтропии устной речи может быть связано с тем, что в разговоре мы зачастую употребляем больше повторений одних и тех же слов (меньше заботимся о «красоте стиля») и нередко добавляем довольно много «лишних» (т. е. не несущих содержательной информации) слов — это делается как для облегчения восприятия речи, так и просто затем, чтобы говорящий имел время обдумать, что он хочет сказать дальше. В частности, очень высока избыточность разговоров при высоком уровне помех (например, в гудящем самолете, вагоне электрички или в метро), а также разговоров пьяниц, упрямо повторяющих одни и те же (как правило, далекие от «высокой» литературы) слова и выражения — последнее связано с тем,

что в этом случае затруднено и само произнесение речи, а не только ее восприятие.

Определив среднее число букв, произносимых за единицу времени, можно приближенно оценить количество информации, сообщаемое при разговоре за 1 *сек*; обычно оно, по-видимому, имеет порядок 5—6 бит (это количество информации, естественно, сильно зависит от «скорости разговора», которая может меняться весьма значительно: «очень быстрая» речь почти в 5 раз скорее «очень медленной» [1])). Эти данные согласуются с данными физиологической акустики, позволяющими оценить общее число произнесенных человеком в единицу времени «различимых звуков» (ср. обзор Дж. Миллера [95]).

Однако эта оценка скорости передачи информации при разговоре относится лишь к «смысловой информации», которую можно извлечь и из записи сказанных слов. На самом деле живая речь всегда содержит, кроме того, еще довольно значительную дополнительную информацию, которую говорящий сообщает нам иногда добровольно, а иногда и прямо против своего желания; эта дополнительная информация может и противоречить «смысловой информации», причем в таких случаях она, как правило, заслуживает большего доверия. Так, из разговора мы можем судить о настроении говорящего и об его отношении к сказанному; мы можем узнать говорящего, если даже никакие другие источники информации (включая сюда и «смысловую информацию») не указывают нам его; мы можем во многих случаях определить место рождения незнакомого нам человека по его произношению (последнее обстоятельство играет основную роль в завязке действия пьесы Б. Шоу «Пигмалион»); мы можем оценить громкость устной речи, которая в случае передачи голоса по линии связи (телефон, радио) во многом определяется чисто техническими характеристиками линии передачи, и т. д. Количественная оценка всей этой информации представляет собой очень сложную задачу, требующую значитель-

---

[1]) Мы не говорим здесь, разумеется, о разговорах с особо высокой избыточностью, типа обсуждавшихся выше: так, в случае переговоров между пилотом и дежурным на аэродроме скорость передачи информации не превосходит 0,2 бит/сек, т. е. намного меньше, чем для самого медленного разговора на общие темы.

но бо́льших знаний об языке, чем имеется в настоящее время; в частности, здесь нужны весьма обширные и разнообразные статистические данные, пока почти полностью отсутствующие.

Исключением в этом отношении является сравнительно узкий вопрос о логических ударениях, подчеркивающих в фразе отдельные слова; эти ударения также несут определенную информационную нагрузку, которую (для частного случая разговоров по телефону) можно оценить количественно. Необходимые для этого статистические данные были получены английским связистом Дж. Берри [107], проанализировавшим ряд «типичных английских телефонных разговоров»; в частности, согласно данным Берри ударение чаще всего падает на наиболее редко употребляемые слова (что, впрочем, довольно естественно — ясно, что вряд ли кто будет выделять логическим ударением наиболее распространенные слова — например, предлоги или союзы). Если вероятность того, что данное слово W_r находится под ударением, мы обозначим через q_r, то средняя информация, заключающаяся в сведениях о наличии или отсутствии ударения на этом слове, будет равна

$$-q_r \log q_r - (1 - q_r) \log (1 - q_r).$$

Пусть теперь $p_1, p_2, \ldots, p_K$ — вероятности (частоты) всех слов $W_1, W_2, \ldots, W_K$ (здесь K — общее число всех употребляемых слов; вероятности $p_1, p_2, \ldots p_K$, играющие основную роль во всех статистических теориях языка, приводятся в так называемых «частотных словарях» — ср. выше, стр. 263). В таком случае для средней информации H, заключенной в логическом ударении, можно написать следующую формулу:

$$H = p_1 [-q_1 \log q_1 - (1 - q_1) \log (1 - q_1)] + \\ + p_2 [- q_2 \log q_2 - (1 - q_2) \log (1 - q_2)] + \ldots \\ \ldots + p_K [-q_K \log q_K - (1 - q_K) \log (1 - q_K)].$$

Подставив сюда данные Берри, французский ученый Б. Мандельброт [108] подсчитал, что средняя информация, которую мы получаем, выяснив, на какие слова падает логическое ударение, по порядку величины близка к 0,65 бит/слово.

Что же касается всей вообще разнообразной «несмысловой» информации, содержащейся в устной речи, то

имеющиеся данные позволяют дать лишь весьма грубую и несовершенную оценку ее суммарной величины. Такая оценка была получена немецким связистом Кюпфмюллером в интересном исследовании [82] об устной и письменной немецкой речи, о котором мы уже упоминали раньше. В своей работе Кюпфмюллер и не пытался учитывать сложные статистические закономерности интонаций, тонов голоса и других особенностей речи; по существу он ограничился лишь «нулевой энтропией» H_0, связанной с числом различных возможностей, а затем грубо ориентировочно принял соответствующую избыточность равной 50%. Наряду с информацией, содержащейся в интонации, Кюпфмюллер отдельно оценил информацию, связанную с индивидуальными особенностями голоса говорящего человека, а также информацию, доставляемую громкостью речи; сумма трех полученных при этом величин сопоставлялась с содержащейся в той же речи «смысловой информацией». Для оценки общего числа распознаваемых степеней громкости и общего числа «речевых мелодий» (типов интонации, определяемых небольшими изменениями основной частоты голосовых колебаний) были привлечены данные физиологической акустики [1]; общее число различаемых человеком индивидуальных голосов определялось, так сказать, «на глаз». Естественно, что найденные на этом пути оценки «общего числа возможных исходов» не могут претендовать на особенно большую точность; однако, так как информация определяется логарифмом этого числа, то даже грубые оценки позволяют подсчитать информацию с весьма приличной точностью (ведь при общем числе возможных исходов порядка 1 000 для того, чтобы преувеличить информацию в два раза, пришлось бы преувеличить это число возможностей в 1000 раз!). Подобные подсчеты привели Кюпфмюллера к выводу, что дополнительная информация, содержащаяся в интонации, громкости и особенностях индивидуального голоса при нормальном разговоре не должна

[1]) Может показаться, что громкость и интонация могут меняться непрерывным образом, так что здесь должно иметься бесконечно много разных возможностей. На самом деле, однако, человеческое ухо различает лишь конечное число разных степеней громкости и конечное число интонаций; подробнее об этом мы еще будем говорить ниже (см. стр. 290 и след.).

превосходить 75% от «смысловой информации»; при очень быстром разговоре она составляет не более 30% от смысловой информации, а при очень медленном — не более 150% (существенное различие этих чисел частично может объясняться тем, что при быстром разговоре мы можем распознать заметно меньше разных голосов и меньше различаем интонацию) [1]).

В работе Кюпфмюллера указаны также «удельная» энтропия и информация устной речи, отнесенные к одной произнесенной букве. Фактически, однако, эти цифры имеют лишь условный характер (они нужны только для сравнения устной речи с письменной); в действительности же во время разговора отдельные буквы никогда не произносятся, а произносятся звуки, существенно отличающиеся от букв. Поэтому основным элементом устной речи (в том же смысле, в каком буква является основным элементом письменной речи) надо считать отдельный звук — ф о н е м у. Осмысленная устная речь составляется из фонем точно так же, как осмысленная письменная речь составляется из букв; при передаче устной речи по линии связи мы должны только проследить, чтобы все фонемы передавались правильно — тогда и смысл всей речи будет передан правильно, т. е. никакая часть «смысловой информации» не будет потеряна. Поэтому во всех случаях, когда нас интересует лишь передача «смысловой информации» устной речи (а таких случаев—большинство), наибольший интерес представляют не энтропия и информация одной «произнесенной буквы» (являющейся чисто условным понятием), а энтропия и информация одной реально произнесенной фонемы.

Список фонем данного языка, разумеется, не совпадает со списком букв алфавита. Общее число фонем заметно превышает число букв, так как одна и та же буква в разных случаях может звучать по-разному (например, произношение гласной существенно зависит от того, находится

---

[1]) По-видимому, это обстоятельство связано с тем, что ведущие от органов слуха к головному мозгу первные каналы могут пропускать за определенное время лишь строго определенное количество информации (см. ниже, стр. 318—320). Поэтому увеличение скорости передачи «смысловой информации» неизбежно влечет за собой уменьшение скорости передачи по тем же каналам информации другого типа.

ли она под ударением или не находится; одна и та же согласная может произноситься и твердо и мягко и т. д.). При этом приходится иметь в виду, что если даже в отношении числа букв алфавита возможны разные точки зрения (так, например, не совсем ясно, следует ли считать *е* и *ё* или *и* и *й* одной или двумя буквами русского алфавита; далее, в относящихся к телеграфии исследованиях обычно принимают ь и ъ за одну букву, что, разумеется, вовсе не всегда можно считать оправданным[1]), то в отношении «алфавита фонем», по поводу самого определения которых (см., например, В. А. Успенский [109]) лингвисты пока не пришли к согласию, расхождения между отдельными авторами являются неизбежными. В частности, американские ученые Е. К. Черри, М. Халле и Р. Якобсон [110] (ссылающиеся на ряд авторитетных советских лингвистов) выделили в *русском* языке 42 различные фонемы и подсчитали частоты отдельных фонем (а также различных комбинаций двух и трех следующих друг за другом фонем), воспользовавшись, в основном, довольно старыми и неполными данными известного русского филолога А. М. Пешковского [111][2]). Исходя из этих данных, они определили значения «максимальной возможной энтропии» $H_0 = \log 42$ одной фонемы, энтропии первого порядка $H_1 = -p_1 \log p_1 - p_2 \log p_2 - \ldots - p_{42} \log p_{42}$ (где $p_1, p_2, \ldots, p_{42}$ — относительные частоты различных фонем) и «условных энтропий» H_2 и H_3 (определяемых в точности так же, как и для письменной речи). Полученные результаты (в битах) сведены в следующую таблицу:

H_0	H_1	H_2	H_3
$\log 42 \approx 5{,}38$	4,77	3,62	0,70

[1]) Это отождествление смазывает существенное различие между частотой букв ь и ъ (первая из которых встречается в тексте много чаще, чем вторая; напротив, при использовании «старой орфографии», принятой в нашей стране до 1917 г., буква ъ оказывается гораздо более частой, чем ь).

[2]) Гораздо более широкое исследование частот отдельных фонем и их парных комбинаций (проведенное на обширном современном материале) было выполнено на кафедре фонетики Ленинградского государственного университета (см. Л. Р. Зиндер [112]); в этом исследовании общее число фонем было принято равным 48 (в первую очередь за счет более детального разграничения гласных звуков).

Поучительно сравнить эти значения с приведенными на стр. 246 значениями величин H_0, H_1, H_2 и H_3 для письменной русской речи (ср. также стр. 254 и др.). Сравнение показывает, что если только данные работы [110] являются обоснованными [1]), то убывание ряда условных энтропий для фонем происходит заметно быстрее, чем в случае букв письменного текста.

В отношении *английской* устной речи некоторые результаты были получены Дж. Блэком и П. Дьенешем (см. [113]). Первый из этих авторов подсчитал энтропии H_0, H_1 и H_2 одной фонемы по статистическим данным, относящимся к совокупности одно- и двусложных английских слов (которая, разумеется, не характеризует еще весь английский язык в целом), причем число фонем он считал равным 41. Второй автор определил относительные частоты фонем и всех их парных сочетаний («диграмм» по данным, относящимся к «среднему английскому языку», и приняв число фонем равным 45 (энтропия H_1 одной диграммы, следующая из данных Дьенеша, приведена в работе [115]). Родственные статистические результаты о фонемах и парах фонем *французской* речи опубликованы Ж. Гатоном и М. Ламоттом [114]. Немецкий ученый В. Эндрес [115] попытался оценить суммарную избыточность одной фонемы *немецкой* и *английской* речи, воспользовавшись спектрограммами фонем (дающими представление фонемы в виде некоторой фигуры на плоскости) и применив затем методы приближенного определения избыточности соответствующих рисунков, родственные использованным в заключительной части работы [135] (о которой см. ниже, стр. 303 и след.) для оценки избыточности изображений букв в машинописном тексте; согласно его данным для обоих языков избыточность фонем имеет порядок 80 — 85% (т. е. близка к избыточности букв письменной речи). Изучению энтропий низких порядков в устной *румынской* речи (и сравнению полученных данных с теми, которые относятся к письменной

---

[1]) К сожалению, в работе [110] не указан точно объем материала использованного для определения частот различных фонем и их двойных и тройных сочетаний. Поэтому можно опасаться, что значение H_3 оказалось сильно заниженным из-за недостаточности статистических данных (ср. ниже сноску на стр. 289).

речи) посвящена работа А. Фрадиса, Л. Михайлеску и И. Воинеску [116]; относящиеся к *татарскому* языку данные имеются в работе Т. И. Ибрагимова [117]. Отметим, наконец, работы И. Воинеску, А. Фрадиса и Л. Михайлеску [118], посвященные сравнению теоретико-информационных характеристик (энтропий H_1 и H_2 одной фонемы, разностей $H_0 - H_1$, а также энтропии $H_1^{(\text{слова})}$; см. выше стр. 263) устной речи здоровых людей и людей больных афазией (т. е. расстройством речи). При этом оказалось, что для речи больных афазией энтропии H_1, H_2 и $H_1^{(\text{слова})}$ все принимают заметно меньшие значения, чем для речи здоровых людей (т. е. избыточность речи здесь заметно повышается), а кроме того указанные энтропии еще, как правило, и гораздо сильнее меняются при переходе от одного больного к другому, чем при переходе от одного здорового человека к другому (особенно резкий характер приобретают указанные явления в применении к величине $H_1^{(\text{слова})}$, существенно зависящей от объема словаря говорящего и от степени равномерности использования им слов этого словаря).

С помощью соображений, использованных нами выше для определения избыточности $R^{(\text{слова})}$, можно также установить связь между избыточностями устной и письменной речи. Из того, что устная речь может быть записана, а письменная — прочитана, следует, что «полная информация», содержащаяся в определенном тексте [1]), не зависит от того, в какой форме — устной или письменной — этот текст представлен, т. е. что

$$H_\infty^{(\text{буквы})} \cdot \text{число букв} = H_\infty^{(\text{фонемы})} \cdot \text{число фонем}$$

(ср. выше, стр. 264). Отсюда вытекает, что

$$H_\infty^{(\text{фонемы})} = H_\infty^{(\text{буквы})} \cdot \omega,$$

где ω есть среднее число букв, приходящихся на одну фонему («средняя длина фонемы»); эта величина является важной статистической характеристикой языка,

[1]) Разумеется, в случае устной речи здесь учитывается лишь содержащаяся в ней «смысловая» информация (ср. выше, стр. 274).

связывающей устную и письменную речь. Из последней формулы следует также, что (ср. стр. 262 и 264)

$$\frac{H_{\infty}^{(\text{фонемы})}}{H_{0}^{(\text{фонемы})}} = \frac{H_{\infty}^{(\text{буквы})}}{H_{0}^{(\text{буквы})}} \cdot \omega : \frac{\log k}{\log n}$$

или

$$(1 - R^{(\text{фонемы})}) = (1 - R^{(\text{буквы})}) \cdot \omega \frac{\log n}{\log k},$$

где k — общее число фонем, а n — число букв; за $R^{(\text{буквы})}$ здесь естественнее принимать $R^{(\text{без проб})}$. Однако использование этой формулы затрудняется отсутствием статистических данных, позволяющих определить величину ω (даже по поводу числа фонем k мы не имеем пока единодушного мнения филологов) [1]).

Музыка

Исследования того же рода могут быть проведены и в отношении музыкальных сообщений. Естественно думать, что связи между последовательными звуками некоторой мелодии, выражающимися отдельными нотными знаками, достаточно сильны: так как одни сочетания звуков будут более благозвучны, чем другие, то первые будут встречаться в музыкальных произведениях чаще вторых. Если мы выпишем ряд нот наудачу, то информация, содержащаяся в каждой ноте этой записи, будет наибольшей; однако с музыкальной точки зрения такая хаотическая последовательность нот не будет представлять никакой ценности. Для того чтобы получить приятное на слух звучание, необходимо внести в наш ряд определенную избыточность; при этом, однако, можно опасаться, что в случае слишком большой избыточности, при которой последующие ноты уже почти однозначно

---

[1]) Сопоставив фонемы английской речи 43-м фонетическим знакам, употребляющимся в англо-русских словарях, мы сможем приближенно определить «среднюю длину фонемы» ω из сравнения длины буквенной записи английских слов и их фонетической транскрипции. При этом получается $\omega \approx 1{,}2$, что даст

$$(1 - R^{(\text{фонемы})}) = (1 - R^{(\text{буквы})}) \cdot 1{,}2 \frac{\log 27}{\log 43} \approx 1{,}05\,(1 - R^{(\text{буквы})}).$$

определяются предшествующими, мы получим лишь крайне монотонную и малоинтересную музыку. Какова же та избыточность, при которой может получиться «хорошая» музыка?

Весьма правдоподобно, что избыточность простых мелодий никак не меньше, чем избыточность осмысленной речи; представляло бы интерес специально изучить вопрос об избыточности различных форм музыкальных произведений или произведений различных композиторов. К сожалению, в настоящее время мы имеем еще мало конкретных данных такого рода. Одни из первых результатов в этом направлении были получены в 1956 г. американским ученым Р. Пинкертоном [119], проанализировавшим с точки зрения теории информации популярный в Америке альбом детских песенок. Для простоты в этой работе предполагалось, что все звуки находятся в пределах одной октавы; так как к тому же в рассматриваемых мелодиях не встречались так называемые хроматизмы, то все эти мелодии могли быть приведены к семи основным звукам: *до*, *ре*, *ми*, *фа*, *соль*, *ля* и *си* (которым на фортепьяно соответствуют белые клавиши). Все анализируемые песенки записывались как последовательности этих «основных элементов», каждый длительностью в одну восьмую; учет звуков, длительностью более одной восьмой, осуществлялся с помощью добавления к семи нотам восьмого «основного элемента» O, обозначающего продление предшествующего звука еще на промежуток времени в одну восьмую (или же паузу в одну восьмую). Таким образом, «максимальная возможная энтропия» H_0 одной ноты здесь равна

$$H_0 = \log 8 = 3 \text{ бита}.$$

Подсчитав частоты (вероятности) отдельных нот во всех 39 анализируемых песенках, Пинкертон нашел, что

$$H_1 = -p(O)\log p(O) - p(до)\log p(до) - p(ре)\log p(ре) - \\ - p(ми)\log p(ми) - p(фа)\log p(фа) - p(соль)\log p(соль) - \\ - p(ля)\log p(ля) - p(си)\log p(си) \approx 2{,}73 \text{ бит};$$

здесь, например, $p(до)$ означает вероятность ноты *до*. Воспользовавшись найденными Пинкертоном вероятностями сочетаний из двух нот, можно подсчитать также условную энтропию H_2; она оказывается близкой к 2,42 бит (впрочем, в статье Пинкертона указываются лишь

определенным образом осредненные вероятности двухнотных сочетаний, так что полученное значение H_2 является завышенным). Разумеется, по одним только значениям H_1 и H_2 еще очень мало что можно сказать о степени избыточности рассматриваемых мелодий (можно лишь сказать, что, по-видимому, она заметно выше, чем $1 - \frac{2{,}42}{3} \approx 0{,}2$). Некоторые косвенные данные, подтверждающие этот вывод, будут приведены ниже.

Еще до появления работы Пинкертона на конференции по теории информации в Лондоне (осень 1955 г.) было сообщено о работе Ф. и К. Аттнив, подсчитавших частоты отдельных нот и двунотных комбинаций в ряде американских ковбойских песен. Значительно более детальное исследование такого рода было выполнено в 1957 г. в лаборатории вычислительных машин Гарвардского университета (см. Ф. Брукс и др. [120]). Здесь были проанализированы отрывки из 37 гимнов различных композиторов и эпох, имеющих одну и ту же метрическую структуру. Применение быстродействующей электронной вычислительной машины позволило авторам отказаться от упрощения, заключающегося в отнесении всех нот к одной и той же октаве; в качестве различных «основных элементов» здесь рассматривались все ноты четырех октав хроматической гаммы (включающей также и пять промежуточных звуков, соответствующих черным клавишам фортепьяно) — всего 49 различных элементов, не считая специальных обозначений для звуков, продолжающихся из предыдущего временно́го интервала. За единицу длительности одного основного элемента была снова выбрана длительность в одну восьмую, так как более короткие ноты ни в одном из рассматриваемых гимнов не встречались.

С помощью современных средств вычислительной техники Брукс и др. подсчитали частоты всех отдельных «основных элементов», всех комбинаций из двух соседних таких элементов, из трех элементов и т. д., вплоть до комбинаций из восьми соседних элементов включительно. Полученные результаты в принципе дают возможность написать приближенные выражения для всех условных энтропий от H_0, H_1, H_2 и до H_8 включительно. Правда, при этом надо иметь в виду, что использованный статистический материал (состоящий из 37 небольших отрывков

из различных гимнов) заведомо недостаточен для получения сколько-нибудь надежных оценок вероятностей сочетаний из большого числа нот; поэтому найденные на этом пути значения энтропий высоких порядков (энтропий H_6, H_7 и H_8, во всяком случае) были бы очень мало обоснованы. Тем не менее уже и значения первых нескольких условных энтропий могут иметь определенный интерес; поэтому можно только пожалеть, что авторы работы [120] не произвели таких вычислений (и не привели никаких данных, позволяющих как-нибудь оценить соответствующие энтропии).

Аналогичный анализ мелодий известного американского сочинителя песен Стефана Фостера (1826—1864) был выполнен (правда, в более скромных размерах) Г. Олсоном и Г. Беларом [121]. Эти авторы рассмотрели 11 наиболее популярных песен Фостера и, положив в основу музыкальную шкалу из 12 разных нот (охватывающих полторы октавы), подсчитали частоты (т. е. эмпирические значения вероятностей) каждой отдельной ноты и всевозможных групп из двух и трех последовательных нот. Ясно, что исходя из полученных данных можно без труда оценить также и условные энтропии H_0, H_1, H_2 и H_3 одной ноты в песенках Фостера (хотя это и не было сделано в статье [121]). Дальнейшие сведения об исследованиях статистических закономерностей музыкальных произведений могут быть найдены в книге Р. Х. Зарипова [122], содержащей обширную библиографию.

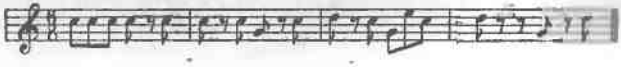
Примеры непосредственного вычисления теоретико-информационных характеристик различных музыкальных произведений имеются в статьях Дж. Юнгблада [123], Дж. Коэна [124], Г. Сиромони и К. Р. Раджагопалана [125], Л. Хиллера и Дж. Бишема [126], М. Ролландя [127] и некоторых других (см. также обзор этого направления в гл. 13 книги [3]). Так, например, в статье [124] (в которой использованы также и результаты Юнгблада и Броули) значения энтропий H_1 и H_2 и соответствующих избыточностей $R_1 = 1 - \frac{H_1}{\log n}$ и $R_2 = 1 - \frac{H_2}{\log n}$ первых двух порядков, отнесенные к одной ноте, вычислены (и сравниваются между собой) на материале музыки отдельных ком-

позиторов-романтиков XIX в. (Шуберта, Мендельсона, Шумана) и всей немецкой романтической музыки XIX в. в целом, а также на материале католических религиозных гимнов и современного американского рок-энд-рола. В статье [127] сравниваются значения избыточности для классической музыки Гайдна и модернистской музыки Шенберга (естественно, что у Шенберга избыточность оказалась меньшей, чем у Гайдна). В работе [126] приведены некоторые результаты анализа одного из произведений близкого к Шенбергу композитора Веберна, а в [125] подсчитаны значения H_1 для ряда произведений южно-индийской музыки XVIII—XIX столетий. В статьях [124] и [126] приводятся также и некоторые данные относительно «ритмической избыточности» различных музыкальных произведений (аналогичной избыточности «стихотворных ритмов» в поэзии). Однако до сих пор все полученные оценки информационных характеристик музыкальных произведений все же должны рассматриваться как предварительные и методы их вычисления требуют еще дальнейшего обсуждения (об этом говорится, в частности, в заключительной части статьи [124]).

Заметим также, что основная цель статистических подсчетов различных вероятностей, описывающих музыкальную структуру, во многих случаях состояла вовсе не в определении энтропии и избыточности. Дело в том, что высокая степень избыточности хорошей музыки позволяет дать совсем другое, довольно неожиданное, применение статистическим таблицам, задающим вероятности и условные вероятности различных нот. Для того чтобы подойти к этому применению, вспомним приведенные на стр. 237, 240, 242, 243 и 244 «модели русских фраз» — последовательности букв русского алфавита, в которых в большей или меньшей степени учитывались имеющиеся в русском языке внутренние связи между соседними буквами. Мы видели, что чем дальше простирались те зависимости, которые учитывались при составлении наших фраз, тем «более русскими» становились эти фразы, т. е. тем более приближались они по звучанию к обычной русской речи. Ясно, однако, что вряд ли можно надеяться получить на этом пути полностью осмысленные выражения — всегда в наших фразах будет иметься некоторый элемент случайности, путающий их смысл. Попробуем теперь применить

эти же методы к музыке. При этом мы будем получать «музыкальные фразы» — последовательности нот, все более и более близкие по своей статистической структуре к тем источникам, которые использовались для вычисления частот различных нот и их комбинаций. Как и в случае «моделей русских фраз», эти новые «музыкальные фразы» не будут точно повторять ни одну из последовательностей, положенных в основу при вычислении частот; однако в то время как в случае языка это обстоятельство делает наши «фразы» бессмысленными, в случае музыки именно оно делает их заслуживающими внимания — ведь они будут представлять собой новые, оригинальные музыкальные произведения!

Разумеется, трудно сказать заранее, насколько интересными могут быть такие «модели музыкальных мелодий»; неясно также, сколь глубокие связи должны быть учтены для получения сочетаний, близких «по духу» к исходному материалу (т. е., например, имитирующих произведения определенного жанра или определенного автора). Существенно заметить, однако, что в силу значительной избыточности музыки мы уже на одном из ранних шагов описанного на стр. 237 и след. процесса можем прийти к достаточно гармоничным звучаниям. Это было убедительно показано еще в первых чисто любительских экспериментах, произведенных Пинкертоном [119]. В этих экспериментах учитывались только вероятности отдельных нот и двухнотных комбинаций, которые к тому же очень сильно округлялись; для нахождения последовательных нот «искусственных музыкальных фраз» использовалось последовательное извлечение бумажек с записанными на них двумя нотами из нескольких «урн» (точнее говоря, просто кучек), каждая из которых содержала всего 12 бумажек, или даже еще более простая и грубая процедура. Накладывая, кроме того, дополнительные связи, обеспечивающие сохранение определенного ритма «музыкальных фраз», Пинкертон смог получить несколько новых мелодий, которые, по утверждению автора, иногда не уступали мелодиям исходных детских песенок из использованного им альбома. Запись одной из таких «случайно получившихся» мелодий приведена ниже:

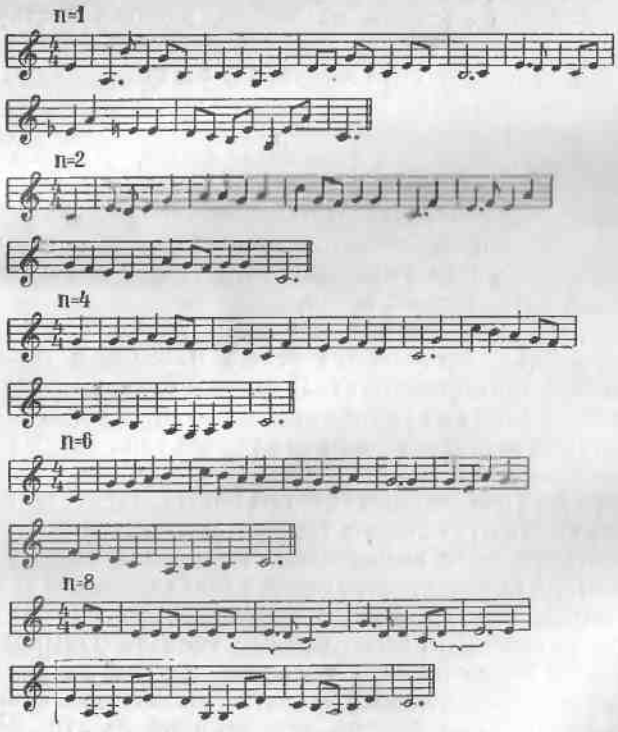


Избыточность этой мелодии может быть сравнительно просто подсчитана, исходя из статистических закономерностей, положенных в основу при ее получении; она оказалась превышающей 63%. По словам Пинкертона, «эта мелодия довольно монотонна, но все же менее монотонна, чем некоторые из настоящих детских мелодий»; отсюда можно заключить, что и в «настоящих» детских песенках избыточность, вероятно, имеет тот же порядок.

Аналогичные попытки получения новых мелодий с помощью опытов типа извлечения билетиков из урны производились Ф. и К. Аттнив в применении к ковбойским песням. При этом также учитывались лишь вероятности отдельных нот и двухнотных комбинаций (т. е. строились «фразы» типа приведенной на стр. 242) и также дополнительно требовалось, чтобы сохранялся определенный ритм. Единственным отличием от работы Пинкертона было то, что ковбойские мелодии оказалось удобнее составлять «с конца», используя подсчитанные условные вероятности того, что заданной ноте будет *предшествовать* та или иная нота. Как было указано на Лондонской конференции по теории информации, среди нескольких десятков «случайных музыкальных фраз», составленных Аттнив, две оказались удачными — похожими на настоящие ковбойские мелодии. Сравнительно малый процент удач естественно объясняется тем, что во внимание принимались лишь самые простые статистические закономерности рассматриваемых песен.

Той же самой была и основная цель работы [120] Брукса и др.— составление новых мелодий с помощью «случайных экспериментов». В данном случае только «извлечение билетика из урны» автоматически осуществлялось электронной машиной; операции такого типа оказываются весьма полезными при многих вычислениях на таких машинах (так называемые «методы Монте-Карло»)— и в настоящее время существуют хорошо разработанные методы их автоматического выполнения. Громадные возможности современной быстродействующей вычислительной техники были продемонстрированы, в частности, тем, что Брукс и др. сумели составить всевозможные «модели музыкальных фраз» — от «фраз первого порядка», в которых учитывались лишь относительные частоты появления отдельных нот (типа «русской фразы», приведенной на стр. 240), и

вилоть до «фраз восьмого порядка» включительно, в которых принимались во внимание частоты всевозможных последовательностей из восьми нот. При составлении «фразы n-го порядка» (где n в разных опытах принимало значения 1, 2, 3, 4, 5, 6, 7 или 8) каждый раз заранее задавалась определенная «ритмическая схема» (касающаяся распределения длительностей нот и пауз), а затем все ноты последовательно выбирались «наудачу», но в соответствии с подсчитанными частотами различных сочетаний из n нот. Если при таком выборе заданная «ритмическая схема» оказывалась не удовлетворенной, то соответствующая нота браковалась и машина автоматически повторяла процедуру «случайного выбора»; если 15 последовательных попыток приводили к «браконанным нотам», то машина останавливалась и составление всего ряда нот начиналось



с самого начала. Всего таким образом было составлено около 600 «новых гимнов» (при общем числе попыток порядка 6000); большой процент неудач объясняется тем, что при некоторых значениях n (в частности, при $n = 5$ и $n = 7$) оказалось очень трудно удовлетворить ритмической схеме. На стр. 288 приведены примеры построенных мелодий с $n = 1, 2, 4, 6$ и 8. При $n = 1$ и $n = 2$ построенные «мелодии» содержат много странных сочетаний нот и неестественных интервалов; несмотря на наличие жесткой ритмической схемы, эти «мелодии» нелегко пропеть. При $n = 4$ и $n = 6$ они заметно более приближаются к обычным гимнам. В случае же $n = 8$ «творчество» электронной машины свелось к малооригинальным компиляциям: целые куски полученных «мелодий» полностью совпадают с отрывками из одного из гимнов и лишь иногда (в местах, где два или более из рассмотренных 37 гимнов имеют одинаковые группы из 7 нот) происходит переход от одного гимна к другому (в частности, записанный выше отрывок составлен из частей трех разных гимнов; места перехода обозначены фигурной скобкой снизу). Это обстоятельство связано с малым объемом материала, использованного при составлении таблиц частот, что, естественно, приводило к чрезвычайно высокой избыточности [1]). Дело в том, что многие комбинации из 8 нот встречались в проанализированных отрывках гимнов лишь по о д н о м у разу; поэтому при $n = 8$ много нот подряд оказывались выбранными из одного гимна.

Родственные попытки были описаны и в статье О л с о н а и Б е л а р а [121], также использовавших анализ частот отдельных нот, их пар и троек в песнях С. Фостера для создания специальной «машины-композитора», сочиняющей (а затем даже и проигрывающей) простенькие музыкальные композиции, аналогичные (с точки зрения своей статистической структуры) мелодиям Фостера. В последующие годы опыты по сочинению искусственных

[1]) Заметим, что в любом отрывке, в котором никакие N соседних нот (или букв, или фонем) не повторяются, энтропия H_N будет равна нулю, т. е. подсчитанная по H_N избыточность будет равна единице. Поэтому надежное определение условной энтропии H_N при большом N требует использования громадного статистического материала.

музыкальных композиций с помощью вычислительных машин, использующие данные статистического анализа различных музыкальных произведений, получили большое развитие в ряде стран; при этом, например, в США мелодии, «сочиненные машиной», многократно передавались по радио и были записаны на пластинки, поступившие в продажу. Мы здесь, однако, не будем задерживаться на обсуждении указанных опытов, лишь косвенно связанных с непосредственным изучением теоретико-информационных характеристик музыкальных текстов, а отошлем интересующихся читателей к книге Р. Х. Зарипова [122], в которой все эти опыты рассмотрены весьма подробно.

Передача непрерывно изменяющихся сообщений. Телевизионные изображения

Прежде чем идти дальше, подчеркнем одно обстоятельство, имеющее очень большое значение и для теории и для практики передачи информации по линиям связи. Ясно, что устная речь или музыка принципиально отличаются от письменной речи в том отношении, что здесь «возможными сообщениями» являются уже не последовательности символов («букв»), могущих принимать конечное число значений, а совокупности звуковых колебаний, могущих меняться непрерывным образом. Поэтому, строго говоря, следовало бы считать, что каждый звук может иметь бесконечно много «значений»; однако в таком случае все формулы нашей книги становятся неприменимыми. Выше мы вышли из этого затруднения, воспользовавшись разбиением всех звуков русского языка на конечное число фонем, а всех музыкальных звуков — на конечное число нот. Но законно ли это?

Для ответа на этот вопрос надо будет разобраться в истинном смысле использованного разбиения. Дело заключается в том, что если нас интересует лишь «смысловая информация», содержащаяся в устной речи, то можно не обращать внимания на любые изменения звуков речи, не препятствующие пониманию сказанного и не меняющие его смысла. Поэтому мы вполне можем объединить большое число схожих между собой звуков, если только замена одного из них другим не изменяет смысла сказанного.

Но фонема фактически и представляет собой как раз такую совокупность близких между собой звуков, имеющих одно и то же смысловое значение (наоборот, замена одной фонемы в устной речи другой может изменить смысл слова; это свойство часто кладется в основу при определении фонемы). Отсюда ясно, что при рассмотрении вопроса о содержащейся в устной речи смысловой информации мы должны считать «основными элементами» речи не все вообще различные между собой звуки (число которых, разумеется, бесконечно), а лишь всевозможные «осмысленные звуки», имеющие различный смысл — фонемы. Точно так же в случае музыки, если интересоваться лишь информацией, содержащейся в самом исполняемом произведении, а не в трактовке его данным исполнителем, то следует отождествить все звуки, выражаемые одной и той же последовательностью нотных знаков, т. е. рассматривать лишь конечное число различных «основных звуков», соответствующих конечному числу имеющихся нот.

Но ведь можно поставить вопрос и шире: в случае речи помимо «смысловой информации» можно рассматривать также и информацию, содержащуюся в интонации и в тоне голоса, а в случае музыки можно специально интересоваться особенностями данного индивидуального исполнения (передача этих особенностей является весьма важной задачей техники связи). Надо ли в этом случае считать, что каждый звук может принимать бесконечное множество значений и поэтому имеет бесконечную энтропию? На этот вопрос мы фактически уже один раз ответили отрицательно — на стр. 276—277, где были указаны конкретные оценки энтропии устной речи с учетом различных форм «несмысловой» информации. Сейчас мы несколько подробнее остановимся на разъяснении этого обстоятельства.

Верно, конечно, что громкость звука или высота тона могут меняться непрерывным образом, т. е. могут принимать бесконечное число различных значений; к тому же в принципе эти значения могут сколь угодно быстро сменять одно другое. Однако наше ухо может различать только не слишком быстро следующие друг за другом звуки; поэтому можно считать, что все звуки, которые мы слышим, имеют определенную минимальную длительность. Кроме того, мы можем различить лишь звуки,

отличающиеся по громкости и по высоте не меньше чем на некоторое определенное конечное значение, и не воспринимаем ни слишком высокие, ни слишком низкие, ни слишком тихие, ни слишком громкие звуки (громкие звуки нас оглушают). Отсюда вытекает, что на самом деле различимо лишь конечное число градаций громкости и высоты тона. Отождествив на этом основании все звуки, громкость и высота тона которых находятся в пределах одной градации, мы снова придем к привычному для нас случаю последовательностей сигналов, могущих принимать лишь к о н е ч н о е число разных значений.

Рассмотренная здесь весьма общая ситуация очень близка к той, с которой мы столкнулись при решении задачи 22 из § 3 гл. II (стр. 112). Там нам также встретился случай опыта β, имеющего бесконечное число возможных исходов; однако оказалось, что при решении задачи опыт β вполне можно заменить новым опытом β_ε, получающимся из β при помощи отождествления всех его исходов, отличающихся друг от друга меньше чем на некоторое малое число ε. Энтропию H_ε этого нового опыта β_ε (в отличие от энтропии самого опыта β являющуюся уже конечной величиной) мы назвали *ε-э н т р о п и е й* опыта β. Во всех вопросах, касающихся передачи сообщений, представляемых *непрерывно меняющимися величинами*, ε-энтропия играет весьма важную роль. При передаче таких сообщений совокупность всевозможных значений передаваемого сигнала всегда разбивается на конечное число градаций («ячеек» в пространстве значений) и все значения в пределах одной градации отождествляются между собой (например, считаются совпадающими с «центром» соответствующей ячейки). Эта операция замены непрерывного сообщения новым сообщением, принимающим лишь конечное число возможных значений, называется в технике связи к в а н т о в а н и е м сообщения. Квантованное сообщение всегда имеет конечную энтропию (представляющую собой один из вариантов ε-энтропии исходного непрерывного сообщения), зависящую от выбора применяемого метода квантования, но характеризующую также и степень неопределенности исходного непрерывного сообщения; это последнее обстоятельство как раз и определяет возможность использования соответствующей величины в технике связи.

Важным классом таких непрерывно меняющихся сообщений являются изображения, передаваемые по телевизионным или фототелеграфным линиям связи. Легко понять, что принципиально здесь мы имеем то же положение, что и в случае передачи звука — наш глаз способен различить лишь конечное число степеней яркости изображения и лишь не слишком близкие его участки; поэтому любое изображение можно передавать «по точкам», каждая из которых является сигналом, принимающим лишь конечное число значений. В случае фототелеграфа во многих случаях можно считать, что каждый «элементарный сигнал» (т. е. мельчайший элемент изображения — «точка») принимает лишь одно из двух значений — является либо «белым», либо «черным»; в телевидении же необходимо учитывать значительное число (несколько десятков) градаций степени почернения («яркости») каждого элемента. Кроме того, фототелеграфные изображения являются неподвижными, а на телеэкране ежесекундно сменяется 25 кадров, создавая впечатление «движения». В обоих случаях, однако, по линии связи фактически передается не исход опыта α_0, состоящего в определении значения непрерывно меняющейся от точки к точке (а в случае телевидения — и во времени) окраски или яркости изображения, а исход совсем другого «квантованного» опыта α_1, состоящего в определении цвета (белого или черного) или градаций яркости в конечном числе «точек». Этот новый опыт α_1 может иметь уже лишь конечное число исходов, и мы можем измерить его энтропию H (являющуюся, по существу, одним из вариантов ε-энтропии исходного опыта α_0).

Общее число элементов («точек»), на которые следует разлагать изображение, определяется в первую очередь так называемой «разрешающей способностью» глаза, т. е. его способностью различать близкие участки изображения. В современном телевидении это число обычно имеет порядок нескольких сотен тысяч (в советских телепередачах изображение разлагается на 400 000—500 000 элементов, в американских — примерно на 200 000—300 000, в передачах некоторых французских и бельгийских телецентров — почти на 1 000 000). Нетрудно понять, что по этой причине энтропия телевизионного изображения имеет огромную величину. Так, если даже считать, что

человеческий глаз различает лишь 16 разных градаций яркости (значение явно заниженное) и что изображение разлагается всего на 200 000 элементов, то мы найдем, что «энтропия нулевого порядка» здесь равна $H_0 = \log 16^{200000} = 800\,000$ бит. Значение истинной энтропии H, разумеется, будет меньше, так как телевизионное изображение имеет значительную избыточность $R = 1 - \frac{H}{H_0}$. Действительно, ведь при вычислении величины H_0 мы предполагали, что значения яркости в любых двух «точках» изображения являются независимыми между собой, в то время как на самом деле яркость обычно очень мало меняется при переходе к соседним элементам того же (или даже другого, но близкого по времени) изображения. Наглядный смысл этой избыточности R заключается в том, что среди наших $16^{200\,000}$ возможных комбинаций значений яркости во всех точках экрана осмысленные комбинации, которые можно назвать «изображениями», будут составлять лишь ничтожно малую часть. Подавляющее же большинство этих комбинаций будет представлять собой совершенно беспорядочную совокупность точек разной яркости, весьма далекую от какого бы то ни было «сюжета». Между тем реальная «степень неопределенности» H телевизионного изображения, разумеется, должна учитывать лишь те комбинации значений яркости, которые имеют хоть какие-то шансы быть переданными, а не все вообще комбинации значений яркости [1]).

[1]) Не следует только думать, что из крайней редкости «осмысленных изображений» автоматически вытекает, что избыточность R обязательно очень велика. В самом деле, предположив, например, что человеческий глаз различает всего 10 различных градаций яркости (так что общее число возможных комбинаций яркости равно $10^{200\,000}$) и что «осмысленные изображения» (которые для простоты мы будем считать все равновероятными) составляют всего 0,00...01% (где вслед за запятой стоит 1997 нулей!) от всех возможных комбинаций яркости, мы легко найдем, что избыточность R близка к $1 - \frac{200\,000 - 2000}{200\,000} = 0{,}01 = 1\%$, т. е. весьма мала (если бы мы увеличили число различаемых градаций яркости, то она бы стала еще меньше). Этот как будто бы неожиданный результат объясняется крайней медленностью изменения функции $\log n$ при больших значениях n, о которой мы уже упоминали на стр. 264 (в связи с оценкой избыточности «иероглифической» письменности) и на стр. 276 (в связи с оценкой «несмысловой» информации устной речи).

Для определения точного значения энтропии H (или избыточности R) телевизионного изображения нужно детально изучить статистические зависимости между яркостями различных точек экрана. Эта задача весьма трудна, и в настоящее время мы имеем лишь несколько относящихся сюда частных результатов. Так, американский инженер У. Ф. Шрейбер [129] нашел значения энтропий H_0, H_1, H_2 и H_3 для двух конкретных телевизионных изображений, первое из которых (изображение A — парк с деревьями и строениями) было более сложным, а второе (изображение $Б$ — довольно темная галерея с прохожими) было более однотонным по цвету и содержало меньше деталей. Шрейбер различал при этом 64 разных градаций яркости элемента телевизионного изображения; поэтому энтропия H_0 (отнесенная к одному элементу, а не ко всему изображению в целом) здесь оказалась равной $H_0 = \log 64 = 6$ бит. Далее с помощью специального радиотехнического устройства он подсчитал для обоих рассматриваемых изображений относительные частоты (вероятности) p_1, p_2, . . ., p_{64} всех различимых градаций яркости и определил «энтропию первого порядка»

$$H_1 = H(\alpha_1) = -p_1 \log p_1 - p_2 \log p_2 - \ldots - p_{64} \log p_{64}$$

(заметим, что непосредственный подсчет частот p_1, p_2, ..., p_{64} без привлечения радиотехники при общем числе элементов экрана порядка 200 000 вряд ли мог бы быть осуществлен). То же самое радиотехническое устройство было применено затем для вычисления относительных частот p_{ij} пар соседних (по горизонтали) элементов, в которых первый элемент имеет i-е значение яркости, а второй j-е, а также относительных частот p_{ijk} троек соседних (также лишь по горизонтали) элементов, в которых первый элемент имел i-е значение яркости, второй j-е, а третий k-е (числа i, j, и k пробегали все значения от 1 до 64). Эти частоты позволили определить «энтропии сложных опытов»

$$H(\alpha_1\alpha_2) = -p_{11} \log p_{11} - p_{12} \log p_{12} - \ldots - p_{64,64} \log p_{64,64}$$

и

$$H(\alpha_1\alpha_2\alpha_3) = -p_{111} \log p_{111} - \ldots - p_{64,64,64} \log p_{64,64,64}$$

а затем и «условные энтропии» (ср. выше, стр. 241—243) $H_2 = H_{\alpha_1}(\alpha_2) = H(\alpha_1\alpha_2) - H(\alpha_1)$ и $H_3 = H_{\alpha_1\alpha_2}(\alpha_3) = H(\alpha_1\alpha_2\alpha_3) - H(\alpha_1\alpha_2)$, последняя из которых, впрочем, была подсчитана лишь для изображения $Б$. Полученные результаты сведены в следующую таблицу:

	H_0	H_1	H_2	H_3
Изображение $А$	6	5,7	3,4	—
Изображение $Б$	6	4,3	1,9	1,5

Из таблицы видно, что энтропия H_1 лишь немного отличается от максимальной энтропии H_0, причем для изображения $А$ она заметно больше, чем для $Б$ (это, очевидно, связано с большей однотонностью изображения $Б$ по сравнению с изображением $А$). Условная энтропия H_2 (т. е. средняя «степень неопределенности» яркости элемента экрана при известной яркости соседнего по горизонтали элемента) уже гораздо больше отличается от H_0; для изображения $Б$ она также и заметно меньше, чем для $А$, что соответствует меньшему обилию деталей в изображении $Б$. Избыточность R, оцененная по величине H_2 (т. е. разность $1 - \frac{H_2}{H_0}$) для изображения $А$ равна 44%, а для изображения $Б$ — 68%; действительное значение избыточности может быть только больше этого. Что же касается условной энтропии H_3 при известных яркостях двух предыдущих элементов той же строки, то она сравнительно мало отличается от H_2 (ей соответствует значение избыточности изображения $Б$, равное 75%); отсюда можно заключить, что знание яркости самого близкого элемента определяет весьма большую часть общей избыточности.

Близкий характер имеют также работы Д. С. Лебедева и Е. И. Шийль [130] (см. также книгу [128]) и Дж. О. Лимба [131]. В статье [130] и книге [128] приведены результаты вычислений (опирающихся на использование несколько более бедного, чем в работе [129], статистического материала и на разбиение возможных значений яркости элемента телевизионного изображения на 8, а не на 64 градаций) энтропий H_0 и H_1 и ряда условных энтропий H_2, H_3 и H_4 одного элемента изображения для следующих четырех спортивных телевизионных сю-

жетов: A — быстро бегущие баскетболисты, $Б$ — лицо одного зрителя на трибуне стадиона крупным планом, B — панорамирование вида зрителей на трибуне и $Г$ — быстро бегущие футболисты. Будем обозначать цифрами 1 и 2 соседние с данным по горизонтали и по вертикали элементы изображения, цифрой 3 — соседний по диагонали элемент, цифрой 4 — тот же, что и рассматриваемый, элемент на предшествующем кадре телевизионной передачи, цифрой 5 — элемент на той же горизонтали,

Рис. 16.

соседний с элементом 1, и, наконец, цифрой 6 — тот же элемент на кадре, предшествующем тому, который содержит элемент 4 (см. рис. 16, *а*), и будем указывать в обозначениях условных энтропий сверху в скобках номера элементов изображения, степень яркости которых считается известной. В таком случае найденные в [130] (см. также [128]) значения энтропий (в битах) могут быть сведены в следующую таблицу:

	H_0	H_1	$H_2^{(1)}$	$H_2^{(2)}$	$H_2^{(4)}$	$H_2^{(6)}$
A	3	1,96	0,69	0,98	—	1,77
$Б$	3	1,95	0,36	0 39	—	—
B	3	2,78	1,34	1,95	2,78	—
$Г$	3	2,45	—	—	2,00	2,08

	$H_3^{(1,\,5)}$	$H_3^{(4,\,6)}$	$H_3^{(1,\,2)}$	$H_4^{(1,\,2,\,3)}$	$H_4^{(1,\,2,\,4)}$
A	0,68	—	0,56	—	—
$Б$	0,35	—	0,27	0,26	—
B	—	—	1,22	1,18	1,19
$Г$	—	1,83	—	—	—

(черточки в таблице означают, что соответствующие энтропии не были сосчитаны). В работе [131] были

проанализированы следующие четыре части (содержащие по 5000 отдельных элементов каждая) двух телевизионных изображений: A — покрытая травой и кустами поверхность земли средним планом, $Б$ — примыкающая к части A и аналогичная ей часть того же пейзажа, B — часть неба, покрытого сравнительно однородными светлыми облаками, и $Г$ — травянистая растительность крупным планом. Изображения были разбиты на 16 градаций яркости; при вычислении условных энтропий элемента изображения с номером 0 использовались данные, относящиеся к элементам 1, 2, 3, 4 и 5 той же и предыдущей строк того же кадра (см. рис. 16, *б*). Полученные в [131] результаты приведены ниже в виде таблицы:

	H_0	H_1	$H_1^{(2)}$	$H_1^{(1)}$	$H_2^{(1,\,2)}$	$H_2^{(1,\,4)}$	$H_3^{(1,\,2,\,3)}$	$H_3^{(1,\,4,\,5)}$
A	4	2,85	2,24	2,38	1,82	2,10	1,46	1,47
$Б$	4	2,51	1 99	1,96	1,66	1,66	1,15	1,28
B	4	1,32	1,04	0,99	0,94	0,97	0,90	0,92
$Г$	4	3,72	2,70	3 10	2,01	2,23	0,87	0,86
A и $Б$	4	2 90	—	2,27	—	2,03	—	1,54
B и $Г$	4	3 29	—	2,17	—	1,65	—	0,91
A, $Б$, B и $Г$	4	3,52	—	2,31	—	2,00	—	1,39

Содержащиеся в [128], [130], [131] данные качественно близки к результатам работы [129] (количественное сравнение здесь затруднено различиями в числе используемых уровней квантования, влияющим на численные значения энтропий), но заметно более полны. В частности, вывод Шрейбера (относящийся к сравнительно однотонному и бедному деталями изображению $Б$) о том, что при известном одном предшествующем элементе изображения знание еще каких-то других элементов уже мало меняет степень неопределенности (т. е. энтропию) данного элемента телевизионного изображения, прекрасно согласуется с данными, относящимися к однотонным и бедным деталями изображениям лица крупным планом (изображение $Б$ работ [130], [128]) и облачного неба (изображение B работы [131]). Заметим, однако, что согласно приведенным в [128] данным указанный вывод неплохо выполняется и для всех других исследованных изображений (включая и наиболее «пестрое» изображение B), в то время как результаты [131], относящиеся к изображениям A, $Б$ и $Г$, его не подтверждают. Анализ данных Лимба позволяет также

заключить, что использование вероятностей (т. е. частот), подсчитанных для большого и весьма неоднородного изображения (моделью которого можно считать объединение разнородных частей A, $Б$, B и $Г$ двух разных кадров), приводит лишь к небольшому увеличению значений условных энтропий (при известных значениях яркости одного, двух или трех предшествующих элементов) по сравнению со средними значениями условных энтропий, подсчитанных для каждой из частей большого изображения в отдельности. Далее, результаты работ [130], [128], относящиеся к условным энтропиям при известных значениях яркостей того же элемента изображения на одном или двух предыдущих кадрах, показывают, что для рассматривавшихся быстро меняющихся изображений эти условные энтропии оказываются заметно превосходящими условную энтропию при известной яркости предшествующего (вдоль строки) элемента того же кадра; поэтому учет связи между значениями яркостей на последовательных кадрах телевизионной передачи здесь не может привести к значительному возрастанию избыточности, определенной из анализа распределения яркостей на одном кадре. Последний вывод, разумеется, не может быть справедлив для телевизионных сюжетов, при которых изображение мало меняется во времени; однако надежные количественные данные, относящиеся к таким случаям, пока еще отсутствуют (некоторые оценки влияния временных связей, основанные на косвенных соображениях, могут быть найдены в книге [132]). Общая избыточность телевизионных изображений по данным работы [131] и в случае богатого деталями изображения («растительность крупным планом»), и в случае бедного деталями однотонного изображения («небо») оказывается не меньшей, чем 80% (но для «средних» изображений A и $Б$ она почему-то оказывается не столь высокой, хотя все же не меньшей, чем 65%). В то же время результаты [130], [128] приводят к выводу, что для бедного деталями изображения («лицо») избыточность не меньше, чем 90%, а для изображения, богатого деталями («зрители»), она не меньше, чем 60%. Заметим, что бо́льшие, чем найденные Шрейбером [129], значения избыточности в работах [128], [130], [131] могут естественно объясняться более грубым делением на градации яркости; что же касается расхождения в выводах

Лебедева и Лимба о различиях избыточности «однотонных» и «пестрых» изображений, то они отражают уже отмеченное выше расхождение в выводах этих авторов о характере убывания ряда энтропий H_0, H_1, H_2, H_3, H_4 для всех не слишком бедных деталями изображений (причины этого расхождения пока неясны, но в целом результаты работ [128], [130] кажутся все же более правдоподобными, чем результаты статьи [131]).

Ясно, что подсчеты того типа, который описан в работах [128] — [131], не могут быть использованы для определения влияния на избыточность изображения связей между большим числом его элементов: уже в случае энтропии H_4 число различных комбинаций значений яркости в четырех точках оказывается огромным (напомним, что в работах [128], [130], [131] применялось сравнительно грубое деление на градации яркости), а при дальнейшем возрастании порядка условной энтропии это число стремительно возрастает и трудности вычислений становятся непреодолимыми. Поэтому заслуживает внимания попытка американского ученого Н. Цаниеса и его сотрудников [133] применить для оценки условной энтропии изображения с учетом также и далеких связей между элементами «метод угадывания», предложенный Шенноном [75] для оценки энтропий высокого порядка письменной речи и описанный выше на стр. 249 и след. В опытах Цаниеса в качестве исходного материала были выбраны 20 фотографий частей лунной поверхности, каждая из которых была представлена в виде совокупности $50 \times 50 = 2500$ отдельных элементов, принимающих одно из восьми возможных значений в зависимости от своей «яркости» (т. е. степени почернения). Далее эти фотографии были разбиты на 4 группы родственных по своему характеру фотографий. Одна из фотографий (вместе с ее числовой формой, представляющей собой квадратную таблицу из 2500 чисел от 0 до 7) давалась отгадывающему лицу (студенту старшего курса университета), которому предлагалось внимательно ее изучить (достигаемое таким путем «ознакомление с изображением», разумеется, мало сравнимо с присущим каждому грамотному человеку знанием структуры родного языка, использовавшемуся в опытах по отгадыванию письменных текстов, но здесь уж ничего поделать нельзя), после чего тот же человек

начинал последовательно отгадывать элементы другой фотографии из той же группы. При отгадывании разрешалось после каждого уже отгаданного элемента двигаться в любом направлении; на каждую догадку давался ответ «*да*» или «*нет*», который считался содержащим один бит информации (на самом деле он часто содержал заметно меньшую информацию, так как оба возможных ответа вовсе не были равновероятны). Таким образом, среднее число вопросов, приходящихся на один элемент изображения, доставляло довольно грубую оценку сверху (т. е. сильно завышенную) средней энтропии одного элемента изображения. В описанных в [133] двух опытах по отгадыванию эта средняя оценка оказалась примерно равной 1,8 бит в одном случае и 1,3 бит во втором; авторы отмечают, что специалист в области изучения фотографий лунной поверхности, предварительно потренировавшись, мог бы, вероятно, получить заметно лучшие результаты (т. е. меньшую оценку энтропии). Во всяком случае и так обе полученные оценки оказались заметно меньшими, чем значение $H_0 = 3$ бита; истинная энтропия H, по-видимому, еще значительно меньше, чем эти оценки. Если, следуя предложению Шеннона, приведенному в сноске на стр. 250, использовать только результат более удачливого из двух отгадывающих лиц, то соответствующая оценка снизу избыточности изображения лунной поверхности будет близка к 60%.

В последнее время в связи с появлением цветного телевидения возникла также потребность в оценке информации, содержащейся в окраске изображения. Первые грубо ориентировочные расчеты такого рода показали, что для цветных телевизионных изображений, приближающихся по качеству к хорошим цветным иллюстрациям в журналах, информация по порядку величины сравнима с удвоенной информацией, содержащейся в соответствующем черно-белом изображении (ср. [132]).

Фототелеграммы

Перейдем теперь к данным, касающимся *фототелеграфа*. Общий принцип передачи изображений здесь близок к принципу телепередач: изображение разлагается на мельчайшие квадратики («растровые элементы»),

после чего по линии передается информация о цвете каждого такого элемента (черный он или белый). Таким образом, по сравнению с черно-белыми телевизионными изображениями рассматриваемые сейчас изображения более просты: для них не существует градаций яркости (т. е. степени почернения), а цвет может принимать лишь два значения. Естественно, что максимальная информация (т. е. энтропия H_0), содержащаяся в сведении о цвете одного элемента, равна $H_0 = \log 2 = 1$ бит; эта информация достигается, когда черные и белые элементы встречаются одинаково часто и цвет каждого элемента независим от цвета всех остальных. На самом деле два цвета обычно встречаются с разной частотой (число белых элементов, как правило, значительно превосходит число черных) и между цветами отдельных элементов существует заметная зависимость; поэтому истинное значение энтропии одного элемента фототелеграммы заметно меньше, чем 1 бит. Чему же оно равно?

Можно подсчитать, что при передаче по фототелеграфу обычного книжного или журнального печатного текста относительная частота p_0 белых элементов близка к 0,8, а частота p_1 черных элементов — к 0,2. Отсюда следует, что энтропия H_1 здесь равна

$$H_1 = -0{,}2 \cdot \log 0{,}2 - 0{,}8 \cdot \log 0{,}8 \approx 0{,}73 \text{ бит},$$

что соответствует избыточности $R = 1 - \frac{0{,}73}{1} = 0{,}27 = 27\%$. Однако это значение избыточности сильно занижено, так как оно не учитывает зависимости между цветами соседних элементов. К сожалению, точный количественный учет этой зависимости (простирающейся на большое число соседних элементов) весьма сложен; поэтому представляют интерес и приближенные методы оценки энтропии H_∞ и избыточности R.

Одна из первых, весьма мало совершенных попыток оценить энтропию $H_\infty = H$ фототелеграфных сообщений описана в работе американского связиста С. Дейча [134]. В этой работе анализировался небольшой отрывок английского текста (порядка нескольких строк), напечатанного сравнительно крупными буквами. Так как записанный на бумаге текст совсем не просто непосредственно

разбить на мельчайшие «растровые элементы», используемые в фототелеграфии, и при таком разбиении анализируемый отрывок оказывается состоящим из громадного числа элементов, что необычайно усложняет арифметический подсчет частот различных комбинаций, то Дейч использовал разложение анализируемого текста на сравнительно большие квадратики, состоящие из многих растровых элементов каждый. Такой квадратик он считал белым или черным в зависимости от того, какой цвет имеет бо́льшая часть квадратика (т. е. если более 50% площади квадратика оказывалась белой, то весь квадратик считался белым; в противном случае он считался черным). Естественно, что в таком случае для «квадратика», как и для растрового элемента, $H_0 = \log 2 = 1$ бит. Далее Дейч подсчитал условные энтропии H_1, H_2 и H_3 для вертикальных «блоков», состоящих из нескольких соседних квадратиков (для горизонтальных «блоков» была подсчитана лишь величина H_2, которая оказалась немного большей соответствующей величины для вертикальных «блоков»). Энтропия H_1 оказалась равной 0,67 бит, что соответствует избыточности R, равной 33%; энтропия H_3 имела уже значение 0,57 бит, т. е. отвечала избыточности $R = 43\%$ [1]). При помощи некоторых косвенных соображений в работе [134] было также показано, что энтропия одного «квадратика» на самом деле должна быть заметно меньше, чем 0,5 бит, так что избыточность R здесь должна значительно превышать 50%. Заметим, впрочем, что все эти цифры не заслуживают особенно большого доверия, так как использованное в работе [134] разбиение текста на сравнительно большие квадраты заметно искажает его статистическую структуру.

Значительно более детальное исследование того же рода выполнил немецкий ученый Г. Кайзер [135]. Он уже

---

[1]) Для вертикальных блоков была подсчитана еще энтропия $H^{(N)}$ блоков из N соседних элементов для $N = 1, 2, 3$ и 7. Любопытно, что отношение $\frac{H^{(N)}}{N}$ при $N = 7$ оказалось равным всего 0,58 бит, т. е. даже несколько большим, чем H_3. Этот факт наглядно показывает, насколько медленнее приближается к величине H_∞ последовательность величин $h_N = \frac{H^{(N)}}{N}$, $N = 1, 2, 3, \ldots$, чем последовательность H_N (ср. сноску на стр. 244).

разбивал напечатанные на пишущей машинке тексты на гораздо более мелкие квадратики со стороной 0,2 *мм* (одна печатная страница при этом оказывалась разбитой примерно на миллион отдельных элементов). Для того чтобы сделать возможными расчеты со столь большими статистическими совокупностями, Кайзер сконструировал специальную измерительную аппаратуру, автоматически выделяющую последовательные «блоки» из небольшого числа N соседних элементов и регистрирующую на счетчиках число блоков различного состава. Эта аппаратура была затем применена к блокам различного направления (горизонтальным, вертикальным и расположенным под углом к печатному тексту), причем оказалось, что все результаты подсчетов мало меняются при изменении направления. Исходя отсюда, Кайзер, в основном, ограничился анализом данных для горизонтальных блоков, в применении к которым он изучил зависимость удельных энтропий $h_N = \frac{H^{(N)}}{N}$, где $N = 1, 2, 3, 4, 5$ и 6, от следующих факторов: а) степени «жирности» (т. е. толщины букв) текста, б) расстояния между строками и в) размера машинописи (т. е. степени увеличения машинописной копии). Полученные им результаты, относящиеся к нормальному по «жирности» и размеру тексту и пяти разным расстояниям между строками (от наиболее густой машинописи «через один интервал» и до наиболее редкой — «через три интервала»), показаны на рис. 17. Из него видно, что избыточность наиболее «густого» (но нормального во всех других отношениях) машинописного текста наверное превосходит 50%, в то время как для наиболее «редкой» машинописи она уже не меньше, чем 80% (причем, по-

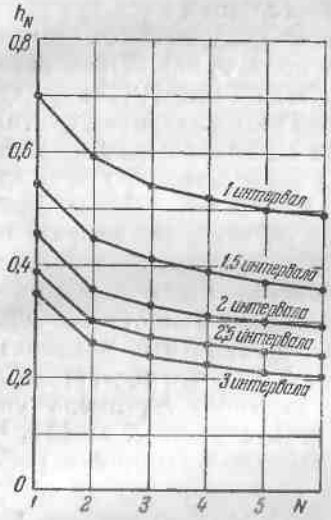


Рис. 17.

видимому, цифры эти сильно занижены, так как h_6 является весьма грубой оценкой величины H_∞). В случае тонко напечатанного текста все энтропии, естественно, оказываются меньшими, а избыточности — бо́льшими, причем особенно заметно уменьшается значение $h_1 = H_1$; с ростом же N значения h_N для тонкого шрифта постепенно приближаются к значениям для обычного шрифта. Для текста, напечатанного очень «жирно», наоборот, все энтропии оказываются бо́льшими, чем для нормального текста, причем самая большая разница снова наблюдается при $N = 1$, а самая маленькая — при $N = 6$. При подобном увеличении машинописной копии значения $h_1 = H_1$, разумеется, не меняются (так как не меняется доля белых и черных элементов), но статистические связи между соседними элементами при этом возрастают, и потому все энтропии h_N с $N > 1$ здесь уменьшаются, а избыточности возрастают. В отношении значений h_N с $N > 6$ в работе [135] приведены лишь некоторые довольно грубые оценки, согласно которым, например, для напечатанного через один интервал нормального машинописного текста $h_{12} \approx \approx 0{,}40 - 0{,}45$ бит.

Ясно, что величины h_N при небольших N никак не характеризуют полной избыточности машинописного текста, обусловленной всеми существующими в таком тексте статистическими зависимостями. Это видно, в частности, из того, что, применив совсем другой метод, Кайзер сразу же получил результаты, сильно отличающиеся от описанных выше. Дело в том, что сконструированная им измерительная аппаратура, конечно, не могла как следует учесть то, что все черные элементы в ее поле зрения на самом деле представляли собой части 26 немецких букв вполне определенной формы. Поэтому Кайзер попытался дополнительно выяснить, какова наименьшая доля квадрата, плотно охватывающего одну букву, по виду которой грамотный человек может уже догадаться, какая же это буква. Поставленные с этой целью опыты показали, что если для каждой буквы выбирать наиболее характерную ее часть, то достаточно показать лишь около 15% площади квадрата. Отсюда можно заключить, что избыточность двумерного рисунка отдельных букв (а значит, и очень тесно напечатанного буквенного текста) в среднем близка к 85% (белые же промежутки между буквами, словами и

строками в напечатанном тексте вообще можно считать целиком избыточными). Кроме того, надо учесть, что показывались лишь части одной изолированной буквы; но ведь если заранее знать весь предшествующий этой букве текст, то очень часто букву можно будет угадать, даже не видя никакой ее части. Поэтому ясно, что доля одной буквы текста, нужная для ее угадывания, в среднем должна быть заметно меньшей, чем 15%. Исходя из данных работы [82], о которой говорилось на стр. 253, Кайзер заключил, что знание предшествующих букв немецкого машинописного текста должно еще примерно втрое уменьшить предельную степень неопределенности H_∞; поэтому он пришел к выводу, что истинная избыточность тесного машинописного текста, по-видимому, близка к 95%. Эта оценка избыточности уже учитывает очень сложные и охватывающие одновременно много «растровых элементов» статистические связи, порожденные и правилами написания букв, и грамматикой, и структурой языка; использование всех их в технике фототелеграфии пока еще остается делом далекого будущего.

В дальнейшем мы больше не будем учитывать смысловую и грамматическую избыточность фототелеграфных текстов, а будем рассматривать лишь статистические закономерности простого чередования черных и белых растровых элементов. В таком случае сравнительно неплохую оценку энтропии H одного растрового элемента можно получить с помощью представления каждой строки фототелеграммы в виде последовательности чередующихся белых и черных участков различной длины. Подсчитав относительные частоты появления всех таких участков, можно определить соответствующую «энтропию первого порядка» $H_1^{(\text{участ})}$; при этом отношение $\frac{H_1^{(\text{участ})}}{w}$, где w — среднее число элементов в одном участке, будет, наверное, больше, чем истинное значение энтропии H одного элемента (ср. сноску [1]) на стр. 248). С помощью этого метода У. Майчел [136] показал, что при передаче текста, густо напечатанного («через один интервал») на пишущей машинке с крупным шрифтом, энтропия H будет меньше, чем 0,3 бит, т. е. избыточность R будет превышать 70%; близкий вывод получен с помощью того же метода и в работе [135]. Более детальное исследование такого рода было

выполнено на очень большом статистическом материале для русского печатного (книжного или журнального) текста В. А. Гармашем и Н. Е. Кириллов ым [137]. Эти авторы подсчитали не только частоты одноцветных участков различной длины, но и частоты всевозможных пар подобных участков и определили по этим данным для участков энтропию первого порядка $H_1^{(\text{участ})}$ и энтропию второго порядка $H_2^{(\text{участ})}$. Подсчитав отношение $\frac{H_1^{(\text{участ})}}{w}$, они выяснили, что при передаче печатного текста $H \leqslant 0{,}33$ бита, т. е. $R \geqslant 67\%$; неравенство $H \leqslant \frac{H_2^{(\text{участ})}}{w}$ позволило еще уточнить эту оценку и показать, что $H \leqslant 0{,}28$ бита и, соответственно, $R \geqslant 1 - 0{,}28 = 72\%$.

Иной метод оценки энтропии H и избыточности R для фототелеграмм был использован Р. Р. Васильевым [138] и В. Г. Фролушкиным [139]. Ясно, что точный подсчет энтропии $H^{(N)}$ опыта, состоящего в определении цвета N последовательных растровых элементов, при большом N будет очень сложен из-за того, что общее число 2^N исходов этого опыта крайне велико. Разобьем поэтому соответствующие 2^N исходов на какие-то n групп, содержащих соответственно $M_1, M_2, \ldots, M_n$ исходов (где $M_1 + M_2 + \ldots + M_n = 2^N$) и будем определять лишь вероятности $q_1, q_2, \ldots, q_n$ того, что последовательные N элементов принадлежат 1-й, 2-й, ..., n-й группе. Предположим теперь, что внутри каждой из групп все исходы являются равновероятными (невыполнение этого предположения может только у м е н ь ш и т ь энтропию $H^{(N)}$!), и при этом предположении определим значение $H^{(N)}$. В таком случае исходам, принадлежащим i-й группе (где i может быть равно 1, 2, ..., n) в выражении для $H^{(N)}$ будут отвечать M_i одинаковых членов $-\frac{q_i}{M_i}\log\frac{q_i}{M_i}$, откуда следует, что

$$H^{(N)} \leqslant -q_1 \log \frac{q_1}{M_1} - q_2 \log \frac{q_2}{M_2} - \ldots - q_n \log \frac{q_n}{M_n} \qquad (*)$$

(знак $\leqslant$ связан с тем, что наш подсчет дает, вообще говоря, завышенное значение $H^{(N)}$). Подобным же образом

предположив, что один из исходов i-й группы имеет вероятность 1, а все остальные имеют вероятность 0, т. е. невозможны (невыполнение этого предположения может лишь увеличить энтропию $H^{(N)}$!), мы получим

$$H^{(N)} \geqslant -q_1 \log q_1 - q_2 \log q_2 - \ldots - q_n \log q_n. \quad (**)$$

Р. Р. Васильев [138] исходил из того, что при передаче печатного текста весьма значительная часть избыточности связана с большой частотой сравнительно длинных участков из N белых элементов (возникающих из-за наличия междустрочных пространств и полей). Соответственно этому у него 1-я группа исходов состоит из единственного исхода — того, при котором все N элементов являются белыми; остальные же $2^N - 1$ исходов составляют 2-ю группу. При этом формулы (*) и (**) дают

$$-q \log q - (1-q) \log \frac{(1-q)}{2^N - 1} \geqslant H^{(N)} \geqslant$$
$$\geqslant -q \log q - (1-q) \log (1-q),$$

где q — вероятность «белого» блока из N растровых элементов. Учитывая еще, что при большом N выражение $2^N - 1$ почти не отличается от 2^N, так что $\log (2^N - 1)$ можно заменить на $\log 2^N = N$, найдем, что

$$\frac{-q \log q - (1-q) \log (1-q)}{N} + (1-q) \geqslant h_N \geqslant$$
$$\geqslant \frac{-q \log q - (1-q) \log (1-q)}{N},$$

где $h_N = \frac{H(N)}{N}$ — приближенное значение «удельной энтропии» одного растрового элемента. Чтобы получить удовлетворительные оценки для $H = H_\infty = \lim\limits_{N \to \infty} h_N$ здесь надо брать N порядка одного или нескольких десятков; при этом q для газетного текста оказывается близким к 0,5 (или даже бóльшим), а для машинописного текста, отпечатанного обычным образом («через два интервала») — близким к 0,7 (или бóльшим). Отсюда ясно, что при передаче газетного текста $H \leqslant \frac{1}{10} + 0{,}5 = 0{,}6$ и $R \geqslant 1 - 0{,}6 = 40\%$;

при передаче обычного машинописного текста

$$H \leqslant \frac{-0{,}3\log 0{,}3 - 0{,}7\log 0{,}7}{10} + 0{,}3 \approx 0{,}33 \text{ и } R \geqslant 1 - 0{,}33 = 67\%.$$

Достоинством такой сравнительно грубой оценки энтропии H является то, что здесь легко указать конкретный метод кодирования, позволяющий вести передачу со скоростью

$$v = \frac{C}{H} = \frac{NC}{-q\log q - (1-q)\log(1-q) + N(1-q)}$$

(в растр. элем./ед. времени), где C — пропускная способность используемой линии связи (см. [138]).

В работе [139] всевозможные блоки из N растровых элементов разбивались на большое число групп, характеризуемых определенными значениями «насыщенности» и «детальности». Под «насыщенностью» здесь понимается просто общее число входящих в состав блока черных элементов (так что для блоков из N элементов «насыщенность» может принимать $N + 1$ значений: 0, 1, 2, . . ., N), а под «детальностью» — число одноцветных участков, на которые разбивается данный блок («детальность» блока из N элементов может равняться 1, 2, 3, . . ., или N, т. е. может иметь N различных значений). Подсчет значений «насыщенности» и «детальности» отдельных блоков производился автоматически, с помощью сконструированной Фролушкиным очень удобной специальной аппаратуры. Значение N в работе [139] принималось равным 100, т. е. оценивалась величина $H^{(100)}$ и энтропия H одного элемента приравнивалась к $h_{100} = \frac{H^{(100)}}{100}$. В связи с таким выбором числа N измерительная схема была снабжена устройством, автоматически включающим ее на промежуток времени, соответствующий передаче по линии 100 растровых элементов фототелеграммы; вслед за тем схема выключалась, значения «детальности» и «насыщенности» записывались и лишь после этого на схему снова подавался другой отрывок фототелеграммы.

Исследованию подвергались отдельно фототелеграммы с рукописным, машинописным и печатным (газетным) текстом, причем во всех случаях бланк фототелеграммы заполнялся текстом максимально плотно — так, как он

обычно заполняется при реальных передачах. Каждый из трех типов текста был представлен 10 образцами, и из каждого образца выбиралось 400 различных блоков из 100 элементов. По полученным данным определялись частоты (приближенные значения вероятностей) различных значений «насыщенности» и «детальности», а также частоты различных комбинаций значения «насыщенности» и значения «детальности». Подсчитав далее число $M_n^{(\text{нас})}$ блоков, имеющих заданную «насыщенность» n, число $M_m^{(\text{дет})}$ блоков, имеющих заданную «детальность» m и, наконец, число $M_{n,m}$ блоков, имеющих одновременно «насыщенность» n и «детальность» m (определение всех этих чисел может быть осуществлено с помощью несложных комбинаторных рассуждений [1])) и, воспользовавшись формулой (*) (стр. 307), мы получим три различные оценки энтропии H (а следовательно, и избыточности $R = 1 - \frac{H}{H_0}$). Ясно, что все эти оценки будут давать несколько завышенное значение H (и заниженное значение R), причем третья из них (отвечающая делению на наибольшее число групп), в принципе должна быть более точной, чем первые две.

В результате проведенного исследования были получены следующие оценки значений H и R для трех типов текста (см. таблицу на следующей странице). Мы видим, что оценка H по данным о «насыщенности» оказывается заметно более грубой, чем оценка по данным о «детальности». Отсюда можно заключить, что предположение о равновероятности всех блоков с одинаковым значением «детальности» лучше соответствует действительности, чем предположение о равновероятности блоков одинаковой «насыщенности» — блоки с одинаковой «де-

---

[1]) Легко понять, что в общем случае блоков из N элементов

$$M_n^{(\text{нас})} = C_N^n = \frac{N!}{n!\,(N-n)!} \quad \text{и} \quad M_m^{(\text{дет})} = 2C_{N-1}^{m-1} = \frac{2\,(N-1)!}{(m-1)!\,(N-m)!}$$

(последняя формула следует из того, что в этом случае $m-1$ «границ» между различными одноцветными участками можно выбрать C_{N-1}^{m-1} разными способами, а после этого можно еще по произволу выбрать первый одноцветный участок либо белым, либо черным). Что же касается числа $M_{n,m}$, то оно задается более сложной формулой, которую мы здесь не приводим.

	Оценка по данным о «насыщенности»		Оценка по данным о «детальности»	
	H (в битах)	R	H (в битах)	R
Рукописный текст	0,37	63%	0,22	78%
Машинописный текст	0,53	47%	0,30	70%
Газетный текст . . .	0,43	57%	0,34	66%
Среднее . . .	0,44	56%	0,29	71%

тальностью» образуют более однородные группы, чем блоки с одинаковой «насыщенностью».

Оценка энтропии H по данным о вероятностях всевозможных комбинаций «насыщенности» и «детальности», требует значительного увеличения объема использованного материала. В самом деле, нетрудно подсчитать, что для блоков из 100 элементов всего можно составить около 5000 (точнее говоря, 5001) различных таких комбинаций. Следовательно, все множество различных блоков (содержащее $2^{100} > 10^{30}$ элементов, т. е. число элементов, выражающееся 31-значным числом!) здесь разбивается на 5001 отдельную группу. Ясно, что вероятности всех этих групп никак нельзя оценить по данным о частотах, полученным при исследовании $400 \times 10 = 4000$ различных блоков. Поэтому третья оценка энтропии в работе [139] дается только для «среднего русского текста» (на основании данных о частотах отдельных групп во всей совокупности исследованных блоков без отношения к тому, из текста какого типа они извлекаются). Эта оценка, полученная с помощью формул (*) и (**), имеет вид

$$0{,}23 \geqslant H \geqslant 0{,}06, \text{ т. е. } 77\% \leqslant R \leqslant 94\%.$$

Истинные значения энтропии H и избыточности R по-видимому должны заключаться где-то между указанными здесь пределами.

До сих пор, говоря о фототелеграммах, мы рассматривали только случаи передачи по фототелеграфу текстового материала (рукописного, машинописного или печатного). Однако фототелеграф может использоваться и для

передачи ряда других типов черно-белых сообщений, и для многих из них значения средней энтропии (на один растровый элемент) и избыточности могут оказаться совсем другими, чем для буквенного текста. Так, например, ясно, что в случае чертежей следует ожидать заметно большей избыточности, чем в случае текста (в первую очередь из-за того, что на чертежах «черное» занимает гораздо меньше места, чем на листе с буквенным текстом). Этот вывод подтверждается уже первыми (весьма грубыми, а именно — заметно завышенными) оценками энтропии H для чертежей, полученными (на основе данных о распределении длин одноцветных участков) в уже цитировавшейся выше работе Майчела [136]. Согласно оценкам Майчела в случае сложных радиосхем с рядом надписей с уверенностью можно утверждать, что $H \leqslant 0{,}12$ бит, т. е. $R \geqslant 88\%$, в то время как для простых чертежей энтропия H вполне может оказаться еще более чем вдвое меньшей (т. е. избыточность будет превосходить 95%). Более аккуратный (но и заметно более сложный) метод приближенной оценки энтропии и избыточности простых чертежей (состоящих из ряда непрерывных линий) предложил У. Фоб [140]. В случае разобранного в работе [140] одного частного модельного примера учет лишь отличия относительной частоты p_1 черных элементов от 1/2 приводил к оценкам $H \leqslant 0{,}08$ бит, $R \geqslant 92\%$ (значение p_1 здесь было близко к 0,01), в то время как использование предложенного автором более точного метода позволяло получить следующий результат: $H \leqslant 0{,}015$ бит, $R \geqslant 98{,}5\%$. Что же касается до передаваемых по фототелеграфу рисунков и фотографий, то эти типы сообщений фактически мало отличаются от черно-белых телевизионных изображений; поэтому на данных об их энтропии и избыточности мы можем специально не останавливаться, отослав читателя к предыдущему разделу настоящего параграфа.

Пропускная способность реальных линий связи

В заключение этого параграфа мы остановимся еще на вопросе о практической ценности оценок энтропии и информации реальных сообщений для техники связи. Роль энтропии в теории передачи сообщений определяется

основной теоремой § 2 (стр. 230—231): максимальная достижимая скорость v передачи по линии связи определяется формулой

$$v = \frac{C}{H} \text{ элементов/ед. времени,}$$

где H — энтропия одного элемента сообщения (будет ли это буква, фонема, нота, элемент телеизображения или растровый элемент фототелеграммы — безразлично), а C — пропускная способность этой линии связи. Поэтому для того, чтобы найти предельную скорость передачи, надо знать не только энтропию H, определению которой для разных случаев была посвящена предшествующая часть настоящего параграфа, но еще и пропускную способность C. Чем же определяется пропускная способность?

В § 2 мы видели, что

$$C = L \log m,$$

где через L обозначено число элементарных сигналов, которые можно передать по линии за единицу времени, а через m — общее число используемых различных сигналов. На практике число m часто выбирается из условия, чтобы для соответствующей линии связи можно было создать достаточно простую и дешевую передающую и принимающую аппаратуру. Так, например, очень часто применяются всего 2 элементарных сигнала (обычно — посылка тока и пауза): дело в том, что задача различения таких двух сигналов на приемном конце является технически наиболее простой и основанные на этом принципе приемные аппараты наиболее дешевы и надежны. В тех случаях, однако, когда нам необходимо передать как можно больше сообщений за единицу времени, естественно пренебречь простотой и дешевизной оборудования линии и стремиться максимально увеличить значения L и m. И здесь на первый взгляд кажется, что возможности совершенно безграничны: ведь обычно сигналы, передаваемые по линии связи, могут изменяться непрерывно, так что их как будто можно выбирать сколь угодно краткими по длительности и сколь угодно мало отличающимися друг от друга. Но это означает, что числа L и m могут быть сделаны сколь угодно большими и, следовательно, пропускная способность любой линии, передающей непрерывные сигналы, факти-

чески безгранична. Какую же роль в таком случае может играть большее или меньшее значение энтропии H?

На самом деле, однако, приведенное здесь рассуждение неверно: любая линия связи, передающая непрерывные сигналы, также имеет строго ограниченную пропускную способность. Прежде всего мы никогда не можем мгновенно изменить значение передаваемого сигнала — на это всегда требуется определенное время. В используемых на практике линиях связи минимальное время, требующееся для ощутимого изменения сигнала, строго регулируется техническими характеристиками самой линии. Это приводит к тому, что для каждой линии лишь значения сигнала, разделенные определенным минимальным промежутком времени τ_0, могут выбираться более или менее произвольно: после того как эти значения выбраны, все значения сигнала в промежуточные моменты времени будут уже однозначно определены. Иначе говоря, максимальное число $L = \frac{1}{\tau_0}$ различных элементарных сигналов, которое можно передать по линии связи за единицу времени, является некоторой технической характеристикой линии, которую нельзя изменить, не внося изменений в самую линию. Это обстоятельство, играющее основную роль во всех применениях теории информации к вопросу о передаче непрерывных сигналов, впервые было четко сформулировано еще до возникновения современной теории информации (в 1933 г.) В. А. Котельниковым. В работе Котельникова число L было выражено также через привычные для техников характеристики линии связи (через так называемую «ширину полосы пропускания»); полученное выражение показывает, что, например, в случае радиосвязи перестройка линии с целью увеличения значения L не может принести выгоды, поскольку она сделает невозможной работу радиолиний, ведущих передачу на близкой длине волны (см., например, [4], [13] или [69]).

Но, может быть, хотя бы число m можно выбрать сколь угодно большим — ведь этого уже достаточно для того, чтобы добиться сколь угодно большой пропускной способности C? К сожалению, это тоже неверно. Прежде всего мы не можем использовать сигналы сколь угодно большой интенсивности, так как при этом нам придется затратить на их создание громадную мощность. Существует строго оп-

ределенная средняя мощность P передаваемого сигнала, однозначно определяемая энергетическим питанием нашей линии связи. Кроме того, мы не можем различить и сигналы, значения которых слишком близки друг к другу. С этим обстоятельством мы уже встречались на стр. 290—292, где максимальная степень близости, при которой сигналы еще можно различить, определялась чисто физиологическими факторами («разрешающей способностью» глаза или уха). В случае технических линий связи прием осуществляется специальными аппаратами, и ценой усложнения и удорожания этих аппаратов их разрешающую способность можно сделать практически сколь угодно высокой, т. е. можно добиться, чтобы наши аппараты различали даже очень близкие между собой сигналы. Но существует еще одна причина, препятствующая различению близких сигналов — помехи. Дело в том, что в любой линии связи существуют помехи, которые никак не могут быть устранены; эти помехи искажают значение передаваемого сигнала. В случае электросвязи, например, эти помехи могут вызываться малыми колебаниями нагрузки в сети, электрическим полем соседней линии, или даже просто тем обстоятельством, что электроны во всех проводниках всегда находятся в случайном «тепловом» движении (зависящем от температуры проводника и вполне аналогичном хаотическому движению молекул газа); в случае радиосвязи они могут создаваться грозовыми разрядами в атмосфере или электрическими разрядами, создаваемыми промышленными или транспортными установками (например, искрением дуги проходящего неподалеку трамвая). Если мы обозначим через W среднюю мощность этих помех (т. е. мощность тех искажений, которым подвергаются наши сигналы в процессе передачи), то те сигналы, разность которых имеет много меньшую чем W мощность, на приемном конце нельзя будет различить никакими аппаратами — небольшая разница между ними будет полностью «смазана» значительно бо́льшими «случайными» искажениями. Поэтому различными здесь оказываются лишь сигналы, отличающиеся не меньше, чем на некоторое определенное значение; так как, кроме того, максимальный уровень наших сигналов (определяющийся средней мощностью сигнала P) также не может быть безгранично велик, то может существовать лишь

конечное число m различных между собой градаций значения сигнала. Количественный анализ возникающей здесь ситуации был произведен Шенноном [1] (см. также [4] или [13]), показавшим, что, вообще говоря, число m можно определить формулой $m = \sqrt{1 + \frac{P}{W}}$. Таким образом, мы приходим к следующему выражению для пропускной способности C произвольной линии, передающей непрерывно изменяющиеся сигналы:

$$C = L_1 \log\left(1 + \frac{P}{W}\right), \quad L_1 = \frac{L}{2} \qquad (*)$$

(где L_1 — некоторая «универсальная» характеристика линии связи, не зависящая от передаваемого сообщения) [1]. Вывод этой замечательной формулы представляет собой один из важнейших вкладов теории информации в общую теорию связи.

Приведенная формула позволяет без труда подсчитать пропускную способность каждой конкретной линии связи; кроме технических характеристик самой линии, при этом надо еще только знать отношение $\frac{P}{W}$ средних мощностей сигнала и помех. Оказывается, что для телетрансляционных линий C обычно имеет порядок десятков миллионов бит/сек.; для телефонных, фототелеграфных и радиотрансляционных линий C измеряется многими тысячами или десятками тысяч бит/сек., а для телеграфных линий — сотнями или несколькими десятками бит/сек. (см., например, [69], [132] или [141]). Существенно при этом, что имеющаяся пропускная способность во всех случаях (кроме, быть может, телеграфа) теоретически позволяет передавать информацию с гораздо большей скоростью, чем та, которая достигается при обычных технических передачах. Так, например, по телеграфу информация обычно передается со скоростью, не превышающей 75 бит/сек.; по телефону — со скоростью, не превышающей 2500 бит/сек; по телевидению — со скоростью, не превышающей 500 000 бит/сек. Таким образом, все реально используемые в настоящее

---

[1]) Мы говорим здесь лишь о пропускной способности линии, передающей *непрерывные* сигналы, поскольку случай передачи дискретных сигналов при наличии помех будет специально рассматриваться в следующем параграфе.

время способы передачи сообщений, как правило, используют лишь небольшую часть пропускной способности существующих линий связи. Более полное использование пропускной способности требует применения значительно более совершенных методов кодирования и декодирования; в этой связи возникает много трудных и научных, и чисто технических проблем, занимающих в настоящее время умы большого числа исследователей во всех странах мира (подробнее об этом будет говориться в заключительном параграфе настоящей главы). Заметим, что достижения последних лет в области теории и практики кодирования и декодирования в принципе позволяют уже сейчас существенно повысить эффективность использования линий связи: так в экспериментальных передачах, специально организованных американскими учеными и инженерами, удалось достигнуть скорости передачи информации по телефону порядка 7500—8000 бит/сек. (см., например, [13], стр. 415, [142] или же [176], стр. 18), а по телевидению — порядка 20 000 000 бит/сек. (см. [142]). Однако и такие скорости передачи информации кажутся все же недостаточными для нужд будущего — общее количество информации, передаваемое по имеющимся линиям связи, в большинстве стран мира стремительно возрастает с каждым годом, а в дальнейшем можно ожидать широкого развития новых типов передачи информации (например, видеотелефона), а также появления двусторонней телевизионной связи между отдельными учреждениями в разных городах и массового использования непосредственной передачи цифровых данных в крупные централизованные вычислительные центры, что приведет к значительному убыстрению этого процесса. Поэтому в настоящее время в ряде лабораторий мира начата разработка совершенно новых видов линий связи, обладающих заметно большими пропускными способностями — в первую очередь металлических и диэлектрических волноводных линий [1]) с пропускными способностями порядка $5 \cdot 10^8$ — $1 \cdot 10^9$ бит/сек. и оптических волноводов из стекловолокна с пропускной способностью

[1]) Волноводы (радио и оптические) представляют собой фактически трубопроводы, по которым распространяются волны. Наличие внешней оболочки позволяет сильно уменьшить уровень помех и вместе с тем использовать очень широкую полосу частот, не создавая препятствий для других линий связи.

порядка 10^8 бит/сек. на одно волокно (о таких проектах говорилось, в частности, в ряде докладов на международной конференции по технике связи в Монреале в июне 1971 г., на международной конференции по теории информации в Цахкадзоре, Арм. ССР, в сентябре 1971 г. и других недавних научных совещаниях, имеющих отношение к науке о связи). Разумеется, реальное внедрение таких новых линий связи требует еще преодоления большого числа технических трудностей,— но сам факт появления такого рода исследований уже представляется знаменательным.

Интересно отметить, что понятие о пропускной способности, возникшее в технике, вполне может быть применено и к тем «линиям связи», по которым каждый живой организм получает информацию от своих органов чувств. В самом деле, мы уже описывали в гл. II специальные психологические эксперименты, показывающие, что время, требуемое для усвоения центральной нервной системой какой-либо информации, прямо пропорционально количеству этой информации; таким образом, здесь выполняются те же закономерности, которые имеют место для всех технических линий связи. В последнее время появились также некоторые работы, обосновывающие применимость к нервным линиям связи в человеческом организме формулы (*) Шеннона (стр. 316); этот вопрос, однако, в настоящее время еще нельзя считать выясненным окончательно.

Пропускная способность C отдельных органов чувств может быть весьма грубо оценена на основе физиологических данных об их разрешающей способности (т. е. об общем количестве объектов, различимых при помощи того или иного органа чувств) и о среднем времени, нужном для восприятия (т. е. о максимальной частоте смены внешних воздействий, при которой эти воздействия все еще могут восприниматься раздельно). Таким образом удается, в частности, показать, что пропускная способность разных органов чувств резко различна: человеческий глаз при благоприятных условиях освещения, по-видимому, способен воспринимать (и передавать центральной нервной системе) информацию со скоростью порядка миллионов (или десятков миллионов) бит/сек., в то время как ухо воспринимает информацию с гораздо меньшей скоростью порядка тысяч или десятков тысяч бит/сек. (см., например, [143] — [146]). Столь различная пропускная способность

отчасти может объясняться резким различием числа нервных волокон, обслуживающих слух и зрение (по современным физиологическим данным число «ушных нервных волокон» имеет порядок 30 000, против примерно 800 000—900 000 «глазных нервных волокон»). Осязание же, повидимому, по своей способности воспринимать и передавать информацию находится где-то в промежутке между зрением и слухом. Надо, однако, отметить, что лишь очень небольшая часть передаваемой органами чувств информации может быть сознательно усвоена человеческим мозгом; это ясно следует, например, из приведенных на стр. 277 данных о скорости восприятия информации при разговоре (мы отмечали там, что при быстром разговоре часть «несмысловой» информации пропадает, так как человек не успевает ее воспринять). Тщательный анализ результатов, касающихся максимальной достижимой скорости разговора, чтения, письма (стенографического) и т. д. показывает, что во всех случаях человек способен усвоить поступающую информацию лишь если скорость ее поступления не превосходит примерно 50 бит/сек. (см., например, [147] и [148]) [1]). Величина того же порядка получается и при определении количества информации, усваиваемой зрителем при просмотре быстро мелькающих на экране кадров [149]. Наконец, специально поставленные опыты по определению минимального времени физиологических реакций (ср. стр. 83 и след.), достигаемого при наиболее благоприятных условиях восприятия, также показывают, что пропускная способность центральной нервной системы человека по порядку величины равна 30—40 бит/сек. (см. [148], [150]). Разумеется, в отношении дальнейшего уточнения этих чисел и выяснения их зависимости от индивидуальных особенностей человека и его физического и психического состояния остается сделать еще очень много [2]); однако сам факт плодотворности применения

---

[1]) Напомним еще, что в соответствии со сказанным на стр. 277 при нормальном разговоре лишь около половины воспринимаемой слушателем информации содержится в записи речи его собеседника; остальная же информация касается голоса говорящего, его эмоций, смысловых ударений и т. д.

[2]) См., в частности, обзор этого вопроса в книге [41] и имеющиеся там ссылки на оригинальную литературу, содержащую множество противоречащих друг другу данных.

общих идей теории информации к изучению первой деятельности человека и животных теперь уже не вызывает сомнений.

Общая схема передачи по линии связи. Передача генетической информации

В заключение настоящего параграфа скажем еще несколько дополнительных слов об общей схеме передачи сообщений по линии связи, с которой мы фактически начали § 1 этой главы. Процесс передачи сообщений по произвольной линии связи можно схематически изобразить следующим образом:

В случае, например, передачи текстовых сообщений по телеграфу сообщения α_1 и β_1 на входе и на выходе записываются на определенном (одном и том же!) языке с помощью соответствующего буквенного алфавита и могут отличаться друг от друга лишь в результате каких-то искажений в процессе передачи, а сигналы α и β на входе и на выходе представляют собой последовательности электрических «элементарных сигналов» (обычно — посылок тока и пауз). Таким образом, операции кодирования и декодирования здесь состоят в преобразовании буквенного сообщения α_1 в последовательность «элементарных сигналов» α и в обратном переходе от принятой последовательности β «элементарных сигналов» к буквенному сообщению β_1. В телефонии сообщение α_1 имеет характер звука, т. е. определенных колебаний давления; кодирование здесь состоит в преобразовании этих колебаний давления в колебания электрического тока, а декодирование — в обратном преобразовании принятых колебаний тока в звук. В линии связи современной электронной вычислительной машины сигнал α_1 на входе представляет собой определенную последовательность чисел, кодирование состоит в

его преобразовании в определенную последовательность α электрических сигналов, непосредственно вводимых в машину, а декодирование — в преобразовании поступивших в машину сигналов β (представляющих собой сумму «вводимых сигналов» α и «искажений в процессе ввода»), приводящем к совсем новому сообщению β_1— ответу задачи, решаемой машиной; здесь уже β_1 принципиально отличается от α_1 и преобразование α_1 в β_1 составляет основную цель нашей линии связи. Точно так же и в случае передачи зрительного «сообщения» по нервным волокнам α_1 и β_1 резко отличаются друг от друга — здесь α_1 состоит из совокупности световых волн разной длины волны (т. е. разного цвета) и разной амплитуды (т. е. интенсивности), а β_1 представляет собой совокупность возбуждений определенных нервных клеток (нейронов) головного мозга (так называемых «зрительных нейронов»), воспринимаемых нами как некоторая зрительная картина. Сигнал α в этом частном случае представляет собой совокупность электрических импульсов, вырабатываемых приемниками света (колбочками и палочками) глаза, а кодирование состоит в преобразовании света в такие импульсы и пока изучено довольно плохо; декодирование же здесь состоит в переходе от электрических импульсов β, дошедших до мозга по нервным волокнам, к возбуждениям нейронов β_1, и его детали известны еще значительно хуже, чем детали кодирования.

Общие вопросы, касающиеся описания произвольной линии связи при наличии помех и определения теоретических границ возможностей использования таких линий для передачи информации, будут нами рассмотрены в следующем § 4, а заключительный § 5 будет посвящен введению в обширную теорию оптимального кодирования и декодирования дискретных сообщений, передаваемых по линиям связи с помехами. Сейчас же мы лишь отметим, что во многих случаях даже вопрос об изучении самой «азбуки», на которой записываются сообщения α_1 и β_1, и о природе передаваемых «элементарных сигналов» α представляет очень большой интерес и является совсем не простым. Наиболее ярким примером здесь является проблема передачи г е н е т и ч е с к о й информации, успехи в изучении которой относятся к числу крупнейших научных достижений двух последних десятилетий.

Ввиду общенаучной важности этой проблемы и ее тесной связи с общей постановкой задачи о передаче информации, быть может уместно остановиться здесь на относящихся сюда результатах немного более подробно. «Линии связи», сопоставляемые явлениям наследственности, играют первостепенную роль в самом существовании органической жизни. По этим линиям постоянно с поразительной точностью передается огромная и крайне важная информация. Всего на Земле зарегистрировано около 2 млн. отдельных видов животных и растений — и по рассматриваемым «линиям связи» безошибочно передаются сигналы, указывающие, какой именно вид должен развиться из единственной зародышевой клетки. При этом передаваемая информация отнюдь не ограничивается одним лишь указанием вида — она содержит также достаточно исчерпывающие сведения об особенностях строения вида и, кроме того, множество данных, касающихся наследственных особенностей индивидуального организма, развившегося из данной клетки. Вся эта информация сохраняется где-то в ничтожном объеме ядра зародышевой клетки и передается какими-то достаточно сложными путями телу («цитоплазме») как исходной клетки, так и всех прочих клеток, возникающих из данной путем деления; она сохраняется также и в процессе дальнейшего воспроизведения последующих поколений аналогичных особей.

Строение соответствующих линий связи и методы передачи информации по ним еще не так давно казались совершенно таинственными и быстрое продвижение в этой области, связанное с громадными успехами молекулярной биологии в период после последней мировой войны, мало кто мог предвидеть. Основную роль здесь сыграло открытие фундаментальной роли колоссальных полимерных молекул так называемой *дезоксирибонуклеиновой кислоты* (сокращенно ДНК), располагающихся в хромосомах ядра клетки. Известно, что эти молекулы состоят из длинной цепи чередующихся углеводных и фосфатных групп одинакового состава, причем к каждой углеводной группе присоединено еще некоторое одно азотистое основание из числа четырех возможных оснований такого типа, называемых *аденин*, *гуанин*, *цитозин* и *тимин*. Все допустимые различия в молекулах ДНК ограничиваются различиями в последовательном чередовании соответствующих оснований (которые, для краткости, можно обозначать их первыми буквами *А*, *Г*, *Ц* и *Т*, а можно и просто занумеровать цифрами 0, 1, 2 и 3). Таким образом, исходное «сообщение» α_i здесь хранится в хромосомах ядра клетки и записано на «четырехбуквенном алфавите» молекул ДНК. Одна молекула ДНК в хромосоме может содержать несколько десятков тысяч или даже более углеводных групп (а, следовательно, и оснований), а число отдельных хромосом в ядре клетки может равняться нескольким десяткам; таким образом, количество информации, которое может быть запасено в хромосомах, имеет порядок

$$\log 4^{100\,000} = 200\,000 \text{ бит}$$

(или еще больше). Этого количества информации с избытком хватает для хранения всех передающихся по наследству данных.

На самом деле строение хромосом является еще несколько более сложным — каждая хромосома представляет собой не одинар-

пую, а д в о й н у ю нить ДНК, составленную из двух таких молекул, свернутых в форме двух спиралей, навивающихся в противоположных направлениях на один (реально не существующий) цилиндр. Эти две молекулы ДНК являются не одинаковыми, а «дополнительными» — аденину в одной из них всегда отвечает в другой тимин, а гуанину — цитозин; соответствующие пары оснований, расположенные на цилиндре друг против друга, связаны между собой сравнительно слабыми водородными связями. Такое «двойное» строение хромосом играет основную роль в процессе их воспроизведения при делении клеток («митозе»), когда каждая из двух новых клеток приобретает свой набор хромосом, идентичный набору хромосом исходной клетки; этот процесс, по-видимому, связан с «развертыванием» двух входящих в хромосому нитей ДНК, при котором две длинные молекулы ДНК расходятся между собой и каждая присоединяет затем к себе еще одну «дополнительную» молекулу, образуя самостоятельную двойную спираль. Происходящая таким путем передача информации от родительских клеток к дочерним играет фундаментальную роль во всех жизненных явлениях; здесь роль передаваемого «сообщения» α_1 играет набор хромосом (набор молекул ДНК) исходной клетки, а в качестве «сообщения на выходе» β_1 выступают наборы хромосом двух новых клеток. Полученное «сообщение на выходе» β_1 непосредственно из «сообщения на входе» α_1 снимает в этом случае вопросы о кодировании и декодировании «сообщений». В то же время вопрос о «помехах» в нашей линии связи является необычайно важным, ибо возникающие в результате этих «помех» (роль которых может играть, например, радиоактивное облучение клетки) искажения представляют собой изменения наследственных признаков («мутации»), играющие основную роль в процессе эволюции органических видов.

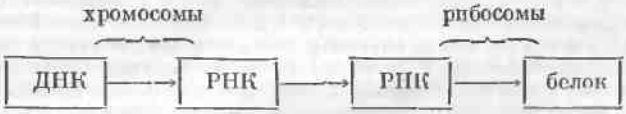
Перейдем теперь к передаче информации от хромосом к телу («цитоплазме») клетки, определяющей процесс построения из одной зародышевой клетки целой особи данного конкретного вида. Основную роль во всех жизненных функциях организма играют *белковые* вещества, в частности *ферменты*, управляющие всеми происходящими в живых организмах биохимическими реакциями. Синтез белка происходит во вкрапленных в цитоплазму клеток так называемых р и б о с о м а х; скорость этого синтеза достигает порядка одной молекулы белка в минуту. При этом строение белковых молекул также является довольно простым — все белки построены из примерно 20 различных *аминокислот*, чередующихся в определенном порядке вдоль линейной молекулы белка; эти аминокислоты перечислены в таблице на следующей странице вместе с принятыми в биохимии сокращениями их названий.

Таким образом, можно сказать, что приемным концом («выходом») рассматриваемой здесь линии связи служат рибосомы; «сообщение на выходе» β_1 представляет собой в этом случае белок и записано оно на «двадцатибуквенном алфавите» аминокислот. Остается еще только установить, как происходит перенос информации от ДНК к белкам, в частности, что надо понимать под «сигналом на входе» α и «сигналом на выходе» β.

На последний вопрос также можно дать сегодня вполне удовлетворительный ответ. Основную роль в процессе передачи информации

Аминокислота	Сокращенное обозначение	Аминокислота	Сокращенное обозначение
Аланин	*Ала*	Лейцин	*Лей*
Аргинин	*Арг*	Лизин	*Лиз*
Аспарагин	*Асн*	Метионин	*Мет*
Аспарагиновая кислота	*Асп*	Пролин	*Про*
Валин	*Вал*	Серин	*Сер*
Глутамин	*Глн*	Треонин	*Тре*
Глутаминовая кислота	*Глу*	Триптофан	*Три*
Глицин	*Гли*	Тирозин	*Тир*
Гистидин	*Гис*	Фенилалалин	*Фен*
Изолейцин	*Иле*	Цистеин	*Цис*

от ДНК хромосом к белковым молекулам играет еще одна нуклеиновая кислота — так называемая *рибонуклеиновая кислота* (сокращенно РНК). Строение РНК весьма близко к строению ДНК — только углеводная группа здесь немного иная и роль тимина играет иное основание — *урацил*, немного отличающееся от тимина по химическому составу. Таким образом, молекулу РНК можно рассматривать как «сигнал», закодированный с помощью четырех «элементарных сигналов» *А*, *Г*, *Ц* и *У* (или 0, 1, 2 и 3'), весьма близких к «буквам» исходного «сообщения» *А*, *Г*, *Ц* и *Т*. На молекулах ДНК хромосом, как на некотором «шаблоне», синтезируются определенные линейные молекулы РНК (так называемая «информационная» РНК или иРНК), которые в дальнейшем выделяются из ядра клетки и проникают в рибосомы; эти молекулы иРНК и играют основную роль в процессе синтеза белка. Таким образом, изображенная на стр. 320 общая схема передачи информации по линии связи в рассматриваемом случае имеет следующий вид:



Здесь роль «сообщения на входе» α_1 и «сообщения на выходе» β_1 играют ДНК и белок, а роль «сигнала на входе» α и «сигнала на выходе» β — молекулы иРНК.

Согласно приведенной схеме «передаваемое сообщение» α_1 записано на «четырехбуквенном алфавите», а «принимаемое сообщение» β_1 — на «двадцатибуквенном алфавите», так что для нашей линии связи число m элементарных сигналов, поступающих на «вход» линии, и число r элементарных сигналов, принимаемых на «выходе», различны ($m = 4$, а $r = 20$); «коды» же, с помощью которых записаны «сигналы» α и β, имеют четыре «элементарных сигнала». Что же касается операций кодирования и декодирования, т. е.

преобразований «сообщения» α_1 в «сигнал» α и «сигнала» β в «сообщение» β_1, то они были изучены, в основном, лишь сравнительно недавно. Естественно, что более простой (а потому — и менее интересной) из перечисленных выше двух операций является операция «кодирования», сводящаяся к преобразованию последовательности чередующихся четырех «букв» *А*, *Г*, *Ц* и *Т* в последовательность четырех «элементарных сигналов» *А*, *Г*, *Ц* и *У*. Здесь можно указать много простых и априорно допустимых систем кодирования: так, например, своеобразная «дополнительность» определенных пар оснований, проявляющаяся, в частности, в строении «двойных» молекул ДНК, предсказывает вариант, при котором гуанин «порождает» цитозин, цитозин — гуанин, тимин — аденин и аденин — урацил. По-видимому, именно такое кодирование, в основном, и осуществляется в природе, хотя, возможно, оно и не является совершенно универсальным [1]).

Значительно больший интерес представляет в нашем случае «декодирование», состоящее в переходе от «четырехбуквенного языка» иРНК к «двадцатибуквенному языку» белка; именно его поэтому и имеют обычно в виду, когда говорят о «генетическом коде». Ясно, что одно основание иРНК, которое может принимать всего четыре «значения»— *А*, *Г*, *Ц* или *У*,— никак не может содержать полной информации об одной из двадцати возможных аминокислот. Поэтому приходится считать, что одну аминокислоту определяет последовательность из *нескольких* соседних оснований в молекуле РНК; такую последовательность оснований, «кодирующую» одну «букву» алфавита аминокислот, принято называть кодоном. Так как число различных последовательностей из *двух* оснований РНК равно $4 \cdot 4 = 16$, что меньше числа разных аминокислот, то кодон должен содержать не меньше трех оснований; *три* же основания он содержать вполне может, так как число всевозможных троек оснований равно $4 \cdot 4 \cdot 4 = 64$, что заметно больше двадцати.

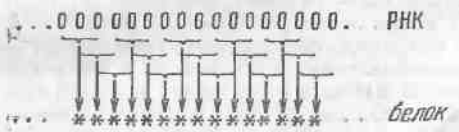
Первая гипотеза о природе генетического кода была предложена в 1954 г. известным американским физиком и астрофизиком Г. Гамовым [151]. Гамов предположил, что заданная аминокислота в белковой цепи определяется некоторой тройкой соседних оснований РНК, скажем — первым, вторым и третьим основаниями, следующая аминокислота — сдвинутой на единицу тройкой, т. е. вторым, третьим и четвертым основаниями, еще следующая — сдвинутой на два основания тройкой и т. д.; такой код с частично перекрывающимися кодонами получил название «перекрывающегося кода» (см. схему на следующей стр., где нуликами обозначены основания, а звездочками — аминокислоты). При этом предполагалось, что аминокислота белка зависит только от состава соответствующего кодона, но не от порядка отдельных оснований в кодоне. Основным аргументом в пользу этой гипотезы для Гамова явилось то, что число различных по составу троек, которые можно образовать

---

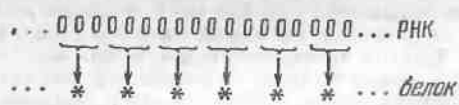
[1]) Так, например, существуют вирусы, у которых вообще роль молекул ДНК играют длинные молекулы РНК, так что здесь «сообщение на входе» α_1 с самого начала записано в «алфавите» *А*, *Г*, *Ц*, *У*.

из четырех оснований, равно

$$\underbrace{C_4^3}_{\text{число троек попарно различных оснований}} + \underbrace{2C_4^2}_{\text{число троек оснований, содержащих два одинаковых основания}} + \underbrace{C_4^1}_{\text{число троек из трех одинаковых оснований}} = 20.$$



Предложенный Г. Гамовым «перекрывающийся код», так же как и выдвинутый совместно Г. Гамовым и М. Ичасом [152] «неперекрывающийся комбинационный код» (см. схему внизу), в котором тоже аминокислота белка определялась единственно составом кодонов, но не расположением в нем оснований,



оказались не соответствующими действительности. Однако четко поставленная Г. Гамовым задача описания процесса синтеза белков в живой клетке как согласующегося с экспериментальными данными «перевода» сигнала β, записанного на четырехбуквенном языке РНК, в сообщение β_1, записанное на двадцатибуквенном языке белков, сыграла большую роль в дальнейших успехах этой области молекулярной биологии.

С «комбинационным кодом» Г. Гамова и М. Ичаса одно время конкурировала выдвинутая знаменитым Ф. Криком и его сотрудниками [153] идея «кода без запятой», довольно долго широко обсуждавшаяся многими учеными разных специальностей (см., например, примыкающую сюда статью математиков С. Голомба, Л. Велча и генетика М. Дельбрюка [154]). Термин «код без запятой» здесь понимается немного иначе, чем на стр. 187, где фактически под этим понимался произвольный однозначно расшифровываемый код — равномерный код, состоящий только из трехбуквенных кодонов, этому последнему условию, очевидно, всегда будет удовлетворять. Но дело в том, что если мы допустим, что код — неперекрывающийся, то неясно, как именно распознается конец одного кодона и начало следующего — ведь в принципе одну и ту же последовательность оснований, скажем, ...*АГЦЦЦА*... можно по-разному разбить на трехбуквенные «кодоны»: ее можно «прочесть» и как ...(*АГЦ*) (*ЦТЦ*) (*А*..., и как ...*АГ*) (*ГЦТ*) (*ЦА*..., и как ...*А*) (*ГГЦ*) (*ТЦА*).... Можно указать три возможности избежать возникающей таким образом неопределенности. В принципе может существовать какой-то особый знак, указывающий начало

считывания последовательности кодонов[1]). Возможно также существование специальной последовательности оснований (быть может, содержащей большее или меньшее число оснований, чем отвечающие аминокислотам кодоны), отделяющей отдельные кодоны друг от друга — подобная последовательность оснований расшифровывается как «запятая», отделяющая друг от друга «слова» (кодоны). Наконец, специалисты по теории связи знают и такие «коды без запятой», что произвольная последовательность «букв» (в нашем случае — оснований ДНК) допускает лишь о д н у возможность ее осмысленного прочтения; другие же варианты разбиения этой последовательности «букв» на отдельные «слова» приводят к последовательности бессмысленных сочетаний «букв», не отвечающих никаким «словам».

Ясно, что так определенный «код без запятой» должен быть «неполным»— в нем должны существовать последовательности букв, не отвечающие никаким «словам» (не составляющие кодонов). Принимая, что каждый кодон состоит из трех оснований (т р и п л е т н ы й код), мы легко найдем наибольшее возможное число осмысленных кодонов. Ясно, что «триплеты», состоящие из трех одинаковых «букв» (оснований), например, *ААА* не могут иметь смысла, ибо иначе длинная последовательность соответствующих «букв»— ...*АААААААА*...— могла бы осмысленно считываться, начиная с л ю б о г о места. Остающиеся 64—4 = 60 разных триплетов можно разбить на 20 групп по 3 триплета, получающихся друг из друга «циклической перестановкой букв» (оснований)— таковы, скажем, триплеты *АГЦ*, *ГЦА* и *ЦАГ* или *ЦЦТ*, *ЦТЦ* и *ТЦЦ*. Ясно, что из этих трех триплетов смысл может иметь только один, ибо в противном случае также нельзя было бы однозначно определить, с какого места надо начать считывание кодонов в длинной последовательности одинаковых триплетов одного из этих видов. Таким образом, наибольшее возможное число осмысленных кодонов в случае триплетного кода без запятой не может превышать 60 : 3 = 20 — и можно показать, что оно в точности р а в н о 20. В этом обстоятельстве Ф. Крик и разделяющие его точку зрения исследователи видели один из веских аргументов в пользу своей гипотезы.

Решение вопроса о строении «генетического кода» было найдено, однако, не за письменным столом, а непосредственно в лабораториях. В начале 60-х годов (в 1961—1963 гг.) группе биохимиков, возглавляемой американцем М. Н и р е н б е р г о м, удалось показать, что синтез цепочек аминокислот, во всем напоминающих белок, можно осуществить и в отсутствии живых клеток, выделив отдельно рибосомы живых клеток, поместив их в органическую среду, содержащую основные компоненты среды цитоплазмы, и добавляя синтетическую РНК заданного состава, в процессе синтеза белка играющую роль информационной РНК живой клетки.

---

[1]) Заметим сразу же, что, видимо, именно этот вариант и реализуется в действительности, хотя детали «указания», предписывающего именно с данного основания начать «считывание» кодонов, пока остаются неясными.

В первом опыте такого рода, осуществленном М. Ниренбергом и Г. Маттеи, синтетическая РНК содержала одно только повторяющееся урацилловое основание; при этом наблюдался синтез искусственного белка, состоящего из многократно повторяющейся аминокислоты фенилаланина (*Фен*). Таким образом РНК ...*УУУУУУУУУ*... порождала белок ...*ФенФенФенФен*..., откуда следовало, что если код является триплетным, то кодону *УУУ* должна соответствовать аминокислота *Фен*. Аналогично было установлено, что кодону *ЦЦЦ* отвечает аминокислота пролин (*Про*).

В течение всех 60-х годов в многочисленных биохимических лабораториях мира велся широкий «штурм» проблемы генетического кода; из числа участвующих в этом исследователей, кроме М. Ниренберга и его сотрудников (из которых особо большую роль сыграл Ф. Ледер), следует упомянуть работающих в США индуса Г. Х. Хорана и мексиканца С. Очоа. Мы не станем рассказывать здесь об этом подробно, отослав желающих к старым обзорам Г. Гамова, М. Ичаса и А. Рича [155], дающим возможность познакомиться с ранними этапами попыток расшифровки генетического кода, к (тоже довольно старым) статьям Ф. Крика, М. Ниренберга и др. [156], рассчитанным на широкого читателя, и, особенно, к обстоятельной монографии М. Ичаса [157], список литературы к которой содержит 869 названий. Трудами многих ученых было установлено, что генетический код действительно является *триплетным* и *неперекрывающимся*; что он является *«вырожденным»* в том смысле, что некоторым аминокислотам отвечают сразу несколько разных кодонов; что существуют *«бессмысленные»* (т. е. не несущие генетической информации)

Кодоны	Аминокислоты	Кодоны	Аминокислоты	Кодоны	Аминокислоты	Кодоны	Аминокислоты
УУУ	*Фен*	*УЦУ*	*Сер*	*УГУ*	*Цис*	*УАУ*	*Тир*
УУЦ	*Фен*	*УЦЦ*	*Сер*	*УГЦ*	*Цис*	*УАЦ*	*Тир*
УУА	*Лей*	*УЦА*	*Сер*	*УГА*	—	*УАА*	—
УУГ	*Лей*	*УЦГ*	*Сер*	*УГГ*	*Три*	*УАГ*	—
ЦУУ	*Лей*	*ЦЦУ*	*Про*	*ЦГУ*	*Арг*	*ЦАУ*	*Гис*
ЦУЦ	*Лей*	*ЦЦЦ*	*Про*	*ЦГЦ*	*Арг*	*ЦАЦ*	*Гис*
ЦУА	*Лей*	*ЦЦА*	*Про*	*ЦГА*	*Арг*	*ЦАА*	*Глн*
ЦУГ	*Лей*	*ЦЦГ*	*Про*	*ЦГГ*	*Арг*	*ЦАГ*	*Глн*
АУУ	*Иле*	*АЦУ*	*Тре*	*АГУ*	*Сер*	*ААУ*	*Асн*
АУЦ	*Иле*	*АЦЦ*	*Тре*	*АГЦ*	*Сер*	*ААЦ*	*Асн*
АУА	*Иле*	*АЦА*	*Тре*	*АГА*	*Арг*	*ААА*	*Лиз*
АУГ	*Мет*	*АЦГ*	*Тре*	*АГГ*	*Арг*	*ААГ*	*Лиз*
ГУУ	*Вал*	*ГЦУ*	*Ала*	*ГГУ*	*Гли*	*ГАУ*	*Асп*
ГУЦ	*Вал*	*ГЦЦ*	*Ала*	*ГГЦ*	*Гли*	*ГАЦ*	*Асп*
ГУА	*Вал*	*ГЦА*	*Ала*	*ГГА*	*Гли*	*ГАА*	*Глу*
ГУГ	*Вал*	*ГЦГ*	*Ала*	*ГГГ*	*Гли*	*ГАГ*	*Глу*

триплеты, которые вообще не являются кодонами в том смысле, что им не отвечает ни одна аминокислота [1]).

Таблица на стр. 328 указывает, как представляют себе сегодня ученые генетический код (черточка в левом столбце означает, что соответствующий триплет не является кодоном).

§ 4. Передача сообщений при наличии помех

В двух первых параграфах настоящей главы на примере телеграфии были рассмотрены некоторые общие вопросы теории передачи сообщений по линиям связи. При этом, однако, все время подразумевалось, что сигналы передаются по линии связи без всяких искажений, т. е. что передача ведется в отсутствие помех. Между тем в практике связи так фактически никогда не бывает: всегда возможны некоторые помехи, вызывающие искажение сигнала в процессе передачи. Кратко об этом уже упоминалось в § 3 в связи с анализом работы линий связи, передающих непрерывные сообщения (см. стр. 315—316). В настоящем параграфе мы снова вернемся к простейшей схеме дискретной линии связи, рассмотренной в §§ 1 и 2, т. е. будем предполагать, что по линии передается лишь конечное число различных «элементарных сигналов» постоянной длительности (в самом простом случае лишь два различных сигнала — посылка тока и пауза). Но, в отличие от §§ 1 и 2, теперь мы уже не будем пренебрегать влиянием помех, т. е. будем учитывать возможность путаницы — элементарный сигнал одного типа в результате искажений, вносимых помехами, может быть ошибочно принят на приемном конце как сигнал другого типа (например, посылка тока может восприниматься как пауза, а пауза — как посылка тока). Посмотрим, что может дать теория информации в применении к этому более сложному (но зато и более реальному) случаю.

Будем для простоты, как и в § 2, предполагать, что последовательные «буквы» сообщения взаимно независимы, причем n букв алфавита характеризуются определенными вероятностями $p_1, p_2, \ldots, p_n$ появления на любом месте сообщения той или иной буквы. Рассмотрим линию связи, в которой для передачи используется m различных

---

[1]) Но которые тем не менее имеют определенный генетический смысл (см. по этому поводу гл. VIII книги [157]).

элементарных сигналов $A_1, A_2, \ldots, A_m$, причем за единицу времени может быть передано L таких сигналов (т. е. длительность одного сигнала равна $\tau = \frac{1}{L}$). Тогда, согласно основному результату § 2, *при отсутствии помех сообщение по нашей линии связи можно передавать со скоростью, сколь угодно близкой к величине*

$$v = \frac{C}{H} \text{ букв./ед. времени}$$

(где

$$C = L \log m$$

— пропускная способность линии связи, а

$$H = -p_1 \log p_1 - p_2 \log p_2 - \ldots - p_n \log p_n$$

— энтропия одной буквы передаваемого сообщения); однако *скорость передачи, превосходящая* v, здесь *никогда не может быть достигнута*. При этом для достижения скорости передачи, очень близкой к v, надо лишь разбить передаваемое сообщение на достаточно длинные блоки и воспользоваться, например, для передачи отдельных блоков оптимальным кодом Хаффмана или же каким-либо близким к оптимальному кодом (скажем, кодом Шеннона — Фано или кодом с такими длинами l_i кодовых обозначений, что $-\log p_i/\log m \leqslant l_i < -\log p_i/\log m + 1$). Иначе говоря, для этого надо воспользоваться кодом, для которого избыточность в закодированном сообщении будет наименьшей возможной или, по крайней мере, достаточно близкой к таковой.

При наличии помех в линии связи дело будет обстоять несколько иначе. Естественно, что в этом случае только наличие избыточности в передаваемой последовательности сигналов может помочь нам точно восстановить переданное сообщение по принятым данным: в случае значительных помех мы даже, наоборот, стремимся еще больше увеличить избыточность, например, повторяя каждое переданное слово по нескольку раз или заменяя каждую букву сообщения отдельным словом, начинающимся с этой буквы (передача «по буквам»). Ясно, что использование кода, приводящего к наименьшей избыточности закодированного сообщения, здесь уже будет

нецелесообразным и скорость передачи сообщения должна быть уменьшена. Насколько же придется ее уменьшить?

Для ответа на этот вопрос нам придется предварительно разобрать, как математически описывается линия связи, в которой имеются какие-то помехи. Предположим сперва, что рассматриваемая линия связи использует m различных элементарных сигналов $A_1, A_2, \ldots, A_m$, но из-за наличия помех переданный сигнал A_i (где $i = 1, 2, \ldots,$ или m) может быть иногда принят на приемном конце линии связи за какой-то другой (отличный от A_i) сигнал A_j. Для того чтобы количественно описать эту ситуацию, надо задать вероятность $p_{A_1}(A_1)$ того, что, передавая сигнал A_1, мы на приемном конце получим правильный сигнал A_1 (так что $p_{A_1}(A_1)$ — это вероятность безошибочной передачи сигнала A_1), и вероятности $p_{A_1}(A_2), p_{A_1}(A_3), \ldots$ $\ldots, p_{A_1}(A_m)$ того, что переданный сигнал A_1 будет на приемном конце расшифрован как $A_2, A_3, \ldots, A_m$. Далее надо задать вероятности $p_{A_2}(A_1), p_{A_2}(A_2), \ldots, p_{A_2}(A_m)$ получения на приемном конце сигналов $A_1, A_2, \ldots, A_m$, если на самом деле передавался сигнал A_2, и т. д. вплоть до вероятностей $p_{A_m}(A_1), p_{A_m}(A_2), \ldots, p_{A_m}(A_m)$ того, что на приемном конце будут получены сигналы $A_1, A_2, \ldots$ $\ldots, A_m$, если на самом деле передавался сигнал A_m. Вероятности

$$p_{A_1}(A_1), p_{A_1}(A_2), \ldots, p_{A_1}(A_m);$$
$$p_{A_2}(A_1), p_{A_2}(A_2), \ldots, p_{A_2}(A_m);$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$p_{A_m}(A_1), p_{A_m}(A_2), \ldots, p_{A_m}(A_m)$$

в рассматриваемом случае статистически характеризуют помехи, имеющиеся в нашей линии связи, т. е. они являются математическими характеристиками рассматриваемой линии. Таким образом, полное математическое описание *линии связи с помехами*, приводящими к тому, что передаваемые сигналы могут иногда неправильно расшифровываться на приемном конце, состоит в задании целого числа m, указывающего, сколько различных элементарных сигналов можно передавать по этой линии, числа L (или $\tau = \frac{1}{L}$), определяющего скорость передачи

элементарных сигналов, и еще m^2 неотрицательных чисел $p_{A_i}(A_j)$ (которые, очевидно, должны удовлетворять m условиям: $p_{A_i}(A_1) + p_{A_i}(A_2) + \ldots + p_{A_i}(A_m) = 1$ при всех значениях $i = 1, 2, \ldots, m$), характеризующих влияние помех. Напомним в этой связи, что в §§ 1 и 2 разные линии связи различались лишь числом m используемых элементарных сигналов и (см. конец § 2) скоростью их передачи L.

Приведенное описание линии связи с помехами можно также еще несколько обобщить, допустив, что помехи могут иногда так исказить передаваемый сигнал, что на приемном конце его нельзя будет отождествить н и с о д н и м из m используемых элементарных сигналов A_i. Для того чтобы учесть также и такую возможность, целесообразно допустить, что на приемном конце могут быть получены не обязательно те же m элементарных сигналов $A_1, A_2, \ldots A_m$, которые передавались по линии, а совсем другие какие-то r (где r может быть и больше m, и меньше m, и равно m) элементарных сигналов $B_1, B_2, \ldots, B_r$ (все или некоторые из которых могут отличаться от сигналов $A_1, A_2, \ldots, A_m$; ср. пример 4° ниже). В таком случае помехи будут статистически характеризоваться mr неотрицательными числами

$$p_{A_1}(B_1),\ p_{A_1}(B_2), \ldots, p_{A_1}(B_r);$$
$$p_{A_2}(B_1),\ p_{A_2}(B_2), \ldots, p_{A_2}(B_r);$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$p_{A_m}(B_1),\ p_{A_m}(B_2), \ldots, p_{A_m}(B_r),$$

удовлетворяющими m условиям: $p_{A_i}(B_1) + p_{A_i}(B_2) + \ldots + p_{A_i}(B_r) = 1$ при всех $i = 1, 2, \ldots, m$; через $p_{A_i}(B_j)$ здесь обозначается вероятность того, что на приемном конце будет принят сигнал B_j, если на самом деле был передан сигнал A_i. Вся же линия связи будет теперь характеризоваться целыми числами m и r, числом L (или $\tau = \frac{1}{L}$) и mr числами $p_{A_i}(B_j)$. Использование такого более общего описания линии связи нисколько не усложняет всех последующих рассуждений по сравнению с тем случаем, когда полагается, что $r = m$ и сигналы на при-

емном конце линии совпадают с передаваемыми сигналами $A_1, A_2, \ldots, A_m$; именно его мы в дальнейшем и будем применять [1]).

Предположим теперь, что $p(A_1)$ — это вероятность того, что передаваемым сигналом является сигнал A_1, $p(A_2)$ — вероятность, что им является сигнал $A_2, \ldots$ $\ldots, p(A_m)$ — вероятность, что им является сигнал A_m (где, очевидно, $p(A_1) + p(A_2) + \ldots + p(A_m) = 1$). В таком случае опыт β, состоящий в определении того, какой именно сигнал передается, будет иметь энтропию $H(\beta)$, равную

$$H(\beta) = -p(A_1)\log p(A_1) - p(A_2)\log p(A_2) - \ldots$$
$$\ldots - p(A_m)\log p(A_m).$$

Опыт α, состоящий в выяснении того, какой сигнал при этом будет принят на приемном конце, будет, очевидно, опытом с r исходами, зависимым от опыта β; условная вероятность исхода B_j этого нового опыта при условии, что опыт β имел исход A_i (где $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots$ $\ldots, r$), как раз и равняется $p_{A_i}(B_j)$. Средняя информация об опыте β, содержащаяся в опыте α, равна

$$I(\alpha, \beta) = H(\beta) - H_\alpha(\beta),$$

где $H_\alpha(\beta)$ — условная энтропия, определяющаяся из формул, приведенных на стр. 90—91 (с заменой в этих формулах k и l на m и r). Разумеется, информация $I(\alpha, \beta)$ всегда не больше энтропии $H(\beta)$ опыта β, т. е. той наибольшей информации об опыте β, которую только можно получить и которая содержится, например, в самом этом опыте. Информация $I(\alpha, \beta)$ равна энтропии $H(\beta)$ только в том случае, когда исход опыта α однозначно определяет исход опыта β, т. е. когда по принятому сигналу всегда можно однозначно выяснить, какой сигнал

---

[1]) Вообще говоря, можно даже еще несколько обобщить и это описание, допустив, что на приемном конце может быть получено произвольное (т. е., например, бесконечное или даже непрерывное) множество различных сигналов B. На этот случай также можно перенести почти все указанные ниже результаты, но только здесь уже ряд формул будет выглядеть более сложно; поэтому указанного обобщения понятия линии связи мы ниже вовсе не будем касаться.

был передан (с практической точки зрения это означает, что помехи здесь совсем не препятствуют нормальному приему); информация $I(\alpha, \beta)$ равна нулю в том случае, когда опыт α не зависит от β (т. е. когда принятые сигналы вовсе не зависят от того, какие сигналы передавались — из-за очень сильных помех никакой передачи сообщений фактически вообще не происходит).

Напомним теперь, что при отсутствии помех пропускная способность C линии связи определялась как *наибольшее количество единиц информации, которое можно передать по этой линии за единицу времени* (см. стр. 231). Постараемся обобщить это определение на случай линии связи с помехами. Для такой линии среднее количество информации, получаемой на приемном конце при приеме одного элементарного сигнала, равно величине

$$I(\alpha, \beta) = H(\beta) - H_\alpha(\beta),$$

зависящей от вероятностей $p(A_1), p(A_2), \ldots, p(A_m)$ того, что передан сигнал $A_1, A_2, \ldots, A_m$. Пусть

$$c = \max I(\alpha, \beta)$$

есть самое большое значение информации $I(\alpha, \beta)$, которое может быть достигнуто при помощи изменения вероятностей $p(A_1), p(A_2), \ldots, p(A_m)$, и пусть это значение достигается при значениях $p^0(A_1), p^0(A_2), \ldots, p^0(A_m)$ этих вероятностей (ср. ниже конкретные примеры вычисления величины c и вероятностей $p^0(A_1), p^0(A_2), \ldots, p^0(A_m)$). Величина c определяет наибольшее количество информации, которое можно получить на приемном конце при приеме одного элементарного сигнала. Если же желать получать наибольшее количество информации в течение определенного промежутка времени (скажем, в течение единицы времени), то естественно все это время выбирать значения передаваемых элементарных сигналов с одними и теми же вероятностями $p^0(A_1), p^0(A_2), \ldots, p^0(A_m)$, не зависящими от того, какие именно сигналы были уже переданы раньше (см. по этому поводу мелкий шрифт на стр. 383—384, где будет строго доказано, что, выбирая последовательные значения передаваемых сигналов взаимно зависимыми, невозможно увеличить общее количество передаваемой ин-

формации). При такой передаче каждый принимаемый элементарный сигнал будет содержать c единиц информации, т. е. количество информации, переданное за единицу времени, будет равно

$$C = Lc = L \max I(\alpha, \beta).$$

Эта величина C и называется пропускной способностью линии связи с помехами. Так как наибольшее значение $I(\alpha, \beta)$ не может превосходить $H(\beta)$, а $H(\beta)$ всегда не больше, чем $\log m$ (см. стр. 73—74), то ясно, что пропускная способность линии связи с помехами всегда не больше, чем пропускная способность линии без помех, по которой за единицу времени может быть передано то же число элементарных сигналов и которая использует то же число разных сигналов. Следовательно, помехи могут только уменьшить пропускную способность линии связи, как это и должно быть по здравому смыслу.

Примеры.

1°. В случае, когда $r = m$, сигналы $B_1, \ldots, B_r$ совпадают с $A_1, \ldots, A_m$ и $p_{A_i}(A_j) = 1$ при $j = i$, а значит, $p_{A_i}(A_j) = 0$ при $j \neq i$, т. е. если всегда принимается тот же самый сигнал, который был передан (помехи не препятствуют передаче или даже вовсе отсутствуют):

$$H_\alpha(\beta) = 0 \text{ и } c = \max I(\alpha, \beta) = \max H(\beta) = \log m$$

(это наибольшее значение достигается, когда все значения передаваемого сигнала равновероятны, так что здесь $p^0(A_1) = p^0(A_2) = \ldots = p^0(A_m) = \frac{1}{m}$). Итак, в этом случае $C = L \log m$. Отсюда видно, что приведенное в § 2 определение пропускной способности линии связи без помех является частным случаем рассмотренного здесь более общего определения.

2°. Пусть по линии связи можно передавать два элементарных сигнала (скажем, посылку тока A_1 и паузу A_2) и те же два сигнала A_1 и A_2 принимаются на приемном конце. Пусть, далее, вероятность безошибочного приема любого из передаваемых сигналов равна $1 - p$, а вероятность ошибки равна p. В этом случае

$$p_{A_1}(A_1) = p_{A_2}(A_2) = 1 - p, \quad p_{A_1}(A_2) = p_{A_2}(A_1) = p,$$

так что указанная на стр. 331 таблица условных вероятностей здесь имеет вид

$$\begin{matrix} 1-p, & p; \\ p, & 1-p. \end{matrix}$$

Соответствующая линия связи называется двоичной симметричной линией; она схематически изображена на рис. 18, где линии со стрелками указывают, в какие принимаемые сигналы могут перейти передаваемые сигналы A_1 и A_2, а рядом с линиями выписаны вероятности соответствующих переходов.

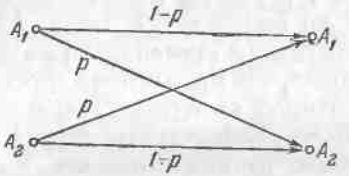


Рис. 18.

Для вычисления величины c мы воспользуемся равенством

$$I(\alpha, \beta) = H(\alpha) - H_\beta(\alpha).$$

Из приведенной выше таблицы условных вероятностей видно, что если передается сигнал A_1, то на приемном конце мы с вероятностью $1-p$ получим тот же сигнал A_1, а с вероятностью p — сигнал A_2; если же передается сигнал A_2, то мы с вероятностью p получим сигнал A_1, а с вероятностью $1-p$ — сигнал A_2. Поэтому

$$H_{A_1}(\alpha) = H_{A_2}(\alpha) = -(1-p)\log(1-p) - p\log p$$

и

$$H_\beta(\alpha) = p(A_1) H_{A_1}(\alpha) + p(A_2) H_{A_2}(\alpha) = -(1-p)\log(1-p) - p\log p$$

независимо от значений вероятностей $p(A_1)$ и $p(A_2)$ (ибо всегда $p(A_1) + p(A_2) = 1$). Следовательно, в рассматриваемом случае $H_\beta(\alpha)$ вовсе не зависит от вероятностей $p(A_1)$ и $p(A_2)$ и для вычисления

$$c = \max I(\alpha, \beta) = \max [H(\alpha) - H_\beta(\alpha)]$$

надо только определить максимальное значение $H(\alpha)$. Но величина $H(\alpha)$ — энтропия опыта α, могущего иметь всего два исхода,— никак не может превосходить 1 бит (см. стр. 74). С другой стороны, значение $H(\alpha) = 1$ наверное достигается при $p(A_1) = \frac{1}{2}$, $p(A_2) = \frac{1}{2}$, так

как в таком случае и оба исхода опыта α также будут иметь одинаковые вероятности (в общем случае эти вероятности, очевидно, равны $q(A_1) = p(A_1)\cdot(1-p) + p(A_2)\cdot p$ и $q(A_2) = p(A_1)\cdot p + p(A_2)\cdot(1-p)$). Отсюда вытекает, что в рассматриваемом случае

$$p^0(A_1) = p^0(A_2) = \frac{1}{2},$$

$$c = 1 + (1-p)\log(1-p) + p\log p$$

и

$$C = L[1 + (1-p)\log(1-p) + p\log p].$$

Мы получили явную формулу, показывающую, как зависит пропускная способность двоичной симметричной линии связи от вероятности p ошибки при передаче. График функции $C(p)$ изображен на рис. 19. Наибольшее значение (равное L) эта функция принимает при $p = 0$ (т. е. при отсутствии помех) и при $p = 1$ (т. е. в случае помех, переводящих каждый передаваемый сигнал A_1 в A_2, а каждый сигнал A_2 — в A_1; ясно, что такие помехи нисколько не мешают понять, какой именно сигнал был передан). Вообще при $p > \frac{1}{2}$ мы всегда можем в принятом сообщении заменить каждый принятый сигнал A_1 на A_2, а каждый принятый сигнал A_2 — на A_1; при этом мы придем к линии связи с вероятностью ошибки $1 - p < \frac{1}{2}$. Отсюда ясно, что при замене p на $1-p$ значение пропускной способности C не может измениться (это видно и из полученной выше формулы), т. е. график функции C должен быть симметричен относительно прямой $p = \frac{1}{2}$. При $p = \frac{1}{2}$ пропускная способность C равна нулю; это связано с тем, что при $p = \frac{1}{2}$, независимо от того, какой сигнал был передан, мы

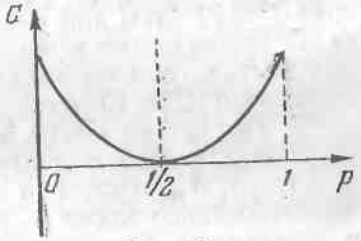


Рис. 19.

получим на приемном конце с вероятностью $\frac{1}{2}$ сигнал A_1 и с вероятностью $\frac{1}{2}$ — сигнал A_2, так что принятый сигнал не будет содержать никакой информации о том, какой сигнал был передан [1]). При значениях p промежуточных между 0 и $\frac{1}{2}$ (или между $\frac{1}{2}$ и 1), мы будем иметь положительную пропускную способность, меньшую чем L, причем при возрастании p (в случае $p < \frac{1}{2}$) или возрастании $1 - p$ (при $p > \frac{1}{2}$) эта пропускная способность быстро убывает. Так, например, если $L = 100$, то при $p = 0{,}01$ (т. е. в случае, когда из 100 передаваемых двоичных сигналов в среднем один сигнал принимается с ошибкой) $C \approx 92$ бита; при $p = 0{,}1$ (т. е. если 10 из 100 сигналов претерпевают искажение) $C \approx 53$ бита, а при $p = 0{,}25$ (т. е. если четверть всех сигналов принимается неправильно) $C \approx 19$ бит.

3°. Рассмотрим теперь более общий пример линии связи, использующей m различных элементарных сигналов $A_1, A_2, \ldots, A_m$, причем те же сигналы принимаются и на приемном конце линии (т. е. $r = m$, $B_i = A_i$ при всех i) и вероятность безошибочной передачи каждого из этих сигналов равна $1 - p$, а в случае ошибки переданный сигнал может с одинаковой вероятностью $\left(\text{равной } \frac{p}{m-1}\right)$ быть воспринятым как любой из $m - 1$ отличных от него сигналов. Таблица условных вероятностей здесь имеет вид

$$\begin{array}{cccc} 1-p, & \frac{p}{m-1}, & \frac{p}{m-1}, \ldots, & \frac{p}{m-1}; \\ \frac{p}{m-1}, & 1-p, & \frac{p}{m-1}, \ldots, & \frac{p}{m-1}; \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots & & & \\ \frac{p}{m-1}, & \frac{p}{m-1}, & \frac{p}{m-1}, \ldots, & 1-p, \end{array}$$

[1]) Вместо использования линии связи здесь можно с тем же успехом бросать у приемного конца монету и считать, что в случае выпадения «герба» принимается сигнал A_1, а в случае выпадения «цифры» — сигнал A_2.

а соответствующая линия связи называется m-ичной симметричной линией. Воспользуемся опять представлением $I(\alpha, \beta)$ в виде $H(\alpha) - H_\beta(\alpha)$; при этом, очевидно,

$$H_{A_1}(\alpha) = H_{A_2}(\alpha) = \ldots = H_{A_m}(\alpha) =$$
$$= -(1-p)\log(1-p) - (m-1)\cdot\frac{p}{m-1}\log\frac{p}{m-1}$$

и, следовательно,

$$H_\beta(\alpha) = -(1-p)\log(1-p) - p\log\frac{p}{m-1}.$$

Итак, как и в случае примера 2°, мы опять получаем, что $H_\beta(\alpha)$ не зависит от вероятностей $p(A_1)$, $p(A_2)$, ..., $p(A_m)$ и для нахождения пропускной способности надо лишь определить наибольшее значение $H(\alpha)$. Это максимальное значение находится вполне аналогично случаю примера 2°: оно равно $\log m$ и достигается, когда все исходы опыта α (т. е. все возможные значения сигнала, поступающего на приемный конец) будут равновероятны (для чего надо только, чтобы и вероятности $p(A_1)$, $p(A_2)$, ..., $p(A_m)$ посылки сигналов $A_1, A_2, \ldots, A_m$ были все одинаковыми). Поэтому здесь

$$p^0(A_1) = p^0(A_2) = \ldots = p^0(A_m) = \frac{1}{m},$$

$$c = \max I(\alpha, \beta) = \log m + p\log\frac{p}{m-1} + (1-p)\log(1-p)$$

и

$$C = L\left[\log m + p\log\frac{p}{m-1} + (1-p)\log(1-p)\right].$$

График функции $C(p)$ (для случая $m = 4$) изображен на рис. 20 (стр. 340). Эта функция достигает максимального значения (равного $L\log m$) при $p = 0$ (при отсутствии помех), а при возрастании p от нуля до значения $p = \frac{m-1}{m}$ она плавно уменьшается до нуля. То, при что $p = \frac{m-1}{m}$ пропускная способность оказывается равной пулю, совершенно естественно: в этом случае при любом значении посылаемого сигнала на приемном конце мы можем получить каждый из сигналов $A_1, A_2, \ldots, A_m$ с одинаковой

вероятностью $\frac{1}{m}$, так что никакой передачи информации о посылаемом сигнале здесь не происходит. При дальнейшем увеличении p мы снова получаем (правда, небольшую) положительную пропускную способность: в этом случае, приняв сигнал A_1, мы сможем отсюда сделать вывод, что

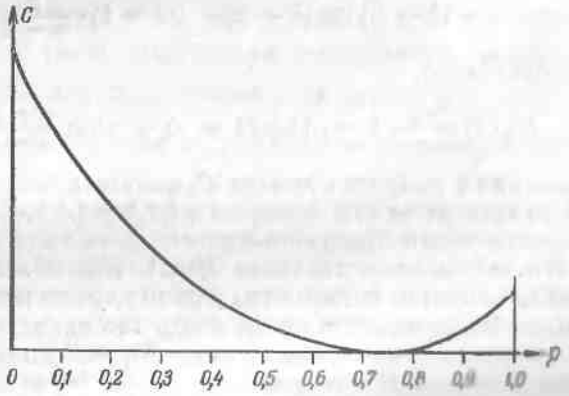


Рис. 20.

более вероятной является передача какого-либо о т л и ч н о г о от A_1 сигнала, т. е. будем иметь все же некоторую информацию о том, какой именно сигнал был передан. При этом при возрастании p от $\frac{m-1}{m}$ до единицы пропускная способность снова увеличивается; при $p = 1$ она равна $L \log \frac{m}{m-1}$.

4°. Рассмотрим теперь снова двоичную линию связи, по которой могут передаваться два сигнала A_1 и A_2, но теперь уже допустим, что на приемном конце полученный сигнал иногда может быть расшифрован как один из тех же двух сигналов, но иногда он так искажается, что его становится совершенно невозможно узнать. В этом последнем случае приходится считать, что принят какой-то совсем новый сигнал A_3, появление которого можно интерпретировать как событие: передаваемой сигнал стерся и не допускает расшифровки (поэтому такая линия связи называется д в о и ч н о й л и н и е й с о с т и р а-

ниям). Мы здесь ограничимся рассмотрением простейшей двоичной симметричной линии связи со стиранием, для которой вероятность «стирания» обоих передаваемых сигналов A_1 и A_2 равна одному и тому же числу q (т. е. $p_{A_1}(A_3) = p_{A_2}(A_3) = q$), и, кроме того, если стирания не произошло, то оба сигнала A_1 и A_2 с одной и той же вероятностью $1 - p - q$ будут правильно расшифрованы на приемном конце, а с вероятностью p они будут перепутаны (т. е. сигнал A_1 будет принят за A_2, соответственно, сигнал A_2 будет принят за A_1). Таким образом, в случае двоичной симметричной линии со стиранием $m = 2$, $r = 3$ и таблица условных вероятностей $p_{A_i}(B_j) = p_{A_i}(A_j)$ имеет вид

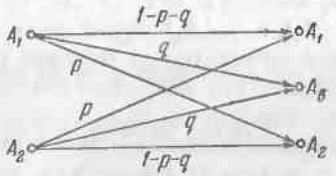


Рис. 21.

$$1 - p - q, \quad p, \quad q;$$

$$p, \quad 1 - p - q, \quad q$$

(см. рис. 21).

Ясно, что какой бы сигнал ни передавался, на приемном конце мы с вероятностью q получим сигнал A_3, в то время как из двух остальных сигналов один будет иметь вероятность $1 - p - q$, а второй — вероятность p. Следовательно, здесь

$$H_{A_1}(\alpha) = H_{A_2}(\alpha) = -(1 - p - q)\log(1 - p - q) - p\log p - q\log q.$$

а значит и

$$H_\beta(\alpha) = -(1 - p - q)\log(1 - p - q) - p\log p - q\log q,$$

так что

$$I(\alpha, \beta) = H(\alpha) + (1 - p - q)\log(1 - p - q) + p\log p + q\log q.$$

Так как опыт α в рассматриваемом случае может иметь три исхода A_1, A_2 и A_3, то $H(\alpha) \leqslant \log 3$; поэтому

$$c = \max I(\alpha, \beta) \leqslant \log 3 + (1 - p - q)\log(1 - p - q) + p\log p + q\log q.$$

Но может ли энтропия опыта α равняться $\log 3$? Легко видеть, что вообще говоря, н е м о ж е т ни при каких вероятностях $p(A_1)$ и $p(A_2)$ передачи сигналов A_1 и A_2. В самом деле, равенство $H(\alpha) = \log 3$ будет выполняться лишь тогда, когда все три исхода опыта α равновероятны (т. е. все имеют вероятность $1/3$); в нашем же случае вероятность исхода A_3 («стирания») при любом выборе $p(A_1)$ и $p(A_2)$ будет равна числу q, которое характеризует линию связи и вовсе не обязано равняться $1/3$. Следовательно, энтропия опыта α имеет вид

$$H(\alpha) = -q_1 \log q_1 - q_2 \log q_2 - q \log q,$$

где q фиксировано, а $q_1 = p(A_1)(1-p-q) + p(A_2)p$ и $q_2 = p(A_1)p + p(A_2)(1-p-q)$ — это вероятности появления на приемном конце линии сигналов A_1, соответственно A_2, которые зависят от значений $p(A_1)$ и $p(A_2)$. Ясно, что $q_1 + q_2 = 1 - q$ при всех значениях $p(A_1)$ и $p(A_2)$. Но легко видеть, что максимум выражения $-q_1 \log q_1 - q_2 \log q_2$, где $q_1 + q_2 = 1 - q$ (здесь q — фиксированное число; разумеется, $0 < q < 1$), достигается при $q_1 = q_2 = \frac{1-q}{2}$ [1]). Кроме того легко видеть, что значения $q_1 = q_2 = \frac{1-q}{2}$ являются возможными: для этого надо только принять, что $p(A_1) = p(A_2) = \frac{1}{2}$. Таким образом, в рассматриваемом случае

$$p^0(A_1) = p^0(A_2) = \frac{1}{2},$$

---

[1]) В самом деле, добавив к $-q_1 \log q_1 - q_2 \log q_2$ постоянное слагаемое $(1-q)\log(1-q) = (q_1+q_2)\log(1-q)$ и затем умножив полученную сумму на постоянный множитель $\frac{1}{1-q}$, мы получим выражение $-\frac{q_1}{1-q}\log\frac{q_1}{1-q} - \frac{q_2}{1-q}\log\frac{q_2}{1-q}$, представляющее собой энтропию опыта с двумя исходами, имеющими вероятности $\frac{q_1}{1-q}$ и $\frac{q_2}{1-q}$. Эта энтропия, очевидно, принимает наибольшее значение при $q_1 = q_2$; следовательно, наибольшее значение исходного выражения $-q_1 \log q_1 - q_2 \log q_2$ также достигается при $q_1 = q_2$.

$$c = \max I(\alpha, \beta) = -(1-q)\log\frac{1-q}{2} + \\ + (1-p-q)\log(1-p-q) + p\log p = \\ = (1-q)(1-\log(1-q)) + \\ + (1-p-q)\log(1-p-q) + p\log p$$

и, значит,

$$C = L\{(1-q)[1-\log(1-q)] + \\ + (1-p-q)\log(1-p-q) + p\log p\}.$$

Полученная пропускная способность C зависит от двух чисел p и q, характеризующих вероятности ошибок различного типа в нашей линии связи. Нетрудно показать, что C будет уменьшаться и при возрастании q, и при возрастании p (при естественном предположении, что $p < \frac{1}{2}$). Заметим еще, что в реальных двоичных линиях связи со стиранием обычно справедливо неравенство $p < q$, т. е. вероятность такого искажения передаваемого сигнала, при котором его невозможно узнать, обычно превосходит вероятность искажения, при котором он оказывается по форме похожим на второй из используемых сигналов. В ряде случаев вероятность p вообще оказывается так мала, что ею можно пренебречь, т. е. можно считать, что единственно возможные вредные искажения сигнала помехами это те, при которых сигнал на выходе невозможно расшифровать (т. е. в ходе передачи он «стерся»). Если допустимо считать, что $p = 0$, то формула для пропускной способности C приобретает особенно простой вид:

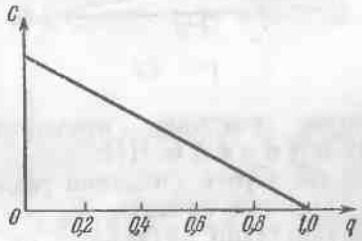


Рис. 22.

$$C = L(1-q)$$

(см. рис. 22). Последний результат представляется совершенно естественным: при $p = 0$ из L двоичных сигналов, передаваемых по нашей линии связи за единицу времени, в среднем Lq сигналов будет «стираться», т. е. не будет переносить никакой информации, в то время как остальные

$L(1-q)$ сигналов будут точно расшифровываться на приемном конце, так что каждый из них будет содержать ровно 1 бит информации.

То обстоятельство, что во всех предыдущих примерах пропускная способность C достигалась при одинаковых вероятностях передачи любого из используемых элементарных сигналов, разумеется, имеет случайный характер: оно объясняется просто тем, что для простоты расчетов во всех этих примерах таблица условных вероятностей $p_{A_i}(A_j)$, характеризующих линию связи, выбиралась очень симметричной. Для иллюстрации того, что дело может обстоять и иначе, приведем еще результаты, относящиеся к следующему несколько более сложному примеру, впервые рассмотренному Шенноном [1]:

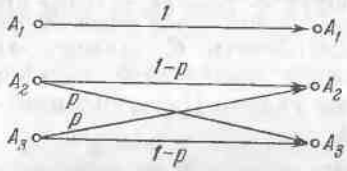


Рис. 23.

5°. Пусть по линии связи можно передавать три элементарных сигнала A_1, A_2 и A_3, причем первый сигнал значительно отличается от двух других и всегда может быть безошибочно угадан на приемном конце линии, а каждый из двух других сигналов с вероятностью $1-p$ принимается правильно, а с вероятностью p воспринимается как второй из них. Иначе говоря, мы считаем, что $m = r = 3$ и что таблица условных вероятностей $p_{A_i}(A_j)$ имеет вид

$$\begin{matrix} 1, & 0, & 0; \\ 0, & 1-p, & p; \\ 0, & p, & 1-p \end{matrix}$$

(см. рис. 23). Следовательно, здесь

$$H_{A_1}(\alpha) = 0, \quad H_{A_2}(\alpha) = A_{A_2}(\alpha) = -(1-p)\log(1-p) - p\log p$$

и

$$H_\beta(\alpha) = [p(A_2) + p(A_3)]\,[-(1-p)\log(1-p) - p\log p],$$

$$I(\alpha, \beta) = -q(A_1)\log q(A_1) - q(A_2)\log q(A_2) - q(A_3)\log q(A_3) + [p(A_2) + p(A_3)]\,[(1-p)\log(1-p) + p\log p],$$

где $q(A_1) = p(A_1)$, $q(A_2) = p(A_2)(1-p) + p(A_3)p$ и $q(A_3) = p(A_2)p + p(A_3)(1-p)$ — вероятности исходов A_1, A_2, A_3 опыта α.

Отметим, что $H_\beta(\alpha)$ зависит не от всех трех вероятностей $p(A_1)$, $p(A_2)$, $p(A_3)$, а лишь от $p(A_2) + p(A_3) = 1 - p(A_1)$. Воспользовавшись рассуждением, приведенным в сноске на стр. 342, легко показать, что при фиксированном $p(A_1) = q(A_1)$ энтропия $H(\alpha)$ (а значит, и информация $I(\alpha, \beta)$) будет наибольшей, если вероятности $q(A_2)$ и $q(A_3)$ (а следовательно, и $p(A_2)$ и $p(A_3)$) будут равны между собой:

$$p(A_2) = p(A_3) = q(A_2) = q(A_3) = \frac{1 - p(A_1)}{2}.$$

После этого остается только определить, при каком значении $p(A_1)$ выражение

$$I(\alpha, \beta) = -p(A_1)\log p(A_1) - [1 - p(A_1)]\left[\log \frac{1 - p(A_1)}{2} - (1-p)\log(1-p) - p\log p\right],$$

где p — заданное неотрицательное число, не превосходящее единицы, будет наибольшим. Последняя задача является довольно сложной, если пользоваться лишь методами элементарной математики, но легко решается с помощью дифференциального исчисления [1]). Оказывается, что искомое значение $p(A_1)$ равно

$$p^0(A_1) = \frac{1}{1 + 2p^p(1-p)^{1-p}}.$$

Итак, в рассматриваемом случае

$$p^0(A_1) = \frac{1}{1 + 2p^p(1-p)^{1-p}},$$

$$p^0(A_2) = p^0(A_3) = \frac{p^p(1-p)^{1-p}}{1 + 2p^p(1-p)^{1-p}};$$

---

[1]) Известно, что точка x отрезка $0 \leqslant x \leqslant 1$, в которой функция $y = -x\log x - (1-x)\left[\log \frac{1-x}{2} - \log a\right]$ (где $a = p^p(1-p)^{1-p}$ и все логарифмы — двоичные) принимает наибольшее значение, совпадает с той, в которой обращается в нуль производная этой функции.

подставляя эти значения вероятностей в выражение для $I(\alpha, \beta)$ и умножая результат на число L сигналов, передаваемых за единицу времени, легко находим пропускную способность нашей линии связи:

$$C = L \log [1 + 2\, p^p\, (1 - p)^{1-p}].$$

График функции $C = C(p)$ приведен на рис. 24. При $p = 0$ эта функция принимает наибольшее значение: при $p \to 0$, как нетрудно показать, $p^p\, (1 - p)^{1-p} \to 1$, и, следовательно, здесь $p^0(A_1) = p^0(A_2) = p^0(A_3) = \frac{1}{3}$ и $C = L \log 3$. Этот результат, разумеется, очевиден; при

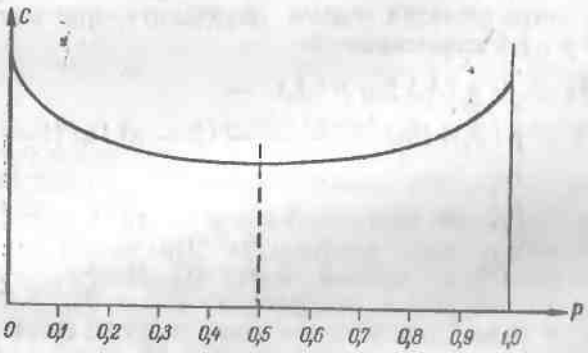


Рис. 24.

$p = 0$ мы будем иметь просто линию связи без помех, использующую три различных элементарных сигнала (см. пример 1°). При возрастании p от 0 до $\frac{1}{2}$ пропускная способность C уменьшается, так как при передаче второго или третьего сигнала мы здесь будем терять часть информации из-за наличия помех; в этой связи и вероятность $p^0(A_1)$ здесь оказывается несколько большей чем $\frac{1}{3}$ (т. е. первый сигнал здесь выгодно передавать несколько чаще, чем второй или третий). При $p = \frac{1}{2}$ пропускная способность принимает наименьшее значение, равное $C = L$

$\left(\text{так как } \left(\frac{1}{2}\right)^{\frac{1}{2}} \cdot \left(\frac{1}{2}\right)^{\frac{1}{2}} = \frac{1}{2}\right)$. Для достижения этой пропускной способности первый сигнал следует передавать в половине всех случаев ($p^0(A_1) = \frac{1}{2}$), а второй и третий — во второй половине случаев (фактически сигналы A_2 и A_3 здесь следует рассматривать как один общий сигнал, так как на приемном конце все равно их никак нельзя различить, а можно лишь утверждать, что передан был какой-то из них, а не сигнал A_1; поэтому случай $p = \frac{1}{2}$ равносилен случаю линии без помех, использующей д в а различных сигнала). При дальнейшем увеличении p от $\frac{1}{2}$ до 1 значение $C(p)$ снова возрастает, причем $C(p) = C(1-p)$ (по тем же причинам, что и в случае примера 2°).

Другой пример линии связи, для которой вероятности $p^0(A_i)$ не равны между собой, можно получить, предполагая, что $m = r = 2$, но что вероятности ошибки при передаче двух используемых сигналов не одинаковы (случай д в о и ч н о й н е с и м м е т р и ч н о й линии). В этом случае, однако, все формулы оказываются заметно более сложными, чем в рассмотренных выше примерах; поэтому мы на нем не будем останавливаться.

Будем теперь считать, что пропускная способность C линии связи нам известна. В случае отсутствия помех, как мы видели в § 2, знание величины C позволяет весьма точно оценить возможную скорость передачи сообщений по данной линии: ни при каком методе кодирования эта скорость не может превзойти величины

$$v = \frac{C}{H} \text{ букв/ед. времени}$$

(где H — энтропия одной буквы передаваемого сообщения); однако скорость передачи, сколь угодно близкая к v, всегда может быть достигнута. При наличии помех, кроме скорости, следует учитывать еще и степень точности передачи, характеризуемую вероятностью ошибки в определении каждой отдельной переданной буквы. Легко

понять, что *при скорости передачи* v_1 *букв/ед. времени, превышающей величину* $v = C/H$ (где C — это определенная выше пропускная способность линии связи с помехами), *точная передача* (позволяющая безошибочно восстановить все буквы переданного сообщения) *никак не может иметь места* (это утверждение представляет собой не вполне четкую формулировку так называемой обратной теоремы о кодировании при наличии помех, о которой мы подробно будем говорить на стр. 362—363). В самом деле, при безошибочной передаче со скоростью v_1 количество информации о буквах сообщения, передаваемое по линии за единицу времени, будет равно полной степени неопределенности v_1-буквенного «блока», т. е. равно произведению v_1H (напомним, что отдельные буквы мы считаем независимыми); следовательно, переданное за единицу времени количество информации о посылаемых кодовых обозначениях (т. е. о сигналах, являющихся исходами опытов β) тем более не может быть меньше чем v_1H (ср. стр. 123). Но так как $v_1H > C$ при $v_1 > v = \frac{C}{H}$, то из самого определения величины C вытекает, что безошибочная передача сообщений со скоростью $v_1 > v$ букв за единицу времени не может быть осуществлена. Исходя из этих соображений, можно даже точно оценить наименьшую вероятность ошибки, обязательно имеющуюся даже при «наилучшей» передаче сообщения с данной скоростью $v_1 > v$ (см. ниже, стр. 361 и след.).

Заметим далее, что если не накладывать вообще никаких ограничений на скорость передачи сообщений, то в большинстве случаев без труда можно добиться того, чтобы вероятность ошибки в определении каждой переданной буквы оказалась сколь угодно малой; как правило, для этого достаточно просто очень много раз повторять каждый передаваемый сигнал (или каждую группу таких сигналов). Заранее, однако, можно было бы думать, что для того, чтобы добиться очень малой вероятности ошибки, необходимо очень сильно уменьшить скорость передачи (такое резкое уменьшение скорости, в частности, будет иметь место, если вероятность ошибки мы будем уменьшать при помощи многократного повторения сигналов). Точнее говоря, на первый взгляд кажется

естественным думать, что любое уменьшение вероятности ошибки в определении каждой передаваемой буквы должно быть неизбежно связано с уменьшением также и скорости передачи, причем неограниченного уменьшения вероятности ошибки никак нельзя добиться не уменьшая неограниченно скорость передачи. Оказалось, что в действительности дело обстоит совсем не так. А именно, Шеннон доказал, что *для любой линии связи с помехами всегда можно подобрать специальный код, позволяющий передавать сообщения по этой линии с заданной скоростью, сколь угодно близкой к*

$$v = \frac{C}{H} \text{ букв/ед. времени}$$

(но обязательно все же несколько меньшей, чем эта величина!) *так, чтобы вероятность ошибки в определении каждой переданной буквы оказалась меньше любого заранее заданного числа* ε (например, меньшей чем 0,001, или чем 0,0001 или чем 0,000001). Разумеется, код, о котором здесь идет речь, будет зависеть от ε и чем меньше ε, тем он будет, как правило, более сложным. Предложение, выделенное курсивом, обобщает сформулированную в § 2 основную теорему о кодировании; его можно назвать о с н о в н о й т е о р е м о й о к о д и р о в а н и и п р и н а л и ч и и п о м е х. Существенную роль в доказательстве этой теоремы играет использование кодирования сразу очень длинных «блоков» из большого числа букв; поэтому передача сообщений со скоростью, близкой к v, и с очень малой вероятностью ошибки обычно будет сопряжена со значительным запаздыванием при расшифровке каждой переданной буквы.

Прежде чем идти дальше, отметим, что здесь, так же как и в случае рассматривавшейся в § 2 этой главы основной теоремы о кодировании при отсутствии помех, предположение о том, что отдельные буквы текста взаимно н е з а в и с и м ы, на самом деле не является существенным. В дальнейшем мы им почти не будем пользоваться, а будем лишь использовать то частное обстоятельство, что при достаточно большом N из n^N различных N-буквенных блоков (где каждая буква может принимать n различных значений) только 2^{HN} являются «вероятными» (и имеющими примерно одинаковую вероятность). В случае,

когда последовательные буквы текста взаимно независимы, последнее обстоятельство будет уже неверным; однако, как было отмечено на стр. 228—229, в этом случае при широких условиях также среди всевозможных N-буквенных блоков, где N достаточно велико, можно будет выделить сравнительно небольшую долю примерно одинаково вероятных блоков, имеющих весьма близкую к единице суммарную вероятность. Общее число «вероятных» блоков из N взаимно зависимых букв согласно сказанному на стр. 229 имеет порядок $2^{H_\infty N} \approx 2^{H(N)}$, где $H^{(N)}$ — энтропия N-буквенного блока, а $H_\infty = \lim\limits_{N\to\infty} \dfrac{H^{(N)}}{N}$ — удельная энтропия, приходящаяся на одну букву текста. Таким образом, если буквы текста являются зависимыми, то, вообще говоря, мы должны лишь заменить во всем дальнейшем энтропию H одной буквы на меньшую чем H удельную энтропию H_∞. Точно так же и в случае передачи со скоростью v_1, превышающей $v = \dfrac{C}{H_\infty}$ букв/ед. времени, мы можем воспользоваться тем, что суммарное количество информации, содержащейся в $v_1 T$ буквах передаваемого текста (где T — время передачи), ни при каком T не может быть меньше, чем $v_1 T H_\infty$ бит. Отсюда сразу вытекает, что утверждение, выделенное курсивом на стр. 349, будет справедливо и в случае передачи сообщений, буквы которых взаимно зависимы, если только заменить скорость $v = \dfrac{C}{H}$ букв/ед. времени на $v = \dfrac{C}{H_\infty}$ букв/ед. времени.

Будем теперь для простоты снова предполагать, что отдельные буквы передаваемых сообщений взаимно н е з а в и с и м ы (т. е. будем всюду использовать обычную энтропию H одной буквы, а не удельную энтропию H_∞). К сожалению, строгое математическое доказательство основной теоремы Шеннона о кодировании при наличии помех и в этом случае является все же довольно сложным. В работе [1], положившей начало всей теории информации, такое строгое доказательство вообще отсутствует; Шеннон здесь ограничился лишь изложением ряда общих соображений, весьма наглядно объясняющих причины, по которым эта теорема должна иметь место. Позже Ф а й н с т е й н (см., например, [5]) дал строгое доказательство

основной теоремы Шеннона, идея которого кое в чем отклоняется от первоначальных соображений Шеннона. Полное доказательство этой теоремы, близкое к выводу, вкратце намеченному в [1], содержится в работе Шеннона [158], в которой одновременно показано, что на том же пути можно получить и заметно более сильные результаты, о которых мы еще скажем ниже. В настоящей книге мы начнем с разбора самых простых, идущих от Шеннона, рассуждений, заставляющих считать основную теорему о кодировании правдоподобной, а в дальнейшем на стр. 373 и след. обрисуем также и метод строгого ее доказательства, опирающийся на более глубокие соображения работы [1]; кроме того, учитывая очень большую важность рассматриваемой теоремы, мы дополнительно изложим мелким шрифтом в самом конце настоящего параграфа (на стр. 385—390) еще одно строгое ее доказательство для частного случая двоичной симметричной линии связи, идея которого родственна той, которую использовал Файнстейн.

Пусть β — опыт, состоящий в выборе (и последующей передаче по линии связи) одного из m элементарных сигналов $A_1, A_2, \ldots, A_m$ с вероятностями $p^0(A_1), p^0(A_2), \ldots, p^0(A_m)$, которым отвечает наибольшая информация $I(\alpha, \beta)$ (т. е. для которых реализуется пропускная способность нашей линии связи). Теорема Шеннона утверждает, что существует способ кодирования сообщений, позволяющий вести передачу со скоростью, сколь угодно близкой к

$$v = L\frac{c}{H} \text{ букв/ед. времени,}$$

где

$$c = H(\beta) - H_\alpha(\beta) = H(\alpha) - H_\beta(\alpha)$$

(но несколько меньшей этой величины!), так что вероятность ошибки при расшифровке принятого сообщения будет мала (меньше произвольного заданного заранее малого числа). Так как за единицу времени мы можем передать L элементарных сигналов, то для достижения такой скорости передачи требуется, чтобы кодовое обозначение «блока» из N букв «в среднем» содержало бы около $\frac{H}{c}N$ (но несколько более этого числа) элементарных сигналов; при этом LT элементарных сигналов, переданных за

большое время T, будут содержать примерно $\frac{LT}{\frac{H}{c}N} = \frac{vT}{N}$ кодовых обозначений, отвечающих сообщению из примерно vT букв.

Мы знаем (см. стр. 217—226), что на самом деле нам нет необходимости заботиться, чтобы длину, близкую к $\frac{H}{c}N$ сигналам, имели кодовые обозначения в с е х $n^N = 2^{\log n \cdot N}$ различных N-буквенных сообщений (где n — число букв алфавита). В самом деле, лишь 2^{HN} из этих сообщений являются «вероятными»; что же касается остальных $2^{\log n \cdot N} - 2^{HN}$ сообщений, то суммарная вероятность их появления при большом N будет очень мала, —и поэтому если даже их кодовые обозначения будут значительно длиннее, то это все равно не уменьшит заметно скорости передачи (остающейся близкой к $L\frac{c}{H}$ букв./ед. времени). Заметим также, что для достижения высокой точности передачи надо лишь позаботиться, чтобы вероятность ошибки при расшифровке принятого кодового обозначения каждого из 2^{HN} «вероятных» N-буквенных сообщений была мала, так как все остальные сообщения сами встречаются крайне редко и их учет мало что изменит.

Мы будем искать способ кодирования, при котором длина кодового обозначения N-буквенного блока будет равна $\frac{H}{c_1}N = N_1$ элементарным сигналам [1]); здесь c_1 — заранее выбранное число, которое д о л ж н о у д о в л е т в о р я т ь е д и н с т в е н н о м у у с л о в и ю:

$$c_1 < c$$

(но c_1 может быть сколь угодно близко к c!). Число всех различных цепочек из $\frac{H}{c_1}N$ элементарных сигналов равно $m^{\frac{H}{c_1}N} = 2^{\frac{\log m}{c_1}HN}$; так как $c_1 < c \leqslant H(\beta) \leqslant \log m$, то оно, наверное, больше, чем 2^{HN}, и поэтому каждому из

[1]) Как обычно, если число $\frac{H}{c_1}N = N_1$ не целое, то его надо заменить ближайшим к нему ц е л ы м числом. Это же замечание относится и ко всем другим встречающимся ниже числам, которые по своему смыслу обязательно должны быть целыми.

2^{HN} «вероятных» N-буквенных сообщений может быть сопоставлена в качестве кодового обозначения своя цепочка из $N_1 = \frac{H}{c_1} N$ элементарных сигналов. Однако нам надо еще добиться, чтобы вероятность ошибки при расшифровке всех переданных кодовых обозначений была мала. Ясно, что для этого наши 2^{HN} используемых кодовых обозначений должны сильно отличаться друг от друга — только при таком условии можно надеяться, что, несмотря на возможные искажения сигналов в процессе их передачи по линии связи, мы все равно сможем достаточно надежно отличить друг от друга кодовые обозначения на приемном конце линии.

Для того чтобы оценить возможное число таких с и л ь н о о т л и ч а ю щ и х с я д р у г о т д р у г а N_1-членных кодовых обозначений, можно рассуждать таким образом. Каждая цепочка из $N_1 = \frac{H}{c_1} N$ передаваемых элементарных сигналов A_i (где $i = 1, 2, \ldots,$ или m) на приемном конце линии будет восприниматься как цепочка из некоторых N_1 элементарных сигналов B_j (где $j = 1, 2, \ldots, r$; ср. выше, стр. 332). Разумеется, передавая много раз о д н у и т у ж е цепочку $A_{i_1}A_{i_2} \ldots A_{i_{N_1}}$, мы будем получать на приемном конце много р а з н ы х цепочек $B_{j_1}B_{j_2} \ldots B_{j_{N_1}}$ — в этом как раз и проявляется случайный характер помех, воздействующих на передачу. Однако, передавая одну цепочку $A_{i_1}A_{i_2} \ldots A_{i_{N_1}}$, мы будем получать на приемном конце линии связи разные цепочки $B_{j_1}B_{j_2} \ldots B_{j_{N_1}}$ с разной частотой: одни из таких цепочек будут появляться при этом сравнительно часто, другие же — крайне редко [1]). Следующие соображения позволяют оценить примерное число цепочек $B_{j_1}B_{j_2} \ldots B_{j_{N_1}}$, которые с не слишком малой вероятностью могут возникнуть при передаче данной цепочки $A_{i_1}A_{i_2} \ldots A_{i_{N_1}}$.

[1]) Например, в случае двоичной симметричной линии связи, рассмотренной на стр. 335—336, при передаче N_1 сигналов на приемном конце линии, очевидно, с очень большой вероятностью появится одна из цепочек, отличающихся от переданной цепочки $A_{i_1}A_{i_2} \ldots A_{i_{N_1}}$ не меньше, чем на $N_1\,(p - \delta)$, и не больше, чем на $N_1\,(p + \delta)$, сигналов, где δ — некоторое малое число (см. обсуждение закона больших чисел в § 4 гл. I).

Предположим, что мы последовательно передаем по нашей линии связи элементарные сигналы A_i, каждый раз выбирая передаваемый сигнал случайно (и независимо от всех ранее переданных сигналов) с вероятностями $p^0(A_1)$, $p^0(A_2)$, ..., $p^0(A_m)$. В таком случае, согласно сказанному на стр. 225, при большом N_1 среди всех N_1-членных цепочек вида $A_{i_1}A_{i_2}\dots A_{i_{N_1}}$ только $2^{H(\beta)N_1}$ цепочек будут «вероятными» (причем они все будут иметь примерно одинаковую вероятность); суммарная же вероятность того, что переданной окажется одна из остальных $m^{N_1} - 2^{H(\beta)N_1} = 2^{\log m \cdot N_1} - 2^{H(\beta)N_1}$ цепочек, будет крайне мала. Условимся выбирать все нужные нам N_1-членные кодовые обозначения из числа $2^{H(\beta)N_1}$ «вероятных» N_1-членных цепочек, а остальные такие цепочки вовсе не будем рассматривать; это возможно, поскольку

$$H(\beta)N_1 = \frac{H(\beta)}{c_1}HN > HN$$

(ибо $c_1 < c \leqslant H(\beta)$), и, следовательно, общее число «вероятных» цепочек также превосходит требуемое число 2^{HN} кодовых обозначений.

Рассмотрим теперь всевозможные цепочки вида $A_{i_1}A_{i_2}\dots$ $\dots A_{i_{N_1}}B_{j_1}B_{j_2}\dots B_{j_{N_1}}$, состоящие из N_1 переданных элементарных сигналов A_i и тех N_1 сигналов B_j, в которые преобразовались эти сигналы A_i в процессе их передачи по линии связи. Общее число таких $2N_1$-членных цепочек, очевидно, равно

$$m^{N_1}r^{N_1} = 2^{(\log m + \log r)N_1}.$$

К ним также можно применить приведенные на стр. 225 соображения, из которых вытекает, что если все передаваемые сигналы A_i выбираются так, как это объяснено выше, то только $2^{H(\alpha\beta)N_1}$ из общего числа $2^{(\log m+\log r)N_1}$ наших цепочек будут «вероятными» (и будут иметь примерно одинаковую вероятность); суммарная же вероятность всех остальных $2^{(\log m+\log r)N_1} - 2^{H(\alpha\beta)N_1}$ цепочек будет крайне мала [1]). Следовательно, число «вероятных»

---

[1]) В самом деле, $2N_1$-членная цепочка $A_{i_1}A_{i_2}\dots A_{i_{N_1}}B_{j_1}B_{j_2}\dots B_{j_{N_1}}$ может рассматриваться как цепочка $(A_{i_1}B_{j_1})(A_{i_2}B_{j_2})\dots(A_{i_{N_1}}B_{j_{N_1}})$, состоящая из N_1 последовательных исходов составного опыта $\alpha\beta$ (с mr возможными исходами), имеющего энтропию $H(\alpha\beta)$.

$2N_1$-членных цепочек $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}B_{j_1}B_{j_2}\ldots B_{j_{N_1}}$ превосходит число «вероятных» передаваемых N_1-членных цепочек $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}$ в

$$2^{H(\alpha\beta)N_1}:2^{H(\beta)N_1}=2^{[H(\alpha\beta)-H(\beta)]N_1}=2^{H_\beta(\alpha)N_1}$$

раз. Отсюда можно заключить, что каждой «вероятной» передаваемой цепочке $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}$ отвечает целая группа из $2^{H_\beta(\alpha)N_1}$ цепочек $B_{j_1}B_{j_2}\ldots B_{j_{N_1}}$ принимаемых сигналов, в одну из которых цепочка $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}$ перейдет с очень большой (т. е. очень близкой к единице)

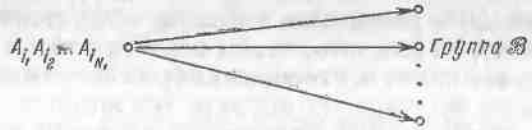


Рис. 25.

вероятностью. Эту группу из $2^{H_\beta(\alpha)N_1}$ цепочек $B_{j_1}B_{j_2}\ldots B_{j_{N_1}}$, отвечающих цепочке $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}$, мы для краткости далее будем называть о т в е ч а ю щ е й $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}$ г р у п п о й $\mathcal{B}$ (см. схематический рис. 25). Комбинируя каждую из $2^{H(\beta)N_1}$ «вероятных» передаваемых цепочек $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}$ с $2^{H_\beta(\alpha)N_1}$ цепочками отвечающей ей группы $\mathcal{B}$, мы как раз и получим все $2^{H(\alpha\beta)N_1}$ «вероятных» цепочек $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}B_{j_1}B_{j_2}\ldots B_{j_{N_1}}$.

Две N_1-членные цепочки передаваемых сигналов $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}$ и $A_{i_1'}A_{i_2'}\ldots A_{i_{N_1}'}$ следует считать «сильно отличающимися друг от друга», если соответствующие им две группы $\mathcal{B}$ н е п е р е с е к а ю т с я между собой. В самом деле, сообщение $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}$ при передаче по нашей линии связи «почти наверное» (т. е. с очень близкой к единице вероятностью) перейдет в одну из цепочек $B_{j_1}B_{j_2}\ldots B_{j_{N_1}}$, принадлежащих первой группе $\mathcal{B}$, в то время как сообщение $A_{i_1'}A_{i_2'}\ldots A_{i_{N_1}'}$ «почти наверное» перейдет в одну из цепочек, принадлежащую второй группе $\mathcal{B}$.

Поэтому если указанные две группы $\mathcal{B}$ не пересекаются между собой и известно, что передано было или сообщение $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}$, или сообщение $A'_{i_1}A'_{i_2}\ldots A'_{i_{N_1}}$, то можно, например, во всех случаях, когда на приемном конце линии связи принимается одна из цепочек первой группы $\mathcal{B}$, считать, что передавалось сообщение $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}$, а когда принимается какая-то из прочих цепочек (включающих также и все цепочки второй группы $\mathcal{B}$) — считать, что было передано $A'_{i_1}A'_{i_2}\ldots A'_{i_{N_1}}$. Ясно, что при этом вероятность ошибки при расшифровке принятого сообщения будет очень малой. Аналогично этому, если требуется подобрать 2^{HN} различных кодовых обозначений из N_1 сигналов A_i, то для того, чтобы вероятность ошибки при расшифровке принятых сообщений была очень мала, достаточно иметь возможность выбрать эти кодовые обозначения так, чтобы *все* 2^{HN} *отвечающих им групп* $\mathcal{B}$ *не пересекались между собой*. Так как каждая группа $\mathcal{B}$ содержит $2^{H_{\mathcal{B}(\alpha)}N_1} = 2^{\frac{H_{\mathcal{B}(\alpha)}}{c_1}HN}$ цепочек $B_{j_1}B_{j_2}\ldots B_{j_{N_1}}$, то в 2^{HN} групп $\mathcal{B}$ будет входить

$$2^{\frac{H_{\mathcal{B}(\alpha)}}{c_1}HN}\cdot 2^{HN} = 2^{\left(\frac{H_{\mathcal{B}(\alpha)}}{c_1}+1\right)HN}$$

цепочек. Поскольку при этом все такие цепочки $B_{j_1}B_{j_2}\ldots B_{j_{N_1}}$ входят в «вероятные» $2N_1$-членные последовательности $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}B_{j_1}B_{j_2}\ldots B_{j_{N_1}}$, то и сами они естественно будут «вероятными», т. е. будут относиться к числу тех, которые не слишком редко возникают при последовательной передаче сигналов $A_1, A_2, \ldots, A_m$ с вероятностями $p^0(A_1), p^0(A_2), \ldots, p^0(A_m)$ (независимо от того, какие сигналы передавались раньше). Число таких «вероятных» цепочек $B_{j_1}B_{j_2}\ldots B_{j_{N_1}}$ (т. е. «вероятных» цепочек из N_1 последовательных исходов опыта α), как известно, равно $2^{H(\alpha)N_1} = 2^{\frac{H(\alpha)}{c_1}HN}$.

Составим теперь отношение общего числа $2^{\frac{H(\alpha)}{c_1}HN}$ «вероятных» цепочек $B_{j_1}B_{j_2}\ldots B_{j_{N_1}}$ к суммарному числу

$2^{\left(\frac{H_\beta(\alpha)}{c_1}+1\right)HN}$ таких цепочек, входящих в 2^{HN} групп $\mathcal{B}$:

$$\frac{2^{\frac{H(\alpha)}{c_1}HN}}{2^{\left(\frac{H_\beta(\alpha)}{c_1}+1\right)HN}} = 2^{\left(\frac{H(\alpha)}{c_1}-\frac{H_\beta(\alpha)}{c_1}-1\right)HN} = 2^{\left(\frac{H(\alpha)-H_\beta(\alpha)}{c_1}-1\right)HN} =$$

$$= 2^{\left(\frac{c}{c_1}-1\right)HN}.$$

Мы видим, что если бы c_1 было б о л ь ш е, чем c, то это отношение было бы меньше единицы, т. е. полное число цепочек в наших 2^{HN} группах $\mathcal{B}$ было бы больше, чем общее число всех «вероятных» цепочек $B_{j_1}B_{j_2}\dots B_{j_{N_1}}$; поэтому ясно, что при $c_1 > c$ кодовые обозначения никак н е л ь з я подобрать так, чтобы отвечающие им группы $\mathcal{B}$ не пересекались. Разумеется, этого и следовало ожидать, так как мы уже знаем, что со скоростью $L\frac{c_1}{H}$ букв/ед. времени, где $c_1 > c$, по нашей линии связи нельзя передавать сообщения так, чтобы вероятность ошибки при их расшифровке на приемном конце линии была бы сколь угодно мала. Но если c_1 м е н ь ш е c, то выписанное нами отношение оказывается бо́льшим единицы (так как в этом случае $\frac{c}{c_1}-1>0$); более того, при очень большом N оно оказывается равным числу 2, возведенному в очень большую степень, т. е. о ч е н ь б о л ь ш и м. Таким образом, при большом N полное число цепочек в 2^{HN} группах $\mathcal{B}$ будет составлять н и ч т о ж н у ю ч а с т ь всего числа «вероятных» цепочек из N_1 сигналов B_j; это обстоятельство делает очень правдоподобным предположение о том, что 2^{HN} *кодовых обозначений длины* $\frac{H}{c_1}N$ *можно выбрать так, чтобы отвечающие им группы* $\mathcal{B}$ *не пересекались между собой*. А такой выбор кодовых обозначений, как мы знаем, при достаточно большом N как раз и обеспечивает возможность расшифровки полученных сообщений со сколь угодно малой вероятностью ошибки.

Приведенные выше соображения делают теорему Шеннона весьма правдоподобной, но, разумеется, их нельзя рассматривать как ее строгое доказательство (это обстоятельство еще будет дополнительно разъяснено на стр. 373—374).

Тем не менее пока мы ограничимся сказанным и перейдем к разбору некоторых других связанных с теоремой Шеннона вопросов; в дальнейшем, однако, мы приведем на стр. 374—383 красивое (но не очень простое) рассуждение (идея которого также принадлежит Шеннону [1]), убедительно показывающее, что действительно должен существовать такой выбор 2^{HN} кодовых обозначений, который обеспечивает если не полное отсутствие пересечений соответствующих 2^{HN} груш $\mathcal{B}$, то, по крайней мере, достаточную малость этих пересечений, не препятствующую тому, чтобы вероятность ошибки при расшифровке могла быть сделана сколь угодно близкой к нулю. Еще более подробно мы разберем в самом конце настоящего параграфа (на стр. 384—390) другое строгое доказательство основной теоремы о кодировании, относящееся, впрочем, лишь к специальному случаю двоичной симметричной линии связи. Мы предоставляем читателю самому решить, стоит ли ему тратить время на разбор всего этого материала (и когда — сейчас же или позже, следуя принятому в книге порядку изложения) или же он предпочитает ограничиться лишь приведенными выше нестрогими соображениями; в этом последнем случае весь конец настоящего параграфа (начиная с начала стр. 373 и до стр. 390) читатель может просто опустить. Предупредим только заранее, что оба излагаемых в конце параграфа доказательства теоремы Шеннона (так же как и все другие известные ее доказательства) являются неэффективными: из них следует, что при достаточно большом N обязательно существует такой способ выбора кодовых обозначений, который гарантирует, что вероятность ошибки при расшифровке каждой буквы полученного сообщения не будет превосходить заданного (произвольно малого) числа ε, но они ничего не говорят о том, как можно найти такой способ выбора кодовых обозначений (ср., впрочем, начало следующего параграфа книги, где это обстоятельство будет разъяснено более точно). Вопросу о том, как на самом деле следует выбирать кодовые обозначения для того, чтобы сделать вероятность ошибки при расшифровке достаточно малой, будет посвящен последний параграф нашей книги.

Теорема Шеннона не только не позволяет указать, как именно надо выбирать кодовые обозначения для того, чтобы сообщения можно было передавать по используемой линии связи с заданной скоростью

$$v_1 < v = L\frac{c}{H} \text{ букв/ед. времени,}$$

и притом так, чтобы вероятность ошибки при передаче не превосходила заданного малого числа ε — она даже не позволяет сказать, как велико должно быть число N букв в блоке, которому сопоставляется одно кодовое обозначение, для того, чтобы такая передача оказалась возможной. Из этой теоремы следует лишь, что если разрешить выбирать N с к о л ь у г о д н о б о л ь ш и м, то передача со скоростью v_1 и вероятностью ошибки, не большей ε, будет в о з м о ж н а, какими бы ни были $v_1 < v$ и $\varepsilon > 0$. Так как, однако, при возрастании N существенно возрастает сложность расшифровки кода и увеличивается запаздывание при расшифровке, то для практики небезынтересно уметь оценить также и наименьшее значение вероятности ошибки ε, которое может быть достигнуто при передаче с заданной скоростью v_1 с помощью кода, сопоставляющего отдельные кодовые обозначения *не более чем N-буквенному блоку*, где N — какое-то з а д а н н о е число. Последнему вопросу посвящено большое число работ К. Шеннона, А. Файнстейна, П. Элайеса, Дж. Вольфовица, Р. Г. Галлагера, Р. Л. Добрушина и других ученых; подробное изложение полученных ими результатов может быть найдено, например, в статьях [158] — [161] и книгах [5] — [7], [21] и [22], которые все заметно сложнее настоящей книги. Не вдаваясь в подробности, мы ограничимся тем, что укажем здесь основной факт, вытекающий из всех этих работ.

Напомним, что передача N-буквенных блоков со скоростью $v_1 = L\frac{c_1}{H}$ букв/ед. времени, где $c_1 < c$, достигается в случае использования кодовых обозначений отдельных блоков, состоящих из $N_1 = \frac{H}{c_1}N$ элементарных сигналов каждый. Таким образом, числа N и N_1 пропорциональны друг другу; при вычислении вероятности ошибки,

соответствующей данным значением $v_1 = L\frac{c_1}{H}$ и N, удобно вместо v_1 и N использовать значения c_1 и N_1, более непосредственно описывающие процесс передачи информации по линии связи. Оказывается, что *при фиксированных* $c_1 < c$ *и* N_1 *всегда существует такой метод передачи* (т. е. метод кодирования — выбора $2^{c_1 N_1}$ N_1-членных кодовых обозначений — и метод декодирования — правила расшифровки принимаемых N_1-членных цепочек элементарных сигналов), *при котором вероятность ошибки при расшифровке каждого передаваемого кодового обозначения не превосходит величины*

$$\varepsilon = \frac{1}{a^{N_1}},$$

где a *— некоторое число, большее единицы* [1]). Число a, разумеется, зависит от c_1 — чем меньше c_1 (т. е., фактически, чем меньше скорость v_1 передачи информации по линии связи), тем оно больше. Вообще говоря, можно бы было думать, что при приближении c_1 (а, значит, и v_1) к нулю число a будет неограниченно возрастать (так как неограниченно уменьшая скорость передачи информации можно добиться сколь угодно малой вероятности ошибки при любом фиксированном N). На самом деле, однако, вывод приведенной выше формулы для ε при очень малых скоростях передачи оказывается довольно грубым и выте-

---

[1]) Приведенную здесь формулу можно, конечно, записать и так: $\varepsilon = \frac{1}{a_1^N}$, где $a_1 = a^{\frac{H}{c_1}}$ — новое число (так же большее единицы). При этом, однако, a_1 оказывается уже зависящим и от энтропии H передаваемого сообщения, в то время как a определяется лишь значением c_1 и характеристиками используемой линии связи. Читателям, знакомым с натуральными логарифмами, полезно также иметь в виду, что в научной литературе формула для ε обычно записывается в виде $\varepsilon = e^{-EN_1}$, где $e = 2{,}718\ldots$ — основание натуральных логарифмов, а $E = \ln a$ — натуральный логарифм (при основании e) числа a. Поскольку функция $y = e^{-Ex}$ в высшей математике называется экспоненциальной, последняя формула для величины ε, ограничивающей вероятность ошибки при передаче, часто называется экспоненциальной границей вероятности ошибки или даже просто экспоненциальной границей ошибки.

кающие из него результаты обычно указывают на стремление a к конечному значению при $c_1 \to 0$. При приближении же c_1 к c (т. е. скорости передачи v_1 к v) число a стремится к единице, так что и ε с ростом v_1 все более и более приближается к единице. Значение a при заданном c_1 будет разным для разных линий связи; схематический вид зависимости a от c_1 для фиксированной такой линии изображен на рис. 26.

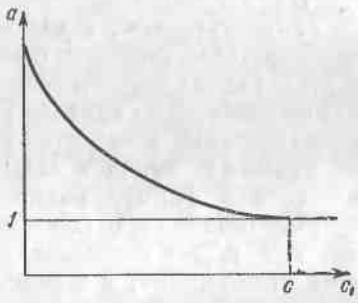


Рис. 26.

Ясно, что теорема Шеннона о кодировании при наличии помех непосредственно вытекает из приведенной формулы для ε и того факта, что $a > 1$ при любом $c_1 < c$. Более того, эта формула представляет собой заметное усиление теоремы Шеннона, утверждающей лишь что ε можно сделать сколь угодно малым, если только N (или, что то же самое, N_1) будет выбрано достаточно большим (но ничего не говорящей о том, как именно убывает ε с ростом N). Последнее обстоятельство мы как раз и имели в виду на стр. 351, когда отмечали, что в работе [158] были получены результаты более сильные, чем основная теорема о кодировании.

Перейдем теперь к случаю передачи сообщений со скоростью v_1, б о л ь ш е й предельной скорости $v = L\frac{c}{H}$ букв/ед. времени. Этот случай, вообще говоря, менее интересен, чем случай передачи со скоростью $v_1 < v$, а относящиеся к нему результаты менее неожиданны, чем основная теорема Шеннона; тем не менее он тоже заслуживает рассмотрения. На стр. 348 мы уже отмечали, что б е з о ш и б о ч н а я передача информации не может происходить со скоростью $v_1 > v$ букв/ед. времени; аналогичное утверждение можно найти и на стр. 357, где указывалось, что если $c_1 > c$, то 2^{HN} групп $\mathfrak{B}$, отвечающих кодовым обозначениям всевозможных «вероятных»

N-буквенных блоков, никак нельзя выбрать так, чтобы они не пересекались между собой. На самом деле, однако, приведенные на стр. 348 и 357 рассуждения сами по себе позволяют сделать лишь довольно поверхностные выводы. Верно, конечно, что безошибочная передача сообщений не может быть осуществлена со скоростью, превышающей $v = \frac{C}{H}$ букв/ед. времени; однако ведь и в случае передачи со скоростью $v_1 < v$ мы не утверждали, что возможна безошибочная передача сообщений, а говорили лишь, что в этом случае вероятность ошибки при передаче каждой буквы может быть сделана сколь угодно малой (с помощью использования в качестве кодовых обозначений достаточно длинных цепочек элементарных сигналов) [1]). Поэтому точная формулировка теоремы, обратной основной теореме Шеннона, должна состоять вовсе не в утверждении, что при $v_1 > v$ невозможна безошибочная передача информации, а в утверждении, что *для любого фиксированного* $v_1 > v$ *можно найти такое положительное число* $q_0 > 0$ (которое, видимо, должно зависеть от v_1 и при увеличении v_1 возрастать), *что в случае передачи информации по линии связи со скоростью* v_1 *вероятность ошибки* q *при расшифровке каждой переданной буквы сообщения при любом методе кодирования и декодирования* (независимо от значений N и N_1) *будет не меньше, чем* q_0. Предположение о справедливости такой обратной теоремы о кодировании при наличии помех было высказано еще в работе Шеннона [1], а затем было строго подтверждено Фано [6]; к рассмотрению ее доказательства (основывающегося на идеях Фано) мы сейчас и перейдем.

[1]) Заметим в этой связи, что в работе [162] Шеннон ввел также понятие пропускной способности C_0 линии связи при нулевой ошибке, определив ее как наибольшую скорость (в бит/ед. времени), с которой по данной линии можно вести точно *безошибочную* передачу информации. Рассуждение на стр. 348 показывает лишь, что ни для какой линии связи C_0 не может превосходить определенной на стр. 334 пропускной способности C — обстоятельство, которое кажется почти очевидным. На самом деле пропускная способность при нулевой ошибке обычно заметно меньше, чем C, причем любопытно, что она оказывается более сложной величиной, чем обычная пропускная способность — ее, вообще говоря, значительно труднее вычислить и она имеет куда менее наглядный смысл.

Прежде всего, однако, нам надо немного уточнить саму формулировку рассматриваемой теоремы. Легко видеть, что если при определении вероятности ошибки в расшифровке переданной буквы мы будем считать точно известным, какая именно буква передавалась, то приведенная выше формулировка никак не сможет оказаться верной. Действительно, мы можем, например, условиться расшифровывать все принимаемые буквы как первую букву алфавита — при этом вероятность ошибки будет равна нулю во всех случаях, когда на самом деле передавалась именно первая буква. С другой стороны ясно также, что расшифровывать все принимаемые буквы как первую букву нецелесообразно — здесь мы, фактически, вообще никак не используем линию связи и будем ошибаться каждый раз, когда передавалась любая буква, отличная от первой; поэтому средняя вероятность ошибки в этом случае будет велика. В то же время наиболее естественным представляется понимать вероятность q ошибки при расшифровке одной переданной буквы именно как с р е д н ю ю в е р о я т н о с т ь ошибки — и в дальнейшем именно так мы и будем поступать.

Итак, предположим, что передается текст, записанный с помощью n-буквенного алфавита $a_1, a_2, \ldots, a_n$, причем вероятности появления на произвольном (но фиксированном) месте в этом тексте букв $a_1, a_2, \ldots, a_n$ равны, соответственно, $p_1, p_2, \ldots, p_n$. Под q мы будем понимать *среднее значение вероятности ошибки*, т. е. величину

$$q = p_1q_1 + p_2q_2 + \ldots + p_nq_n, \qquad (*)$$

где q_1 — вероятность того, что буква a_1 после передачи по линии связи будет ошибочно принята за отличную от a_1 букву алфавита, и аналогичный смысл имеют величины $q_2, \ldots, q_n$. Существенно, что это же среднее значение q может быть вычислено и иначе. Пусть $p_1', p_2', \ldots, p_n'$ — вероятности обнаружить буквы $a_1, a_2, \ldots, a_n$ на произвольном (по фиксированном) месте сообщения, полученного на приемном конце линии связи с помощью расшифровки пришедшей по этой линии последовательности элементарных сигналов. Обозначим далее через q_1' вероятность того, что полученная на приемном конце буква a_1 была расшифрована неправильно (т. е. что на соответствующем месте

переданного сообщения стояла отличная от a_1 буква), а через $q_2', \ldots, q_n'$ — аналогичные же вероятности ошибок, относящиеся к случаям приема букв $a_2, \ldots, a_n$. Ясно, что вероятности $p_1', p_2', \ldots, p_n'$, вообще говоря, не будут совпадать с $p_1, p_2, \ldots, p_n$ (они зависят и от вероятностей $p_1, p_2, \ldots, p_n$, и от методов кодирования и декодирования, и от характеристик линии связи). Однако среднее значение вероятности ошибки при передаче одной буквы можно выразить и через них [1]):

$$q = p_1'q_1' + p_2'q_2' + p_n'q_n'. \qquad (**)$$

Именно формулой (**) мы, в основном, и будем пользоваться ниже.

Переходя к доказательству обратной теоремы о кодировании при наличии помех, начнем с простейшего случая, когда передаваемое сообщение записано с помощью двухбуквенного алфавита (буквы алфавита в этом случае удобно обозначать через «*а*» и «*б*»). Пусть β — это опыт, состоящий в определении передаваемой буквы сообщения (не передаваемого по линии связи элементарного сигнала, как это было на стр. 333, а именно буквы!), а α — опыт, состоящий в расшифровке буквы, полученной на приемном конце линии связи. Тогда оба эти опыта могут иметь два исхода («*а*» и «*б*»), причем вероятности двух возможных исходов опыта α равны p_1' и p_2' (так что $p_1' + p_2' = 1$), вероятности двух исходов β при условии, что опыт α имел исход «*а*», равны $1 - q_1'$ и q_1', вероятности тех же двух исходов β при условии, что опыт α имел исход «*б*», равны q_2' и $1 - q_2'$. Следовательно,

$$H_a(\beta) = -q_1' \log q_1' - (1 - q_1') \log (1 - q_1'),$$

$$H_б(\beta) = -q_2' \log q_2' - (1 - q_2') \log (1 - q_2'),$$

где $H_a(\beta)$ и $H_б(\beta)$ — условные энтропии опыта β при условии, что опыт α имел исход «*а*» и, соответственно, «*б*».

---

[1]) Нетрудно понять, что и правая часть формулы (*), и правая часть формулы (**) определяет с р е д н ю ю ч а с т о т у о ш и б о к в последовательности расшифровок большого числа букв переданного сообщения.

Введя, как и на стр. 75, в рассмотрение функцию

$$h(p) = -p \log p - (1-p) \log (1-p),$$

можно переписать последние равенства в виде

$$H_a(\beta) = h(q_1'), \qquad H_b(\beta) = h(q_2').$$

Отсюда видно, что

$$H_a(\beta) = p_1' H_a(\beta) + p_2' H_b(\beta) = p_1' h(q_1') + p_2' h(q_2').$$

Воспользуемся теперь тем, что функция $h(p)$ (график которой изображен на рис. 8, стр. 75) является выпуклой функцией в смысле, разъясненном в Приложении I на стр. 441. Поэтому в силу теоремы 2 этого Приложения (стр. 444) при любых неотрицательных p_1' и p_2' таких, что $p_1' + p_2' = 1$,

$$p_1' h(q_1') + p_2' h(q_2') \leqslant h(p_1' q_1' + p_2' q_2') = h(q),$$

где $q = p_1' q_1' + p_2' q_2'$. Таким образом,

$$H_a(\beta) \leqslant h(q) \qquad (\Lambda)$$

и

$$I(\alpha, \beta) = H(\beta) - H_a(\beta) \geqslant H(\beta) - h(q).$$

Вспомним теперь, что $I(\alpha, \beta)$ — это информация, содержащая в произвольной букве текста, принятого на приемном конце линии связи, относительно соответствующей буквы переданного сообщения. За единицу времени по линии передается v_1 букв, т. е. передается количество информации, равное $v_1 I(\alpha, \beta)$ (последовательные буквы сообщения мы считаем взаимно независимыми). Но ведь количество информации, переданное за единицу времени, не может превзойти пропускную способность C нашей линии связи [1]); поэтому, тем более,

$$v_1 |H(\beta) - h(q)| \leqslant C.$$

---

[1]) Напомним, что C равно максимальной информации о передаваемых элементарных сигналах, которую можно извлечь из принимаемых за единицу времени на приемном конце элементарных сигналов. Если кодирование последовательности букв сообщения в последовательность элементарных сигналов неоднозначно (например, если используется описанное ниже на стр. 375 «случайное кодирование»), то переход от опыта α к опыту α_1, состоящему в определении передаваемых элементарных сигналов,

Так как $\frac{C}{H(\beta)} = v$, то последнее неравенство удобно переписать в виде

$$1 - \frac{h(q)}{H(\beta)} \leqslant \frac{v}{v_1}. \qquad (Б)$$

Рассмотрим график функции $1 - \frac{h(q)}{H(\beta)} = g(q)$ (см. рис. 27, *а*, *б*, на которых эта функция изображена для

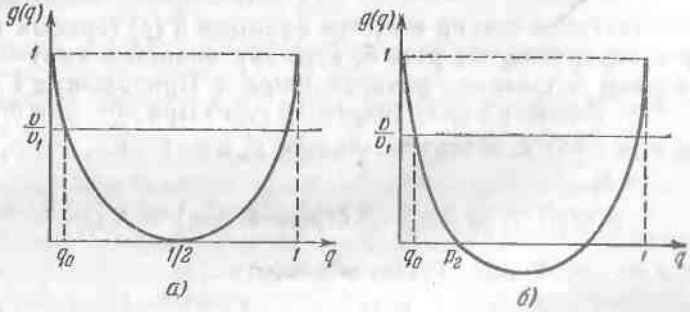


Рис. 27.

случая, когда $H(\beta) = 1$, т. е. когда исходы α_0 и α_1 опыта β равновероятны, и для одного из случаев, когда $H(\beta) < 1$). Он показывает, что если $v_1 \leqslant v$, т. е. если $\frac{v}{v_1} \geqslant 1$, то неравенство (Б) может быть удовлетворено при всех значениях q, включая и значение $q = 0$. Если же $v_1 > v$, т. е. $\frac{v}{v_1} < 1$, то это неравенство может выполняться лишь тогда, когда значение q принадлежит некоторому интервалу значений, расположенному левее точки q_0, где $q_0 > 0$.

Таким образом, *при* $v_1 > v$ *средняя вероятность ошибки* q *не может быть меньше некоторого* $q_0 > 0$, т. е. мы доказали то утверждение, которое выше было названо обратной теоремой о кодировании при наличии помех.

будет сопряжен с какой-то потерей информации; аналогичный эффект будет иметь и неоднозначность декодирования. Для нас здесь, однако, важно лишь то, что в любом случае информация $v_1 I(\alpha, \beta)$ о передававшихся буквах, содержащаяся в принятых буквах, не может быть больше, чем C (ср. стр. 123).

С ростом v_1 (т. е. с уменьшением $\frac{v}{v_1}$) значение q_0 увеличивается; при $v_1 \to \infty$ (т. е. $\frac{v}{v_1} \to 0$) оно, очевидно, стремится к вероятности p_2 той из передаваемых букв («a» или «$б$»), которая передается р е ж е, чем вторая буква. Впрочем, последний результат является совершенно естественным: ведь при очень большой скорости передачи

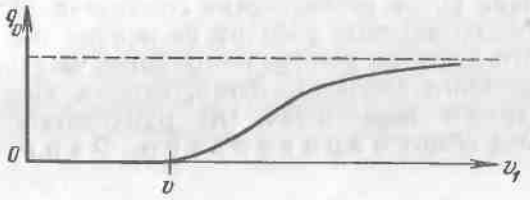


Рис. 28.

мы не сможем передать по нашей линии почти никакой полезной информации, и поэтому наиболее разумным методом расшифровки в этом случае будет метод, при котором в с е принимаемые буквы расшифровываются как буква, имеющая бо́льшую вероятность быть переданной. Но при такой расшифровке средняя вероятность ошибки q будет, очевидно, равна вероятности более редко употребляемой буквы (заметим, что для указанной «расшифровки» линия связи вообще не нужна). Если же вероятности появления обоих букв текста одинаковы, то при очень большой скорости передачи, когда линия связи вообще оказывается бесполезной, у нас вовсе не будет никаких оснований для выбора того или другого значения принятой буквы, так что здесь расшифровку вполне можно производить «наудачу». Средняя вероятность ошибки q в этом случае при $v_1 \to \infty$ будет стремиться к $\frac{1}{2}$, так как э т о и есть вероятность ошибки при расшифровке «наудачу» (и одновременно вероятность «более редкой» буквы). Схематически график зависимости нижней границы q_0 вероятности ошибки от скорости передачи v_1 изображен на рис. 28. То обстоятельство, что при $v_1 < v$ этот график совпадает с осью абсцисс (т. е. $q_0 = 0$), очевидно, соответствует основной теореме Шеннона о кодировании — тому,

что при $v_1 < v$ вероятность ошибки можно сделать сколь угодно малой (но, разумеется, наш вывод, доказывающий лишь, что средняя вероятность ошибки не может быть меньше, чем q_0, сам по себе еще не дает оснований утверждать, что при $v_1 < v$ величина q действительно может быть сделана сколь угодно малой). Положительность же q_0 при всех $v_1 > v$ как раз и составляет содержание обратной теоремы о кодировании.

Случай, когда передаваемое сообщение записано на языке, использующем алфавит из n букв $a_1, a_2, \ldots, a_n$, немного сложнее разобранного выше частного случая двухбуквенного алфавита. Здесь, однако, вместо совсем элементарного неравенства (A) приходится использовать более общее неравенство Фано, имеющее вид

$$H_\alpha(\beta) \leqslant h(q) + q \log (n-1), \qquad (A')$$

где α и β имеют тот же смысл, что выше, а q — это опять средняя вероятность ошибки.

Неравенство Фано (A') имеет очень простой и наглядный смысл. В самом деле, $H_\alpha(\beta)$ — это средняя степень неопределенности исхода опыта β при известном исходе опыта α. Но исход опыта β при известном исходе α можно найти с помощью следующих двух вспомогательных опытов. Прежде всего мы выясняем, *произошла ли или не произошла ошибка при передаче соответствующей буквы сообщения*. Это значит, что мы производим опыт γ, могущий иметь лишь два разных исхода (ответы «да, произошла» и «нет, не произошла»). Средняя вероятность положительного исхода опыта γ (ответа «да») равна q; используя выпуклость функции $h(p)$, отсюда легко вывести, что средняя степень неопределенности результата нашего первого вспомогательного опыта не может превзойти $h(q)$ (см. на стр. 365 неравенство, предшествующее неравенству (A), а также аналогичный общий вывод на стр. 391). Далее, ясно, что если ошибки при передаче не было (т. е. если исход опыта γ был отрицательным), то результаты опытов γ и α уже однозначно определяют исход β. Если, однако, исход γ окажется положительным (что происходит в среднем в доле q всех случаев), то знание исхода опыта γ еще не снимает неопределенности в исходе β — здесь нам придется произвести второй вспо-

могательный опыт γ_1, чтобы выяснить какой же именно из $n-1$ отличных от принятой букв была в действительности переданная буква. Этот второй опыт γ_1 может иметь $n-1$ различных исходов; поэтому степень его неопределенности (энтропия опыта γ_1) не может превзойти $\log(n-1)$. Понятно, что общая степень неопределенности $H_\alpha(\beta)$ должна равняться степени неопределенности первого вспомогательного опыта γ, сложенной со степенью неопределенности второго опыта γ_1, умноженной на среднюю частоту случаев, в которых этот второй опыт оказывается нужным. Отсюда сразу следует неравенство Фано (A′) (подробнее об этом см. текст, напечатанный мелким шрифтом на стр. 390—392).

Заметим теперь, что из неравенства Фано сразу следует неравенство

$$I(\alpha, \beta) \geqslant H(\beta) - h(q) - q\log(n-1).$$

Поэтому

$$v_1 [H(\beta) - h(q) - q\log(n-1)] \leqslant C,$$

где $C = v \cdot H(\beta)$, т. е.

$$1 - \frac{h(q) - q\log(n-1)}{H(\beta)} \leqslant \frac{v}{v_1}. \qquad (Б')$$

В частном случае, когда $H(\beta) = \log n$, функция $g_n(q) = 1 - \frac{h(q) - q\log(n-1)}{H(\beta)}$ лишь постоянным множителем отличается от функции $C(p)$, изображенной (в предположении, что $n=4$) на рис. 20 на стр. 340; для удобства мы воспроизводим аналогичный график и на следующей странице. Рядом на том же рис. 29 изображен схематический вид графика функции $g_n(q)$ при $H(\beta) < \log n$ (т. е. в случае, когда не все буквы алфавита равновероятны). Мы видим, что если $v_1 < v$ (т. е. $\frac{v}{v_1} > 1$), то неравенство (Б′) оказывается справедливым при любом $q \geqslant 0$; если же $v_1 > v$ (т. е. $\frac{v}{v_1} < 1$), то оно будет выполняться лишь для значений q, больших некоторого положительного числа q_0. Тем самым доказано, что обратная теорема о кодировании верна и в общем случае n-буквенного алфавита. Зависимость значения q_0 от скорости передачи v здесь основа имеет вид, схематически изображен-

ный на рис. 28; предельное значение q_0 при $v_1 \to \infty$ (т. е. при $\frac{v}{v_1} \to 0$) в случае, когда $H(\beta) = \log n$, равно $\frac{n-1}{n}$, а при уменьшении $H(\beta)$ оно уменьшается [1]).

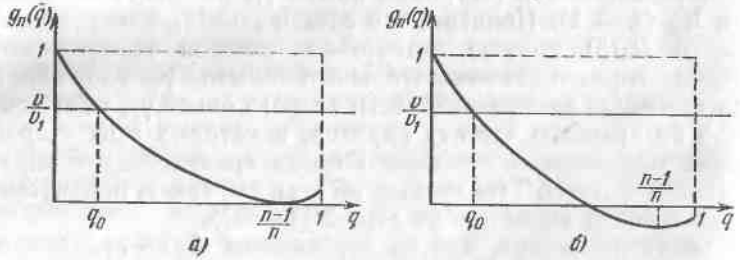


Рис. 29.

Заметим, что доказанные в настоящем параграфе основная теорема о кодировании при наличии помех и обратная теорема о кодировании сильно различаются как по методу доказательства, так и по самому своему характеру. Правда в формулировке обеих теорем речь идет о вероятности ошибки в определении одной переданной буквы; однако на самом деле при рассмотрении основной теоремы о кодировании исходные буквенные сообщения лишь вкратце упоминались в начале обсуждения, а затем рассматривались лишь непосредственно передаваемые по линии связи цепочки из N_1 элементарных сигналов. Существенным здесь было лишь то, что при использовании кодовых обозначений, представляющих собой цепочки («блоки») из N_1 элементарных сигналов, передача со скоростью $v_1 = L\frac{c_1}{H}$ букв/ед. времени тре-

[1]) Если v_1 очень велико, то линия связи становится бесполезной, и поэтому здесь остается расшифровывать все принимаемые буквы, как самую вероятную из передаваемых букв; при этом средняя вероятность ошибки q будет равна $1 - p_1$, где p_1 — наибольшая из вероятностей букв алфавита. Так как, однако, неравенство (Б′) не является точным, то получаемая из него оценка нижней границы q_0 средней вероятности ошибки не должна, вообще говоря, обязательно совпасть с наименьшим реально достижимым значением q.

бует, чтобы эти обозначения сопоставлялись N-буквенным сообщениям, где $N = \frac{c_1}{H} N_1$, т. е. чтобы (в случае достаточно большого N_1) имелось не менее чем $2^{c_1 N_1} = 2^{HN}$ «вероятных» (и притом имеющих практически одинаковую вероятность) кодовых обозначений. Таким образом, нам надо было лишь доказать, что *если* $c_1 < c$ (где $c = \max I(\alpha, \beta)$), *то при достаточно большом* N_1 *всегда можно выбрать* $2^{c_1 N_1}$ *кодовых обозначений длины* N_1 *таким образом, чтобы вероятность ошибки при расшифровке принятой на приемном конце линии связи цепочки из* N_1 *элементарных сигналов была меньше произвольного (но заранее заданного) числа* ε, *независимо от того, какое именно кодовое обозначение передавалось* (число ε здесь естественно выбирать очень малым — скажем, равным 0,001, или 0,0001, или 0,000001). Именно это утверждение (относящееся лишь к линии связи и передаваемым по ней длинным цепочкам элементарных сигналов, но никак не связанное с исходными буквенными сообщениями) и составляет суть основной теоремы о кодировании. Что же касается до обратной теоремы о кодировании, то она уже по существу относится к буквам исходного сообщения, но не к передаваемым по линии связи цепочкам элементарных сигналов.

Существует, однако, и другая теорема, также обратная основной теореме о кодировании, которая уже касается лишь линии связи и длинных цепочек передаваемых по ней элементарных сигналов. Согласно этой теореме, *если* $c_1 > c$ *и* N_1 *достаточно велико, то как бы мы ни выбирали* $2^{c_1 N_1}$ *равновероятных кодовых обозначений* (т. е. цепочек элементарных сигналов) *длины* N_1 *и как бы мы ни расшифровывали принимаемые цепочки из* N_1 *сигналов, все равно средняя вероятность того, что мы ошибемся в расшифровке принятой цепочки, будет больше произвольного (но заранее заданного) меньшего единицы числа* p_0 (число p_0 здесь естественно выбирать достаточно близким к единице — скажем, равным 0,999, или 0,9999, или 0,999999). При этом, разумеется, чем ближе будет p_0 к единице, тем большими будут требуемые значения N_1. Что же касается до средней вероятности ошибки, фигурирующей в условиях теоремы, то она, очевидно, совпадает

со средним арифметическим

$$\frac{p_{0,1} + p_{0,2} + \cdots + p_{0,2^{c_1 N_1}}}{2^{c_1 N_1}},$$

где $p_{0,i}$ — это вероятность ошибки в расшифровке в том случае, когда передавалось i-е из наших $2^{c_1 N_1}$ кодовых обозначений.

Справедливость сформулированной теоремы тесно связана с тем, что, как было показано на стр. 357, при $c_1 > C$ и очень большом N_1 общее число N_1-членных цепочек в $2^{c_1 N_1}$ группах $\mathcal{B}$ (т. е. в группах принимаемых «вероятных» цепочек, отвечающих имеющимся $2^{c_1 N_1}$ кодовым обозначениям длины N_1) будет в громадное число раз превосходить общее число всех «вероятных» принимаемых цепочек. Поэтому принимаемые N_1-членные цепочки будут, вообще говоря, одновременно принадлежать громадному числу различных групп $\mathcal{B}$, так что вероятность их правильной расшифровки будет очень мала. Эти соображения делают нашу теорему крайне правдоподобной, хотя они и не могут заменить ее строгого доказательства. Такое доказательство может быть найдено, например, в книгах [7], [21] или [22]; оно не очень просто, и мы на нем здесь не задержимся. Сама рассматриваемая теорема была названа Вольфовицем (вскоре ее строго доказавшим) усиленной обратной теоремой о кодировании при наличии помех — и это ее название часто используется в литературе по теории информации; однако оно не очень удачно, так как может создать неверное впечатление, что доказанная выше обычная обратная теорема о кодировании вытекает из этой новой теоремы (на самом же деле ни одна из приведенных здесь двух обратных теорем о кодировании не является следствием другой из них). Поэтому, вероятно, более целесообразно, следуя Галлагеру [22], называть рассматриваемую теорему обратной теоремой о блочном кодировании при наличии помех (т. е. о кодировании, использующем в качестве кодовых обозначений блоки элементарных сигналов фиксированной длины).

Вернемся теперь к более аккуратному разбору доказательства основной теоремы Шеннона о кодировании при наличии помех, о которой шла речь на стр. 349 и след. нашей книги. Начнем прежде всего с того, что, следуя работе С. З а р е м б а [163], приведем пример, наглядно показывающий, что из крайней малости общего числа цепочек $B_{j_1}B_{j_2}\ldots B_{j_N}$ в 2^{uN} группах $\mathfrak{B}$ по сравнению с полным числом таких «вероятных» цепочек еще вовсе не следует, что эти группы можно выбрать так, чтобы они не пересекались между собой. Рассмотрим с этой целью совокупность всевозможных цепочек из 10 элементарных сигналов, каждый из которых может принимать два значения. Ясно, что общее число таких цепочек равно $2^{10} = 1024$. Сопоставим, далее, каждой цепочке группу всех 10-членных цепочек, отличающихся от данной не более чем тремя сигналами. Кроме данной цепочки эта группа, очевидно, будет содержать $C_{10}^1 = 10$ цепочек, отличающихся от данной ровно одним сигналом, $C_{10}^2 = 45$ цепочек, отличающихся от данной двумя сигналами, и $C_{10}^3 = 120$ цепочек, отличающихся от данной тремя сигналами; итого группа будет состоять из $1 + 10 + 45 + 120 = 176$ цепочек. Так как 176 — это почти в ш е с т ь раз меньше, чем 1024, то можно было бы думать, что т р и цепочки здесь без особого труда можно выбрать так, чтобы отвечающие им 3 группы из 176 цепочек не пересекались между собой. Однако это неверно: можно показать, что *группы, отвечающие любым трем цепочкам, обязательно пересекаются*.

Действительно, обозначим два значения наших сигналов цифрами 0 и 1, и пусть, например, одна из групп — это группа, сопоставляемая «нулевой цепочке» из десяти нулей. Легко понять, что с этой группой не будут пересекаться лишь группы, сопоставляемые 10-членным цепочкам, содержащим больше шести цифр 1. Но в любых двух 10-членных цепочках, содержащих по семь или больше цифр 1, не меньше четырех из этих цифр 1 расположены в обеих цепочках на одних и тех же местах. Следовательно, наши две цепочки отличаются между собой не более чем на шесть сигналов, и, значит, отвечающие им группы пересекаются между собой. Разумеется, ничего не изменится, если мы начнем с любой другой

цепочки (а не с «нулевой цепочки» 0000000000): две наши группы из 176 цепочек, не пересекающиеся с одной и той же третьей группой, обязательно пересекаются между собой.

В точности так же показывается, что и *при любом k среди групп $(3k+1)$-членных цепочек, отличающихся от какой-то одной такой цепочки не более чем в k сигналах, нельзя найти более двух непересекающихся групп*. Между тем можно показать, что отношение числа цепочек в такой группе (равного сумме $1+C^1_{3k+1}+C^2_{3k+1}+\ldots$ $\ldots+C^k_{3k+1}$) к общему числу всех вообще $(3k+1)$-членных цепочек (равному 2^{3k+1}) с возрастанием k все время убывает; так, при $k=8$, $3k+1=25$ это отношение будет уже близко к $\frac{1}{20}$, а если выбрать k достаточно большим, то можно даже добиться, чтобы указанное отношение оказалось с к о л ь у г о д н о м а л ы м (меньшим любого заранее заданного малого числа). Таким образом, общее число цепочек в трех группах может составлять ничтожную часть числа всех вообще цепочек — и тем не менее любые три группы обязательно будут пересекаться. Поэтому и в случае теоремы Шеннона нельзя просто мотивировать возможность выбора 2^{HN} непересекающихся групп тем, что общее число цепочек в них очень мало по сравнению с числом всех вообще «вероятных» цепочек; требуется еще строго доказать, что в данном случае дело обстоит не так, как в примере Заремба.

На самом деле до сих пор никому не удалось строго доказать, что 2^{HN} цепочек $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}$ можно выбрать таким образом, чтобы никакие две из отвечающих им 2^{HN} групп $\mathcal{B}$ не пересекались между собой. Однако можно показать, что наверное существует такой выбор этих цепочек, при котором соответствующие группы $\mathcal{B}$ п о ч т и н е б у д у т п е р е с е к а т ь с я,— и поэтому их пересечением можно будет пренебречь. Этот факт может быть сделан очевидным с помощью следующего рассуждения, принадлежащего, в основных чертах, Ш е н н о н у [1]. Начнем с того, что выберем нужные нам 2^{HN} цепочек $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}$ с помощью метода, который сначала может показаться явно неразумным, а именно — н а у д а ч у. Такой выбор «наудачу» можно осуществить

так: перенумеруем все $2^{H(\beta)N_1}$ «вероятных» цепочек $A_{i_1}A_{i_2}\dots A_{i_{N_1}}$ в произвольном порядке, выпишем их номера на $2^{H(\beta)N_1}$ бумажек, сложим эти бумажки в урну и перемешаем, а затем будем 2^{HN} раз подряд вытаскивать из урны по одной бумажке, возвращая после каждого извлечения вытащенную бумажку обратно и снова перемешивая содержание урны. Цепочки $A_{i_1}A_{i_2}\dots A_{i_{N_1}}$ с извлеченными номерами мы и примем за наши 2^{HN} кодовых обозначений (такой метод выбора кодовых обозначений носит название *случайного кодирования*). Ясно, что при случайном кодировании один и тот же номер может оказаться вытянутым два или больше раз, так что некоторые из 2^{HN} отобранных цепочек окажутся просто *совпадающими* между собой и их, разумеется, никак нельзя будет различить на приемном конце линии связи; одно это обстоятельство создает впечатление, что предложенный метод выбора кодовых обозначений безусловно является нерациональным. На самом деле, однако, при большом N вероятность такого совпадения будет ничтожно мала (так как число $2^{H(\beta)N_1} = 2^{\frac{H(\beta)}{c_1}HN}$ различных «вероятных» цепочек при большом N будет в очень много раз больше числа 2^{HN}); как будет видно из дальнейшего, это позволит нам не считаться с возможностью совпадений.

Предположим теперь, что по нашей линии связи были последовательно переданы сигналы A_{i_1}, A_{i_2}, ..., $A_{i_{N_1}}$, совокупность которых как раз и составляет одно из выбранных нами кодовых обозначений. Из-за наличия помех эти сигналы, вообще говоря, будут как-то искажаться при передаче; в результате на приемном конце линии мы получим отличную от $A_{i_1}A_{i_2}\dots A_{i_{N_1}}$ последовательность сигналов $B_{j_1}B_{j_2}\dots B_{j_{N_1}}$. Ясно, что цепочка $B_{j_1}B_{j_2}\dots B_{j_{N_1}}$ с очень близкой к единице вероятностью будет принадлежать отвечающей цепочке $A_{i_1}A_{i_2}\dots A_{i_{N_1}}$ группе $\mathfrak{B}$. Но эта же цепочка $B_{j_1}B_{j_2}\dots B_{j_{N_1}}$ будет одновременно принадлежать также и группам $\mathfrak{B}$, отвечающим целому ряду других цепочек из N_1 передаваемых сигналов; именно это обстоятельство и делает затруднительной расшифровку принятого сообщения.

Общее число различных «вероятных» цепочек $A_{i_1}A_{i_2}\dots A_{i_{N_1}}$ таких, что отвечающие им группы $\mathcal{B}$ содержат заданную цепочку $B_{j_1}B_{j_2}\dots B_{j_{N_1}}$, можно оценить без труда. В самом деле, общее число «вероятных» $2N_1$-членных цепочек $A_{i_1}A_{i_2}\dots A_{i_{N_1}}B_{j_1}B_{j_2}\dots B_{j_{N_1}}$, как мы знаем, равно $2^{H(\alpha\beta)N_1}$, причем входящие в них цепочки $B_{j_1}B_{j_2}\dots B_{j_{N_1}}$ все принадлежат к числу $2^{H(\alpha)N}$, равноправных «вероятных» принимаемых цепочек. Таким образом, число «вероятных» $2N_1$-членных цепочек превосходит число «вероятных» N_1-членных цепочек $B_{j_1}B_{j_2}\dots B_{j_N}$

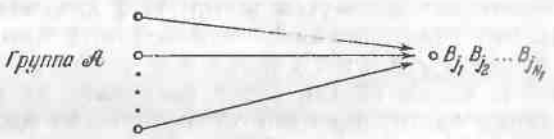


Рис. 30.

в $2^{H(\alpha\beta)N_1}:\ 2^{H(\alpha)N_1} = 2^{H_\alpha(\beta)N_1}$ раз. Отсюда можно заключить, что всевозможные «вероятные» $2N_1$-членные цепочки получаются с помощью комбинирования каждой из $2^{H(\alpha)N_1}$ «вероятных» цепочек $B_{j_1}B_{j_2}\dots B_{j_{N_1}}$ принимаемых сигналов с $2^{H_\alpha(\beta)N_1}$ различными цепочками $A_{i_1}A_{i_2}\dots A_{i_{N_1}}$ передаваемых сигналов. Именно эти $2^{H_\alpha(\beta)N_1}$ передаваемых цепочек и обладают тем свойством, что цепочка $B_{j_1}B_{j_2}\dots B_{j_{N_1}}$ входит в отвечающие им группы $\mathcal{B}$. Совокупность всех этих цепочек $A_{i_1}A_{i_2}\dots A_{i_{N_1}}$ мы назовем *группой $\mathcal{A}$, отвечающей цепочке* $B_{j_1}B_{j_2}\dots B_{j_{N_1}}$ (см. схематический рис. 30, на котором ведущие от цепочек группы $\mathcal{A}$ к цепочке $B_{j_1}B_{j_2}\dots B_{j_{N_1}}$ стрелки означают, что группы $\mathcal{B}$ этих передаваемых цепочек содержат $B_{j_1}B_{j_2}\dots B_{j_N}$ и что, следовательно, существует реальная вероятность приема этих цепочек на приемном конце линии связи как цепочки $B_{j_1}B_{j_2}\dots B_{j_{N_1}}$).

Рассмотрение группы $\mathcal{A}$ из $2^{H_\alpha(\beta)N_1}$ возможных цепочек передаваемых сигналов, отвечающей полученному на приемном конце линии связи сообщению $B_{j_1}B_{j_2}\dots B_{j_{N_1}}$, играет основную роль в том методе расшифровки

этого сообщения, которым мы будем пользоваться. А именно, если указанная группа $\mathcal{A}$ содержит ровно одно из наших кодовых обозначений, то мы будем считать, что именно это обозначение и было передано. В случаях же, когда эта группа $\mathcal{A}$ содержит больше одного кодового обозначения, или не содержит ни одного кодового обозначения, или, наконец, принятое сообщение вообще не входит в число $2^{H(\alpha)N}$ «вероятных» цепочек $B_{j_1} B_{j_2} \dots B_{j_N}$, мы будем считать переданным какое-то одно произвольно выбранное из имеющихся кодовых обозначений (например, будем во всех этих случаях считать, что передавалось кодовое обозначение с номером 1; из дальнейшего будет видно, что на самом деле это соглашение не играет никакой роли).

После того, как мы выбрали метод кодирования (т. е. нахождения 2^{HN} нужных нам кодовых обозначений) и метод декодирования (т. е. расшифровки принимаемых сообщений), можно перейти к определению вероятности ошибки при декодировании. Здесь, однако, нам сразу встретится одно затруднение. Пусть передано было кодовое обозначение $A_{i_1} A_{i_2} \dots A_{i_N}$, а принято было сообщение $B_{j_1} B_{j_2} \dots B_{j_N}$; обозначим через P вероятность того, что, используя описанный выше способ расшифровки поступивших сообщений, мы придем к неверному выводу, т. е. заключим, что было передано какое-то отличное от $A_{i_1} A_{i_2} \dots A_{i_N}$ кодовое обозначение. Ясно, что величина P в принципе может быть различной для различных кодовых обозначений $A_{i_1} A_{i_2} \dots A_{i_N}$; так, например, используемый нами метод декодирования явно ставит кодовое обозначение с номером 1 в особое положение. Надо ли из-за этого отдельно вычислять величину P для различных кодовых обозначений (или хотя бы отдельно для первого и для всех остальных таких обозначений)? Мы увидим ниже, что не надо — оценки, которые будут использоваться, будут справедливы для всех кодовых обозначений без исключения. Но, кроме того, наш метод расшифровки зависит и от выбора используемых кодовых обозначений, а этот выбор, как мы знаем, определяется исходом опыта, состоящего в 2^{HN}-кратном извлечении бумажки из урны, т. е. зависит от некоторого случайного события. Поэтому

и величина P является случайной величиной в том смысле, который был разъяснен на стр. 23. Такая величина может иметь много разных значений; ниже мы вычислим лишь среднее значение этой случайной величины P.

Мы знаем, что если число $N_1 = \frac{H}{c_1} N$ достаточно велико, то сообщение $A_{i_1} A_{i_2} \dots A_{i_{N_1}}$ перейдет в одну из цепочек $B_{j_1} B_{j_2} \dots B_{j_{N_1}}$ из отвечающей этому сообщению группы $\mathfrak{B}$ со сколь угодно близкой к единице вероятностью. Далее мы будем считать, что N_1 настолько велико, что указанная вероятность *не меньше, чем* $1 - \frac{\varepsilon}{4}$, где ε — заранее выбранное малое число. Пусть теперь $B_{j_1} B_{j_2} \dots B_{j_{N_1}}$ — это «вероятная» цепочка принимаемых сигналов, которая входит в группу $\mathfrak{B}$, отвечающую некоторому кодовому обозначению $A_{i_1} A_{i_2} \dots A_{i_{N_1}}$. Через Q мы обозначим вероятность того, что рассматриваемая цепочка входит одновременно также и в группу $\mathfrak{B}$, отвечающую *по крайней мере еще одному кодовому обозначению* (т. е., иначе говоря, вероятность того, что группа $\mathfrak{A}$, отвечающая нашей цепочке $B_{j_1} B_{j_2} \dots B_{j_{N_1}}$, содержит кроме $A_{i_1} A_{i_2} \dots A_{i_{N_1}}$ по крайней мере еще одно кодовое обозначение). Ясно, что Q, так же как и P, будет случайной величиной. Далее ясно, что принятое сообщение $B_{j_1} B_{j_2} \dots B_{j_{N_1}}$ наверное будет безошибочно расшифровано, если выполняются следующие два условия:

A — это сообщение входит в группу $\mathfrak{B}$, отвечающую переданному кодовому обозначению;

B — кроме этой группы оно не входит ни в одну из групп $\mathfrak{B}$, отвечающих прочим используемым кодовым обозначениям.

Поэтому неверная расшифровка может иметь место лишь тогда, когда или не выполняется условие A, или же не выполняется условие B. Но мы знаем, что вероятность суммы $\bar{A} + \bar{B}$ двух событий $\bar{A}$ (это событие состоит в том, что A не имеет места) и $\bar{B}$ (событие B не имеет места) не превосходит суммы вероятностей событий $\bar{A}$ и $\bar{B}$ (см. выше стр. 28). Следовательно, вероятность ошибки при расшифровке принятой N_1-членной цепочки должна

удовлетворять неравенству

$$P \leqslant \frac{\varepsilon}{4} + Q;$$

здесь $\frac{\varepsilon}{4}$ больше или равно вероятности того, что не выполняется условие A (т. е. что имеет место событие $\bar{A}$), а Q равно вероятности невыполнения B (т. е. вероятности события $\bar{B}$). В последнем неравенстве $\frac{\varepsilon}{4}$ — это фиксированное число, а P и Q — случайные величины; отсюда видно, что для оценки среднего значения величины P нам надо только оценить среднее значение величины Q.

Помимо кодового обозначения $A_{l_1}, A_{l_2} \ldots A_{l_{N_1}}$ имеется еще $2^{HN} - 1$ других кодовых обозначений. Перенумеруем заново эти $2^{HN} - 1$ обозначений в произвольном порядке и обозначим через a_i случайное событие, состоящее в том, что группа $\mathcal{A}_i$, отвечающая цепочке $B_{j_1}B_{j_2}\ldots B_{j_{N_1}}$, содержит i-е кодовое обозначение. Условие B не будет выполняться в том и только том случае, когда выполняется, по крайней мере, одно из событий $a_1, a_2, \ldots, a_{2^{HN}-1}$; иначе говоря, событие $\bar{B}$ равно сумме событий $a_1 + a_2 + \ldots + a_{2^{HN}-1}$. Но вероятность суммы событий не может превзойти сумму вероятностей этих событий (см. стр. 28); поэтому

$$Q \leqslant q_1 + q_2 + \ldots + q_{2^{HN}-1},$$

где q_i — это вероятность события a_i.

Попытаемся теперь определить среднее значение вероятности q_i. Так как i-е кодовое обозначение выбирается наудачу (так же как и все остальные кодовые обозначения), то оно с одинаковой вероятностью $\frac{1}{2^{H(\beta)N_1}}$ может совпасть с каждой из $2^{H(\beta)N_1}$ имеющихся «вероятных» цепочек из N_1 передаваемых сигналов A_l. В тех $2^{H_\alpha(\beta)N_1}$ случаях, когда оно оказывается совпадающим с одной из $2^{H_\alpha(\beta)N_1}$ цепочек, входящих в группу $\mathcal{A}_i$, отвечающую цепочке $B_{j_1}B_{j_2}\ldots B_{j_{N_1}}$, событие a_i будет иметь место, т. е. его вероятность будет равна единице; в остальных $2^{H(\beta)N_1} - 2^{H_\alpha(\beta)N_1}$ случаях это событие не будет иметь места, т. е. его вероятность будет равна нулю.

Итак, $q_i = 1$ при $2^{H_\alpha(\beta)N_1}$ равновероятных исходах опыта, состоящего в извлечении бумажки из урны с $2^{H(\beta)N_1}$ бумажками, и $q_i = 0$ при $2^{H(\beta)N_1} - 2^{H_\alpha(\beta)N_1}$ остальных исходах; поэтому ясно, что

$$\text{ср.зн. } q_i = \frac{2^{H_\alpha(\beta)N_1}}{2^{H(\beta)N_1}} \cdot 1 + \frac{2^{H(\beta)N_1} - 2^{H_\alpha(\beta)N_1}}{2^{H(\beta)N_1}} \cdot 0 = \frac{2^{H_\alpha(\beta)N_1}}{2^{H(\beta)N_1}} = 2^{[H_\alpha(\beta) - H(\beta)]N_1}.$$

Но средние значения всех величин q_i одинаковы (ибо все номера i равноправны), а Q не превосходит суммы $2^{HN} - 1$ величин q_i; поэтому ср.зн.Q не больше, чем

$$(2^{HN} - 1) \cdot 2^{[H_\alpha(\beta) - H(\beta)]N_1} < 2^{HN} \cdot 2^{-\frac{H(\beta) - H_\alpha(\beta)}{c_1} HN} = 2^{-\left(\frac{c}{c_1} - 1\right) HN}.$$

Вспомним теперь, что $c_1 < c$; отсюда вытекает, что при большом N стоящее в правой части последнего неравенства выражение будет представлять из себя число 2, возведенное *в очень большую по абсолютной величине отрицательную степень*, т. е. будет очень мало. В частности, как бы ни было мало выбранное число ε, число N можно будет взять столь большим, что это выражение (а значит и ср.зн.Q) будет меньше, чем $\frac{\varepsilon}{4}$.

Но мы знаем, что $P \leqslant \frac{\varepsilon}{4} + Q$; поэтому

$$\text{ср.зн.}P \leqslant \text{ср.зн.}Q + \frac{\varepsilon}{4}.$$

А так как при достаточно большом N

$$\text{ср. зн. } Q < \frac{\varepsilon}{4},$$

то, выбрав N достаточно большим, всегда можно добиться того, чтобы среднее значение вероятности P ошибки при расшифровке любого из 2^{HN} кодовых обозначений (соответствующих 2^{HN} «вероятным» N-буквенным сообщениям) было меньше, чем $\frac{\varepsilon}{2}$, где ε—любое наперед выбранное (сколь бы оно ни было мало!) положительное число.

Полученный результат позволяет без труда доказать основную теорему Шеннона о кодировании при наличии помех. Воспользуемся для этого тем, что *среднее значение любой случайной величины не может быть меньше всех ее возможных значений* (см. стр. 24). В применении к нашему случаю это означает, что *среди* $(2^{H(\beta)N_1})^{2^{HN}}$ *различных возможных выборов наших* 2^{HN} *кодовых обозначений* (т. е. среди всех различных исходов опыта, состоящего в 2^{HN}-кратном вытаскивании бумажки из урны с $2^{H(\beta)N_1}$ бумажками) *наверное будет хоть один, для которого значение* P *окажется меньшим, чем* $\frac{\varepsilon}{2}$.

Последнее утверждение уже весьма близко к тому, которое мы хотим доказать, но для нашей цели оно все же еще недостаточно. Дело в том, что P — это вероятность неправильной расшифровки на приемном конце линии некоторого фиксированного переданного кодового обозначения $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}$; нам же надо доказать, что существует такой выбор этих обозначений, для которого вероятность неправильной расшифровки при передаче по линии связи любого из них будет меньше, чем ε. Обозначим вероятность ошибки при расшифровке переданного i-го кодового обозначения через P_i; тогда $P_1, P_2, \ldots, P_{2^{HN}}$ — это случайные величины, среднее значение каждой из которых может быть оценено точно так же, как выше мы оценивали среднее значение какой-то одной из них — той, которую мы обозначили буквой P. Поэтому средние значения всех величин P_i меньше, чем $\varepsilon/2$; но отсюда еще не следует, что хоть для одного из выборов наудачу 2^{HN} кодовых обозначений значения всех величин $P_1, P_2, \ldots, P_{2^{HN}}$ будут одновременно меньше, чем $\frac{\varepsilon}{2}$.

Последнюю трудность можно, однако, обойти с помощью следующего искусственного приема. Выберем наудачу не 2^{HN} цепочек $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}$, а *в два раза большее* их число, т. е. $2\cdot 2^{HN}$ цепочек. Примем эти $2\cdot 2^{HN}$ цепочек $A_{i_1}A_{i_2}\ldots A_{i_{N_1}}$ за $2\cdot 2^{HN}$ кодовых обозначений и будем передавать все их по нашей линии связи, расшифровывая принятое сообщение $B_{j_1}B_{j_2}\ldots B_{j_{N_1}}$ точно так же, как это было описано выше. Поскольку $2\cdot 2^{HN} = 2^{HN+1} =$

$= 2^{H_1 N}$, где $H_1 = H + \frac{1}{N}$ при достаточно большом N сколь угодно мало отличается от H, то нетрудно видеть, что все предыдущие оценки останутся в силе и в этом случае. Иначе говоря, здесь также можно доказать, что среднее значение вероятности P ошибки при расшифровке цепочки $B_{j_1} B_{j_2} \dots B_{j_{N_1}}$, полученной на приемном конце линии связи, по которой передавалось какое-то одно из наших $2 \cdot 2^{HN} = 2^{H_1 N}$ кодовых обозначений, при достаточно большом N обязательно будет меньше, чем $\frac{\varepsilon}{2}$. Таким образом, если $P_1, P_2, \dots, P_{2 \cdot 2^{HN}}$ — вероятности ошибки при расшифровке переданного по линии связи первого, второго, ..., $2 \cdot 2^{HN}$-го кодового обозначения, то при достаточно большом N средние значения всех этих величин будут меньше, чем $\frac{\varepsilon}{2}$.

Рассмотрим теперь новую случайную величину

$$P_0 = \frac{P_1 + P_2 + \dots + P_{2 \cdot 2^{HN}}}{2 \cdot 2^{HN}},$$

равную среднему арифметическому всех величин P_i. Ясно, что если средние значения всех P_i меньше, чем $\frac{\varepsilon}{2}$, то и среднее значение P_0 также меньше, чем $\frac{\varepsilon}{2}$. Применим теперь утверждение о том, что *среднее значение случайной величины не может быть меньше всех ее значений*, к величине P_0; тогда мы получим, что хоть для одного из возможных выборов наудачу $2 \cdot 2^{HN}$ кодовых обозначений значение P_0 должно быть меньше, чем $\frac{\varepsilon}{2}$. Но все величины $P_1, P_2, \dots, P_{2 \cdot 2^{HN}}$ — это вероятности, которые не могут быть отрицательными; поэтому ясно, что если бы 2^{HN} или больше из этих величин оказались бы не меньше ε, то их среднее арифметическое P_0 было бы не меньше, чем $\frac{\varepsilon}{2}$. Отсюда вытекает, что 2^{HN} *или больше значений величин* P_i, $i = 1, 2, \dots, 2 \cdot 2^{HN}$, *должны быть меньше, чем* ε. Отвечающие соответствующим номерам i цепочки $A_{i_1} A_{i_2} \dots A_{i_{N_i}}$ мы и примем за нужные нам 2^{HN} кодовых обозначений — и будем далее пере-

давать по линии связи только их и расшифровывать принимаемые цепочки $B_{j_1}B_{j_2}\dots B_{j_{N_1}}$ так, как если бы никаких других кодовых обозначений не существовало. Заметим теперь, что во всех тех случаях, когда для принятой цепочки $B_{j_1}B_{j_2}\dots B_{j_{N_1}}$ выписанные на стр. 378 условия A и B оказываются справедливыми в отношении к $2\cdot 2^{HN}$ кодовым обозначениям, они тем более будут справедливы и тогда, когда половина из ранее использовавшихся кодовых обозначений отбрасывается. Поэтому все выведенные выше неравенства для вероятностей ошибок P_t не могут ухудшиться из-за того, что мы отбросили половину из первоначально выбранных $2\cdot 2^{HN}$ кодовых обозначений. Тем самым мы доказали то, что нам было нужно, а именно, что *при достаточно большом N всегда существуют такой выбор 2^{HN} кодовых обозначений $A_{i_1}A_{i_2}\dots A_{i_{N_1}}$ и выбор метода расшифровки принимаемых цепочек $B_{j_1}B_{j_2}\dots B_{j_{N_1}}$, для которых вероятность ошибки при расшифровке будет меньше, чем ε, независимо от того, какое именно кодовое обозначение передавалось по линии связи.*

При определении пропускной способности на стр. 334 мы исходили из предположения о том, что если c — наибольшее количество информации, которое можно получить при приеме одного переданного по линии связи элементарного сигнала, то при приеме L таких сигналов нельзя получить больше Lc единиц информации. Это предположение кажется совершенно естественным; однако строгое его доказательство все же не является очевидным. Сейчас мы вкратце поясним как может быть проведено такое доказательство.

Пусть β — опыт, состоящий в определении значения одного переданного элементарного сигнала, а α — в определении значения сигнала принятого. Тогда по условию $I(\alpha, \beta) \leqslant c$. Требуется доказать, что если $\beta_1\beta_2\dots\beta_L$ — это сложный опыт, состоящий в последовательном осуществлении опытов $\beta_1, \beta_2, \dots, \beta_L$ (т. е. состоящий в последовательной передаче L элементарных сигналов), а $\alpha_1\alpha_2\dots\alpha_L$ — второй сложный опыт, заключающийся в приеме этих L переданных сигналов, то всегда

$$I(\alpha_1\alpha_2\dots\alpha_L,\ \beta_1\beta_2\dots\beta_L) \leqslant Lc.$$

Для этого, разумеется, достаточно доказать, что

$$I(\alpha_1\alpha_2\dots\alpha_L,\beta_1\beta_2\dots\beta_L) \leqslant I(\alpha_1,\beta_1) + I(\alpha_2,\beta_2) + \dots + I(\alpha_L,\beta_L)$$

— ведь каждый член в правой части последнего неравенства равен информации об одном переданном сигнале, содержащейся в соответствующем принятом сигнале, т. е. не может превосходить c.

Будем для простоты считать, что $L = 2$ — это не является ограничением, так как всегда можно подставить в полученное неравенство вместо α_2 и β_2 сложные опыты $\alpha_2\alpha_3 \ldots \alpha_L$ и $\beta_2\beta_3 \ldots \beta_L$, а затем воспользоваться методом математической индукции по числу L. Что же касается доказательства нашего неравенства при $L = 2$, то его можно получить очень быстро, если применить формулу тройной информации (см. выше, стр. 127), согласно которой

$$I(\beta\gamma, \alpha) + I(\beta, \gamma) = I(\alpha\gamma, \beta) + I(\alpha, \gamma).$$

Полагая в этой формуле $\beta = \alpha_1$, $\gamma = \alpha_2$ и $\alpha = \beta_1\beta_2$, получим

$$I(\alpha_1\alpha_2, \beta_1\beta_2) + I(\alpha_1, \alpha_2) = I(\beta_1\beta_2\alpha_2, \alpha_1) + I(\beta_1\beta_2, \alpha_2).$$

Воспользуемся теперь тем, что информация, содержащаяся в сложном опыте $\beta\gamma$ относительно некоторого опыта α, будет равна $I(\beta, \alpha)$, если только условная вероятность исхода α при заданном исходе сложного опыта $\beta\gamma$ на самом деле зависит лишь от исхода β (см. выше, стр. 122). В нашем случае условные вероятности исходов опыта α_1 при заданном исходе опыта $\beta_1\beta_2\alpha_2$, очевидно, могут зависеть лишь от исхода β_1; точно также условные вероятности исходов α_2 при заданном исходе $\beta_1\beta_2$ зависят лишь от исхода β_2. Поэтому

$$I(\beta_1\beta_2\alpha_2, \alpha_1) = I(\beta_1, \alpha_1), \quad I(\beta_1\beta_2, \alpha_2) = I(\beta_2, \alpha_2),$$

а так как $I(\alpha_1, \alpha_2) \geqslant 0$ (информация всегда неотрицательна), то

$$I(\alpha_1\alpha_2, \beta_1\beta_2) \leqslant I(\beta_1, \alpha_1) + I(\beta_2, \alpha_2),$$

что и требовалось доказать [1]).

Перейдем теперь к изложению еще одного метода доказательства основной теоремы Шеннона о кодировании

---

[1]) При выводе равенств $I(\beta_1\beta_2\alpha_2, \alpha_1) = I(\beta_1, \alpha_1)$ и $I(\beta_1\beta_2, \alpha_2) = I(\beta_2, \alpha_2)$ мы фактически воспользовались тем, что условная вероятность исхода B_kB_l опыта $\alpha_1\alpha_2$ при условии, что опыт $\beta_1\beta_2$ имел исход A_iA_j (т. е. вероятность приема пары сигналов B_kB_l, если была передана пара A_iA_j), представима в виде $P_{A_iA_j}(B_kB_l) = P_{A_i}(B_k) \cdot P_{A_j}(B_l)$, где $P_{A_i}(B_k)$ и $P_{A_j}(B_l)$ — известные нам характеристики помех в линии связи. Действительно, именно отсюда вытекает то, что исход α_1 зависит лишь от исхода β_1, а исход α_2 — лишь от исхода β_2. Если теперь мы подставим эти вероятности $P_{A_iA_j}(B_kB_l)$ в выражение для условной энтропии $H_{\beta_1\beta_2}(\alpha_1\alpha_2)$, то с помощью несложных преобразований можно будет непосредственно доказать, что $H_{\beta_1\beta_2}(\alpha_1\alpha_2) = H_{\beta_1}(\alpha_1) + H_{\beta_2}(\alpha_2)$ и, следовательно, $I(\alpha_1\alpha_2, \beta_1\beta_2) = H(\alpha_1\alpha_2) - H_{\beta_1\beta_2}(\alpha_1\alpha_2) \leqslant I(\alpha_1, \beta_1) + I(\alpha_2, \beta_2)$ (так как $H(\alpha_1\alpha_2) \leqslant H(\alpha_1) + H(\alpha_2)$; см. стр. 93). Однако такое доказательство оказывается все же несколько длиннее приведенного выше более искусственного доказательства.

при наличии помех для простейшей *двоичной симметричной линии связи* [1]). По такой линии могут передаваться два элементарных сигнала A_1 и A_2, причем на приемном конце каждый из них с вероятностью $1-p$ расшифровывается правильно, а с вероятностью p принимается за другой сигнал. Как отмечалось на стр. 337, мы без ограничения общности можем считать, что $p < 1/2$. В качестве кодовых обозначений будем использовать последовательности $A_{i_1} A_{i_2} \ldots A_{i_{N_1}}$ из N_1 сигналов; здесь все i_k (где $k = 1, 2, \ldots, N_1$) могут принимать значения 1 или 2, и поэтому всего существует 2^{N_1} различных таких последовательностей. Пусть ε — некоторое заранее заданное малое число; потребуем, чтобы *вероятность ошибки при расшифровке любого переданного кодового обозначения не превосходила* ε. Нас будет интересовать, как много кодовых обозначений можно выбрать, не вступая в противоречие с выделенным курсивом условием. Ниже мы докажем, что при достаточно большом N_1 число K таких кодовых обозначений может быть сделано сколь угодно близким к 2^{cN_1}, где

$$c = 1 + (1-p)\log(1-p) + p\log p$$

— пропускная способность используемой линии связи, отнесенная к одному передаваемому сигналу. Поскольку сообщение о выборе одного обозначения из K возможных может доставить $\log K$ бит информации, отсюда уже будет следовать, что по этой линии можно передавать информацию, со скоростью, сколь угодно близкой к $C = Lc$ бит/ед. времени — и притом так, чтобы вероятность ошибки при расшифровке каждого переданного сигнала не превосходила ε. Тем самым теорема Шеннона будет доказана.

При доказательстве прежде всего требуется указать метод дешифровки получаемых совокупностей сигналов, обеспечивающий то, что вероятность ошибки при расшифровке каждого кодового обозначения не будет превосходить ε. Для этой цели удобно воспользоваться неравенством Чебышева, доказанным в § 4 гл. I. Воспользовавшись формулой(****) на стр. 58, легко показать, что если $N_2 = \sqrt{2N_1 p(1-p)}/\varepsilon$, то вероятность p_0 того, что число x ошибок при расшифровке N_1 последовательно переданных элементарных сигналов A_i не превзойдет $M = N_1 p + N_2$, будет удовлетворять неравенству

$$p_0 = p(x \leqslant N_1 p + N_2) \geqslant 1 - \varepsilon/2. \qquad (*)$$

---

[1]) Как уже отмечалось выше, идея этого доказательства принадлежит Файнстейну, рассмотревшему сразу общий случай определенной на стр. 332 произвольной линии связи. Применение соображений Файнстейна к простейшему частному случаю двоичной симметричной линии рассматривалось Э. Гилбертом [164] и Д. Слепяном [165]; еще один вариант упрощенного доказательства теоремы Шеннона для этого случая может быть найден в статье Г. Барнарда [166].

Заметим еще, что при фиксированных p и ε отношение

$$\frac{N_2}{N_1} = \sqrt{\frac{2p(1-p)}{\varepsilon}} \cdot \frac{1}{\sqrt{N_1}}$$

может быть сделано столь угодно малым, если только N_1 будет выбрано достаточно большим. Поэтому $M = N_1 p + N_2 = N_1 (p + N_2/N_1)$ может быть сделано сколь угодно близким к $N_1 p$. В частности, при $p < 1/2$ и N_1 достаточно большом $M = N_1 p + N_2$ будет меньше, чем $N_1/2$; в дальнейшем N_1 будет считаться таким большим, чтобы последнее условие было выполненным.

Выберем теперь первое кодовое обозначение (которое, для краткости мы обозначим символом A_1) произвольным образом среди 2^{N_1} различных цепочек $A_{i_1} A_{i_2} \ldots A_{i_{N_1}}$. Будем считать обозначение A_1 переданным, если на приемном конце линии будет принято сообщение, отличающееся от N_1-членной цепочки A_1 не более чем в M элементарных сигналах. Совокупность всевозможных N_1-членных цепочек, отличающихся от цепочки A_1 не более чем в M сигналах, мы обозначим символом $R(A_1)$. Таким образом, принятая N_1-членная цепочка будет расшифровываться как цепочка A_1, если она принадлежит совокупности $R(A_1)$; вероятность ошибки при расшифровке кодового обозначения A_1 в силу (*) будет тогда заведомо не превосходить $\varepsilon/2$.

Далее перейдем к выбору второго кодового обозначения A_2. Уговоримся прежде всего считать, что передавалось это обозначение A_2, если на приемном конце линии будет принята N_1-членная цепочка, которая

а) отличается от A_2 не более чем в M элементарных сигналах;

б) не принадлежит совокупности $R(A_1)$.

Нас интересуют только такие кодовые обозначения A_2, вероятность ошибки при расшифровке которых на приемном конце линии связи не превосходит ε. Ясно, что так наверное будет обстоять дело, если *при передаче цепочки A_2 вероятность получения какой-либо из цепочек совокупности $R(A_1)$ будет меньше, чем $\varepsilon/2$*. В тех случаях, когда вовсе не существует N_1-членных цепочек, удовлетворяющих этому последнему условию, мы будем считать, что $K = 1$; если же N_1-членные цепочки, ему удовлетворяющие, существуют, мы примем за A_2 произвольную из них.

Аналогично мы поступим и при выборе третьего кодового обозначения A_3. А именно, если не существует таких N_1-членных цепочек передаваемых сигналов, что *вероятность получения вместо них на приемном конце линии одной из цепочек, принадлежащих или совокупности $R(A_1)$, или же совокупности $R(A_2)$, меньше, чем $\varepsilon/2$*, то мы будем считать, что $K = 2$; в противном случае в качестве третьего кодового обозначения A_3 мы выберем любую из цепочек, удовлетворяющих указанному условию. Аналогично этому, после того, как первые k кодовых обозначений $A_1, A_2, \ldots, A_k$ будут уже выбраны, в качестве $(k+1)$-го кодового обозначения мы выберем произвольную N_1-членную цепочку A_{k+1} такую, что *в случае ее передачи по линии связи вероятность получения на приемном конце одной из цепочек, принадлежащей или $R(A_1)$, или $R(A_2)$, ..., или $R(A_k)$, меньше, чем* $\frac{\varepsilon}{2}$. Выбор всех кодовых обозначений

мы будем считать законченным тогда, когда окажется, что ни одной новой цепочки, удовлетворяющей сформулированному здесь условию, уже выбрать нельзя. При расшифровке принятых сообщений на приемном конце линии связи мы будем считать, что передавалось i-е обозначение A_i, если будет принята цепочка, которая

а′) отличается от A_i не больше чем в M сигналах;

б′) не принадлежит ни совокупности $R(A_1)$, ни $R(A_2)$, ..., ..., ни $R(A_{i-1})$.

Если же будет принята цепочка, которая отличается от в с е х имеющихся кодовых обозначений $A_1, A_2, \ldots, A_K$ больше чем в M сигналах, то ее мы будем расшифровывать произвольно (например, условимся во всех таких случаях считать, что передавалось обозначение A_1). Ясно, что используемое правило расшифровки принимаемых N_1-членных цепочек сигналов гарантирует, что при передаче любого из обозначений $A_1, A_2, \ldots, A_K$ мы правильно расшифруем его на приемном конце с вероятностью, превосходящей $1 - \varepsilon$. Таким образом, нам остается только убедиться, что число K таких обозначений при достаточно большом N_1 будет достаточно большим (а именно, может быть сделано сколь угодно близким к 2^{cN_1}).

Переходя к оценке числа K, начнем с того, что оценим число L_0 цепочек, входящих в совокупность $R(A)$ (где A — произвольная N_1-членная цепочка). Ясно, что совокупность $R(A)$ включает:

0) одну цепочку A;

1) $C^1_{N_1} = N_1$ различных цепочек, отличающихся от A одним сигналом;

2) $C^2_{N_1}$ различных цепочек, отличающихся от A двумя сигналами;

. . . . . .

M) $C^M_{N_1}$ различных цепочек, отличающихся от A какими-то $M = N_1 p + N_2$ сигналами.

Поэтому

$$L_0 = 1 + C^1_{N_1} + C^2_{N_1} + \cdots + C^M_{N_1}.$$

Число слагаемых в правой части последнего равенства можно оценить числом $M = N_1 p + N_2 < N_1/2$ (ибо слагаемое 1 в начале не может повлиять на оценку весьма большого числа L_0); кроме того, известно, что в ряду биномиальных коэффициентов

$$C^0_{N_1} = 1, C^1_{N_1}, C^2_{N_1}, C^3_{N_1}, \ldots, C^{N_1-1}_{N_1}, C^{N_1}_{N_1} = 1$$

члены монотонно возрастают вплоть до середины этого ряда. Поэтому, так как $M < N_1/2$, наибольшим из коэффициентов $C^1_{N_1}, \ldots, C^M_{N_1}$ будет п о с л е д н и й коэффициент; следовательно, можно утверждать, что

$$L_0 < M \cdot C^M_{N_1} < \frac{N_1}{2} \cdot C^M_{N_1}.$$

Воспользовавшись еще неравенством (**) на стр. 221 и учтя, что $N_1 - M = N_1(1-p) - N_2 = N_1 q - N_2$, где $q = 1-p$, получаем

$$L_0 < \frac{N_1}{2} \frac{N_1^{N_1}}{(N_1 p + N_2)^{N_1 p + N_2} (N_1 q - N_2)^{N_1 q - N_2}} = \frac{N_1}{2} \frac{1}{\left(p + \frac{N_2}{N_1}\right)^{N_1 p + N_2} \left(q - \frac{N_2}{N_1}\right)^{N_1 q - N_2}}. \quad (**)$$

Далее нам еще понадобиться оценка числа L_i всевозможных N_1-членных последовательностей принимаемых сигналов, входящих хоть в одну из совокупностей $R(A_1)$, $R(A_2)$, ..., $R(A_K)$. Будем рассуждать следующим образом. Рассмотрим процесс передачи 2^{N_1} всевозможных N_1-членных последовательностей передаваемых сигналов $A_1, A_2, \ldots, A_{2^{N_1}}$, при котором каждая из этих последовательностей передается с одинаковой вероятностью $1/2^{N_1}$[1]). В таком случае вероятность того, что передана будет последовательность, принадлежащая хоть одной из совокупностей $R(A_1)$, $R(A_2)$, ..., $R(A_K)$ очевидно будет равна $L_i/2^{N_1}$ (см. определение вероятности, выделенное курсивом на стр. 21). На приемном конце нашей двоичной линии связи при передаче последовательностей $A_1, A_2, \ldots, A_{2^{N_1}}$ из N_1 сигналов A_1 и A_2 будут приниматься также N_1-членные последовательности тех же сигналов; обозначим через $p(A_i A_j)$ вероятность того, что при передаче последовательности A_i принята будет последовательность A_j. Условимся теперь так нумеровать N_1-членные цепочки, чтобы цепочкам, входящим хоть в одну из совокупностей $R(A_1)$, $R(A_2)$, ..., $R(A_K)$, отвечали первые L_i номеров (т. е. будем считать входящими хоть в одну из этих совокупностей цепочки $A_1, A_2, \ldots, A_{L_i}$, где, разумеется, L_1 много больше, чем K). В таком случае событие, состоящее в том, что передана одна из первых L_1 цепочек A_l, можно будет представить в виде суммы следующих несовместимых событий: передана одна из цепочек A_i, где i пробегает значения $1, 2, \ldots, L_1$, а принята одна из цепочек A_j, где j пробегает все значения $1, 2, \ldots, 2^{N_1}$ (т. е. A_j пробегает всевозможные

---

[1]) Рассмотрение такого процесса передачи играет в данном доказательстве роль, родственную роли процедуры случайного кодирования в доказательстве Шеннона (см. выше, стр. 375). Напомним, что для двоичной симметричной линии связи пропускная способность реализуется для вероятностей $p^0(A_1) = p^0(A_2) = 1/2$; поэтому последовательная передача сигналов A_i независимо от всех предыдущих сигналов и каждый раз с вероятностями p^0 как раз и отвечает передаче всех N_1-членных цепочек с одинаковой вероятностью $1/2^{N_1}$.

N_1-членные цепочки). Таким образом,

$$\frac{L_1}{2^{N_1}} = p(A_1A_1) + p(A_1A_2) + \ldots + p(A_1A_{2^{N_1}}) +$$
$$+ p(A_2A_1) + p(A_2A_2) + \ldots + p(A_2A_{2^{N_1}}) +$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$+ p(A_{L_1}A_1) + p(A_{L_1}A_2) + \ldots + p(A_{L_1}A_{2^{N_1}}).$$

Заметим теперь, что вероятность $p(A_iA_j)$ определяется лишь тем, сколько из сигналов цепочки A_i совпадают с соответствующими сигналами цепочки A_j, а сколько — не совпадают (т. е. числом ошибок при передаче, переводящей цепочку A_i в A_j). Поэтому ясно, что $p(A_iA_j) = p(A_jA_i)$, и, следовательно,

$$\frac{L_1}{2^{N_1}} = p(A_1A_1) + p(A_2A_1) + \cdots + p(A_{2^{N_1}}A_1) +$$
$$+ p(A_1A_2) + p(A_2A_2) + \cdots + p(A_{2^{N_1}}A_2) +$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$+ p(A_1A_{L_1}) + p(A_2A_{L_1}) + \cdots + p(A_{2^{N_1}}A_{L_1}).$$

Ясно также, что сумму слагаемых, входящих в j-й столбец в правой части последнего равенства (т. е. стоящих друг под другом на j-м месте в каждой строке), можно переписать в виде

$$p(A_jA_1) + p(A_jA_2) + \cdots + p(A_jA_{L_1}) =$$
$$= p(A_j)[p(A_1/A_j) + p(A_2/A_j) + \cdots + p(A_{L_1}/A_j)] =$$
$$= \frac{1}{2^{N_1}}\, p(A_1 + A_2 + \cdots + A_{L_1}/A_j),$$

где $p(A_j) = \frac{1}{2^{N_1}}$ — вероятность передачи цепочки A_j, $p(A_i/A_j)$ — условная вероятность приема цепочки A_i при условии, что передавалось цепочка A_j, а $p(A_1 + A_2 + \ldots + A_{L_1}/A_j)$ — условная вероятность приема одной из первых L_1 цепочек при том же условии. Но легко понять, что вероятность приема одной из первых L_1 цепочек при передаче любой N_1-членной цепочки A_j не может быть меньше чем $\frac{\varepsilon}{2}$. В самом деле, если передаваемая цепочка — это одно из выбранных нами K кодовых обозначений $A_1, A_2, \ldots, A_K$, то вероятность приема цепочки совокупности $R(A_j)$ уже больше, чем $1 - \frac{\varepsilon}{2}$, а значит и больше, чем малое число $\frac{\varepsilon}{2}$. Если же для какой-либо из прочих N_1-членных передаваемых цепочек вероятность приема цепочки, принадлежащей хоть одной из совокупностей $R(A_1)$, $R(A_2)$, ..., $R(A_K)$, окажется меньше, чем $\frac{\varepsilon}{2}$, то в таком случае эту цепочку можно будет выбрать в качестве еще одного кодового обозначения, т. е. мы придем в противоречие

с предположением о том, что больше чем K кодовых обозначений выбрать нельзя.

Таким образом, в правую часть многострочного равенства в середине предыдущей страницы входит 2^{N_1} столбцов, сумма членов каждого из которых не меньше, чем $\frac{1}{2^{N_1}} \cdot \frac{\varepsilon}{2}$; поэтому окончательно

$$\frac{L_1}{2^{N_1}} \geqslant 2^{N_1}\left(\frac{1}{2^{N_1}} \cdot \frac{\varepsilon}{2}\right) = \frac{\varepsilon}{2}, \text{ т. е. } L_1 \geqslant \frac{\varepsilon}{2} 2^{N_1}. \qquad (***)$$

Теперь уже совсем легко получить результат, который мы хотим доказать. В самом деле, L_1 цепочек принадлежат K различным (вообще говоря, пересекающимся между собой) совокупностям $R(A_1)$, $R(A_2)$, ..., $R(A_K)$, каждая из которых содержит L_0 цепочек. Следовательно,

$$K \geqslant \frac{L_1}{L_0}.$$

Воспользовавшись оценками (**) и (***) чисел L_0 и L_1, найдем, что

$$K > \frac{\varepsilon}{N_1} 2^{N_1}\left(p + \frac{N_2}{N_1}\right)^{N_1\left(p+\frac{N_2}{N_1}\right)}\left(q - \frac{N_2}{N_1}\right)^{N_1\left(q-\frac{N_2}{N_1}\right)}.$$

При достаточно большом N_1 отношение $\frac{N_2}{N_1}$ будет сколь угодно малым; отсюда вытекает, что

$$\frac{\log K}{N_1} > 1 + \left(p + \frac{N_2}{N_1}\right)\log\left(p + \frac{N_2}{N_1}\right) + \left(q - \frac{N_2}{N_1}\right)\log\left(q - \frac{N_2}{N_1}\right) - \frac{\log N_1}{N_1} + \frac{\log \varepsilon}{N_1}$$

при достаточно большом N_1 будет больше числа, сколь угодно близкого к $c = 1 + p \log p + q \log q$. Но ведь больше, чем 2^{cN_1}, число K не может быть (см. выше, стр. 348 и 362); отсюда видно, что при достаточно большом N_1 число $\frac{\log K}{N_1}$ может быть сделано сколь угодно близким к c. Как уже отмечалось выше, отсюда сразу вытекает справедливость теоремы Шеннона для двоичной симметричной линии связи.

В заключение приведем строгое доказательство выписанного на стр. 368 неравенства Фано (A′): ведь приведенное на стр. 368—369 рассуждение частично опирается на интуитивные представления об информации и потому, строго говоря, не может считаться доказательством. Такое доказательство легко получить, придав точный смысл всем использованным ранее соображениям. Основным для нас являлось то, что *степень неопределенности опыта* β *с* n *исходами* $A_1, A_2, \ldots, A_n$, *имеющими вероятности*

$\pi_1, \pi_2, \ldots, \pi_n$, *равна степени неопределенности опыта* γ, *состоящего в проверке того, имел ли или не имел опыт* β *исход* A_n, *сложенной с умноженной на* $\pi_1 + \pi_2 + \ldots + \pi_{n-1} = 1 - \pi_n$ *степенью неопределенности опыта* γ_1 *с* $n - 1$ *исходами, представляющего собой тот же опыт* β, *но уже при дополнительном условии, что исход* A_n *не имел места.* Но если мы обозначим, как обычно, через $H(\pi_1, \pi_2, \ldots, \pi_n)$ величину

$$-\pi_1 \log \pi_1 - \pi_2 \log \pi_2 - \ldots - \pi_n \log \pi_n,$$

равную степени неопределенности (энтропии) опыта с n исходами и вероятностями $\pi_1, \pi_2, \ldots, \pi_n$ этих исходов, то высказанное утверждение формально будет эквивалентно соотношению

$$H(\pi_1, \pi_2, \ldots, \pi_n) = H(\pi_n, 1 - \pi_n) + (1 - \pi_n) H\left(\frac{\pi_1}{1 - \pi_n}, \frac{\pi_2}{1 - \pi_n}, \ldots, \frac{\pi_{n-1}}{1 - \pi_n}\right).$$

В справедливости последнего соотношения очень просто убедиться с помощью непосредственной проверки. Заметим еще, что мы уже использовали на стр. 132 даже несколько более общее соотношение для $H(\pi_1, \pi_2, \ldots, \pi_n)$ (записанное сверху на этой странице), смысл которого мы тогда разъясняли точно так же, как сейчас.

Предположим теперь, что нам известен исход a_1, или a_2, ..., или a_n опыта α, состоящего в расшифровке одной буквы текста на приемном конце линии связи. Тогда выписанное соотношение можно будет применить к степени неопределенности $H_{a_1}(\beta)$, или $H_{a_2}(\beta)$, ..., или $H_{a_n}(\beta)$ опыта β (состоящего в определении одной буквы передаваемого текста) при известном исходе α. При этом мы будем считать, что исход A_n с вероятностью π_n — это во всех случаях тот исход β, который совпадает с известным исходом α. Так как $H\left(\frac{\pi_1}{1 - \pi_n}, \frac{\pi_2}{1 - \pi_n}, \ldots, \frac{\pi_{n-1}}{1 - \pi_n}\right)$ — это энтропия опыта с $n - 1$ исходами, которая при любых значениях $\pi_1, \pi_2, \ldots, \pi_{n-1}, \pi_n$ не больше чем $\log(n - 1)$, то мы получим

$$H_{a_1}(\beta) \leqslant h(q_1') + q_1' \log(n - 1),$$
$$H_{a_2}(\beta) \leqslant h(q_2') + q_2' \log(n - 1),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$H_{a_n}(\beta) \leqslant h(q_n') + q_n' \log(n - 1),$$

где $h(q) = H(q, 1 - q) = -q \log q - (1 - q) \log (1 - q)$, а $q_1', q_2', \ldots, q_n'$ имеют тот же смысл, что и на стр. 364. Умножим теперь первое из этих неравенств на p_1', второе — на p_2', ..., последнее — на p_n' и сложим отдельно левые и правые части. Так как $h(q)$ — выпуклая функция q при $0 \leqslant q \leqslant 1$, то в силу теоремы 4 Приложения I (стр. 449)

$$p_1' h(q_1') + p_2' h(q_2') + \cdots + p_n' h(q_n') \leqslant h(p_1' q_1' + p_2' q_2' + \cdots + p_n' q_n') = h(q).$$

Поэтому полученный при сложении результат может быть переписан в виде

$$H_\alpha(\beta) \leqslant h(q) + q \log(n-1)$$

— а это и есть то неравенство Фано, которое мы стремились доказать.

§ 5. Коды, обнаруживающие и исправляющие ошибки

Основным результатом предыдущего § 4 бесспорно является теорема Шеннона о кодировании при наличии помех. Согласно этой теореме *для любой заданной линии связи с пропускной способностью* $C = Lc$ *и заданной скорости передачи*

$$v_1 = L\frac{c_1}{H} < L\frac{c}{H} \quad \textit{букв/ед. времени}$$

наверное существует способ выбора кодовых обозначений (представляющих собой «блоки», т. е. длинные цепочки элементарных сигналов), *позволяющий осуществить передачу сообщений со скоростью* v_1 *так, чтобы вероятность ошибки при расшифровке каждой буквы передаваемого сообщения была бы меньше произвольного* (но заранее заданного) *числа* ε. На стр. 352—353 отмечалось также, что теорему Шеннона можно сформулировать и следующим образом: *если* $c_1 < c$, *то* $2^{c_1 N}$ *кодовых обозначений длины* N *при достаточно большом* N *всегда можно выбрать так, чтобы вероятность ошибки при расшифровке полученной на приемном конце линии связи цепочки из* N *элементарных сигналов была бы меньше произвольного* (заранее заданного) *числа* ε *независимо от того, какое именно кодовое обозначение передавалось на самом деле* [1]). Последняя формулировка основной теоремы удобна тем, что она относится только к линии связи, но никак не связана с природой и статистическими свойствами исходных буквенных сообщений; ею мы, в основном, и будем ниже пользоваться.

---

[1]) В § 4 длину кодовых обозначений мы обычно обозначали через N_1, так как буква N там использовалась для обозначения длины кодируемых «блоков» исходного буквенного сообщения. Однако в настоящем параграфе исходные сообщения вообще не будут рассматриваться; поэтому здесь нам будет удобнее считать, что длина кодовых обозначений равна N.

Теорема Шеннона о кодировании, при всей ее простоте и неожиданности, обладает одним очень существенным с практической точки зрения недостатком: она является типичной «теоремой существования» и не содержит никаких указаний на то, как именно следует выбирать кодовые обозначения какой-то приемлемой длины N для того, чтобы обеспечить достаточно малую вероятность ошибки при заданной достаточно высокой (т. е. достаточно близкой к $v = L\frac{c}{H}$) скорости передачи. Вопрос об отыскании практически удобных методов выбора кодовых обозначений для различных линий связи с помехами составляет содержание теории кодирования, развившейся после появления основной работы Шеннона [1] в обширную (и крайне важную для приложений) самостоятельную науку, отличающуюся громадным разнообразием используемых в ней подходов и методов, зачастую заимствованных из казавшихся самыми абстрактными и далекими от запросов практики разделов современной математики [1]). Изложению этой науки только на русском языке посвящено не менее нескольких десятков оригинальных и переводных монографий и сборников статей, из которых мы здесь упомянем лишь широко известные и очень содержательные (но довольно сложные) книги [168] и [169]; ей же посвящены обширные разделы во многих общих курсах теории информации (см., например, [6], [13], [21], [22]) и многочисленные обзорные статьи (например, [165], [167], [170], [171]). В нашей книге,

---

[1]) С этим обстоятельством связано название интересной популярной статьи [167] американского математика Н. Левинсона: «Теория кодирования: противоречащий пример к принадлежащей Г. Х. Харди концепции прикладной математики». Дело в том, что известный английский математик Г. Х. Харди в написанной им в 1940 г. (и затем многократно переиздававшейся) книге «В защиту математика» (G. H. Hardy «A mathematician's apology») разделил математику на «чистую» (или «истинную»), доставляющую громадное эстетическое наслаждение уму своей стройностью, логической законченностью и изяществом, но бесполезную в практической жизни, и «прикладную», нужную для практики, но скучноватую и не содержащую элементов неожиданности. При этом некоторые из наиболее типичных с точки зрения Харди разделов «чистой математики» (например, теория чисел или теория полей Галуа) впоследствии оказались как раз теми, которые играют центральную роль в (безусловно прикладной) теории кодирования!

разумеется, совершенно невозможно даже вкратце охватить хотя бы одни лишь основы современной теории кодирования; однако некоторые относительно простые выводы, относящиеся к этой теории, все же могут быть здесь рассмотрены.

Начнем с небольшого разъяснения, полезного для понимания самой постановки задачи в теории кодирования. Принято утверждать, что все существующие доказательства основной теоремы Шеннона совершенно н е э ф ф е к т и в н ы, т. е. даже в принципе не могут быть использованы для нахождения метода, позволяющего выбрать кодовые обозначения (и метода соответствующей расшифровки принимаемых цепочек элементарных сигналов), обеспечивающих малость вероятности ошибки при заданной скорости передачи. На самом деле, однако, такое утверждение все же нельзя считать вполне справедливым.

Действительно, вспомним, например, намеченный на стр. 374—383 метод доказательства теоремы Шеннона с использованием «случайного кодирования». В ходе этого доказательства предлагалось выбрать $2^{\gamma N}$ кодовых обозначений длины N наудачу (из числа некоторых заранее отобранных $2^{H(\beta)N}$ «вероятных» цепочек длины N) и затем доказывалось, что в таком случае существует метод расшифровки, при котором среднее значение вероятности ошибки при расшифровке будет достаточно малым (меньшим, чем $\frac{\varepsilon}{2}$). Далее мы воспользовались тем, что всегда хоть одно из значений случайной величины будет не превосходить ее среднего значения — для доказательства теоремы этого нам было вполне достаточно. Но ведь можно пойти в том же направлении и немного дальше: ясно, что если среднее значение неотрицательной случайной величины очень мало, то сравнительно малыми должны быть не одно, а п о ч т и в с е ее значения. Последнее обстоятельство находит свое математическое выражение в доказанном на стр. 55 неравенстве Чебышева (**), согласно которому для любой неотрицательной случайной величины α

$$P(\alpha > c) < \frac{a}{c}, \text{ где } a = \text{ср.зн.}\alpha.$$

Поэтому если $a = \text{ср.зн.}\alpha$ настолько мало, что и Ma остается еще малым, где M — какое-то сравнительно большое число, то значение α будет не превосходить малой величины Ma с весьма большой вероятностью (большей, чем $1-1/M$). Исходя из подобных соображений можно доказать, что если мы воспользуемся случайным кодированием (и описанным на стр. 377 методом расшифровки), то при достаточно большом N вероятность ошибки при расшифровке (а не только ее значение при каком-то одном неизвестном нам выборе $2^{c_1 N}$ кодовых обозначений) будет с очень большой вероятностью (т. е. «почти наверняка») очень малой. Тем самым мы сразу получаем как будто бы очень простой метод выбора кодовых обозначений, приводящий практически всегда к малой вероятности ошибки — надо лишь принять N достаточно большим, а затем выбрать $2^{c_1 N}$ кодовых обозначений длины N наудачу (с помощью описанного на стр. 375 опыта с извлечением бумажек с номерами из урны)[1]).

Но как можно реально воспользоваться этим «простым» методом? По-видимому, для получения хороших результатов здесь обычно придется требовать, чтобы N имело, по крайней мере, порядок многих десятков или даже сотен, а если принять, что $N = 100$, а $c_1 = 0{,}5$, то нам надо будет выбрать наудачу $2^{50} \approx 10^{15}$ различных последовательностей из 100 элементарных сигналов и все их надо будет запомнить. Однако это еще самая легкая часть задачи — несравненно большие трудности представляет расшифровка получаемых на приемном конце линии цепочек элементарных сигналов. Согласно сказанному на стр. 355 и след. для такой расшифровки мы должны перебрать все 2^{50} групп $\mathcal{B}$, отвечающих нашим кодовым обозначениям, чтобы выяснить, к каким из них принадлежит принятая цепочка сигналов, а к каким нет — эта задача представляется совершенно нереальной для всех существующих

---

[1]) Слова «практически всегда» здесь означают, что выбранный код может оказаться плохим лишь в крайне маловероятном случае, при «исключительном невезении». Но если N достаточно велико, то этой возможностью можно пренебречь; кроме того, даже и в случае такой неудачи дело можно поправить: убедившись (на примере пробной передачи), что выбранный код плох, можно от него просто отказаться и выбрать кодовые обозначения заново при помощи того же метода.

(и даже для всех могущих появиться в близком будущем) вычислительных машин.

Мы видим таким образом, что основной сложностью в теории кодирования является совсем не то, что вообще невозможно указать метод кодирования (т. е. выбора 2^{cLN} кодовых обозначений длины N) и декодирования (т. е. соответствующей расшифровки принимаемых цепочек из N сигналов), обеспечивающий высокую скорость передачи и, одновременно, малую вероятность ошибки. Наиболее существенно здесь требование, чтобы и кодирование, и — что особенно трудно — декодирование можно было бы сравнительно просто осуществить на практике. Удовлетворить последнему требованию очень нелегко — это как раз и породило громадное число исследований, посвященных разработке тех или иных практически приемлемых методов кодирования и декодирования, хотя и не являющихся о п т и м а л ь н ы м и (т. е. самыми лучшими из всех возможных), но все же достаточно х о р о ш и х (т. е. позволяющих добиться относительно больших скоростей передачи при не слишком больших вероятностях ошибки).

Ограничимся для простоты лишь д в о и ч н ы м и линиями связи, т. е. будем считать, что по линии можно передавать только два элементарных сигнала (скажем, посылку тока и паузу) и что эти же два сигнала могут быть приняты на приемном конце линии. Будем обозначать используемые сигналы цифрами 0 и 1; в таком случае все кодовые обозначения будут последовательностями этих цифр, т. е. *числами, записанными в двоичной системе счисления*. Кодовые обозначения длины N здесь можно выбирать из числа 2^N различных N-значных двоичных чисел — последовательностей $a_0a_1 \ldots a_{N-1}$, где все a_i, $i = 0,1, \ldots, N-1$, принимают значения 0 или 1; набор всех используемых кодовых обозначений мы и будем теперь называть к о д о м. Если в с е 2^N различных N-значных чисел мы примем за кодовые обозначения, то скорость передачи информации будет наибольшей (а именно, равной L бит/ед. времени или, что то же самое, $\frac{L}{H}$ букв/ед. времени), — но зато при этом у нас не будет никакой возможности определить на приемном конце линии связи, имелись ли ошибки при передаче, сколько

их было и какие именно сигналы приняты неправильно. Если, однако, мы ограничимся меньшим числом кодовых обозначений, то возникающая при этом «избыточность кода» может быть использована для дополнительной передачи некоторых сведений об искажениях, внесенных линией связи. Так, например, мы можем воспользоваться простейшим методом N-кратного повторения каждого элементарного сигнала (т. е. использовать в качестве кода лишь два простейших кодовых обозначения $00\ldots0$ и $11\ldots1$ длины N), а на приемном конце линии расшифровывать принятую цепочку длины N как $00\ldots0$, если она содержит больше нулей, чем единиц, и как $11\ldots1$ в противном случае. Ясно, что такой метод передачи при достаточно большом N (и при естественном условии, что вероятность искажения передаваемого элементарного сигнала в процессе его передачи меньше, чем 1/2) обеспечивает очень малую вероятность ошибки при расшифровке переданного сообщения, но зато здесь и скорость передачи также будет крайне мала (за время N/L, нужное для передачи N элементарных сигналов, здесь будет передаваться лишь 1 бит информации, что соответствует скорости передачи, равной $\frac{L}{N}$ бит/ед. времени $= \frac{L}{HN}$ букв/ед. времени). Естественно, что такая низкая скорость передачи во многих случаях нас не будет устраивать; поэтому наибольший интерес представляют промежуточные между рассмотренными классы кодов, обеспечивающие приличную скорость передачи и одновременно позволяющие исправить многие искажения в передаваемых сообщениях.

Сравнительно общий прием использования избыточности в кодовых обозначениях для передачи информации об искажениях может быть проиллюстрирован уже на простейшем случае, когда число кодовых обозначений длины N равно 2^{N-1} (т. е. равно половине числа различных цепочек из N двоичных сигналов). Условимся сопоставлять 2^{N-1} кодовых обозначений всевозможным цепочкам $a_0\,a_1\ldots a_{N-2}$ из $N-1$ цифр 0 и 1, а N-ю цифру a_{N-1} будем каждый раз выбирать так, чтобы сумма $a_0 + a_1 + \ldots + a_{N-1}$ была *четной*. В таком случае наличие о д н о ч н о й ошибки (т. е. ошибки в одном из принятых N элементарных сигналов) приведет к появлению на приемном конце линии связи такой цепочки $a_0'a_1'\ldots a_{N-1}'$,

что сумма $a_0' + a_1' + \ldots + a_{N-1}'$ является нечетной (так как единственно возможные искажения заключаются в том, что 0 принимается за 1, или 1 за 0). Это обстоятельство позволяет легко обнаружить наличие одиночной ошибки, хотя и не позволяет выяснить, какой именно сигнал был принят неверно (точнее говоря, нечетность суммы $a_0' + a_1' + \ldots + a_{N-1}'$ указывает, что заведомо имелось нечетное число ошибок, в то время как четное число ошибок при этом не будет обнаружено). Тем не менее и тех случаях, когда вероятность появления более одной ошибки при передаче N сигналов очень мала, описанный здесь очень простой метод кодирования иногда представляет значительную ценность — ведь если мы наверное знаем, что прием сопровождался ошибкой, то можно просто игнорировать полученное сообщение или, если это допустимо, попросить повторить передачу. С другой стороны, скорость передачи при таком методе кодирования все еще остается очень большой — с максимального значения L бит/ед. времени она убывает всего лишь до $\frac{N-1}{N} L$ бит/ед. времени $= \frac{N-1}{N} \frac{L}{H}$ букв/ед. времени.

Описанный выше прием «проверки на четность» можно применить также несколько раз — и это уже позволяет во многих случаях не только обнаружить наличие ошибок, но и исправить их. Рассмотрим, например, случай, когда $N = 3$, а число используемых кодовых обозначений равно двум. Мы знаем, что в таком случае в качестве кодовых обозначений разумно выбрать тройки 000 и 111; такой выбор с точки зрения использования «проверок на четность» можно обосновать следующим образом. Сопоставим два кодовых обозначения двум возможным значениям первого элементарного сигнала a_0 (т. е. будем считать, что только сигнал a_0 реально содержит информацию), а далее условимся вслед за каждым «информационным сигналом» a_0 передавать еще два «контрольных сигнала» a_1 и a_2, подобранных так, чтобы суммы $a_0 + a_1$ и $a_0 + a_2$ обе были четными (реально это как раз и сведется к выбору в качестве кодовых обозначений цепочек 000 и 111). В таком случае легко видеть, что если только при приеме тройки сигналов не произошло сразу двух или трех ошибок (т. е. если считать возможными лишь правильную передачу и передачу с одиночными

ошибками), то, проверив четность сумм $a_0' + a_1'$ и $a_0' + a_2'$ в принятой на приемном конце тройке $a_0' a_1' a_2'$, можно будет безошибочно установить, какая же именно тройка была на самом деле передана. В самом деле, если обе суммы $a_0' + a_1'$ и $a_0' + a_2'$ окажутся четными, то отсюда сразу будет следовать, что ошибок при передаче не было (напомним, что что возможность двойной ошибки мы исключаем); если нечетной будет лишь одна из них, то это будет значить, что ошибочно принят входящий в эту сумму контрольный сигнал a_1 или a_2, а если обе суммы $a_0' + a_1'$ и $a_0' + a_2'$ — нечетные, то это значит, что неверно принят информационный сигнал a_0. Таким образом, ценой уменьшения скорости передачи втрое (по сравнению с максимальной скоростью L бит/ед. времени) мы можем добиться того, чтобы все о д и н о ч н ы е ошибки в тройках элементарных сигналов были исправлены.

Приведенные выше результаты, разумеется, очевидны (ясно, что, приняв за кодовые обозначения тройки 000 и 111, мы можем добиться исправления всех одиночных ошибок), но они могут быть обобщены и на случай многих бóльших значений N. Так, например, если $N = 7$, а число кодовых обозначений равно $16 = 2^4$, то мы можем принять за «информационные сигналы» первые четыре сигнала a_0, a_1, a_2 и a_3 (так как число различных четверок $a_0a_1a_2a_3$ как раз равно шестнадцати), а последние три «контрольных сигнала» a_4, a_5 и a_6 подобрать так, чтобы были четными суммы

$$s_1 = a_0 + a_1 + a_2 + a_4, \quad s_2 = a_0 + a_1 + a_3 + a_5$$
$$\text{и} \quad s_3 = a_0 + a_2 + a_3 + a_6.$$

При этом «проверка на четность» трех сумм s_1, s_2 и s_3 на приемном конце линии также позволяет однозначно установить, была ли допущена ошибка при приеме (при условии, что возможностью двух и более ошибок при приеме семи сигналов мы пренебрегаем) и если была, то в чем она заключалась. В самом деле, если один из 7 сигналов будет принят неправильно, то хоть одна из сумм наверное окажется нечетной, так что четность трех сумм определенно указывает на отсутствие одиночных ошибок при передаче; далее лишь одна сумма будет нечетной в том (и только том) случае, когда ошибочно принят входящий в эту сумму один из трех «контрольных сигналов»

(a_4, a_5 или a_6); наконец, нечетность двух из трех сумм s_1, s_2 и s_3 будет означать, что неверно принят тот из трех сигналов a_1, a_2 и a_3, который входит в обе эти суммы, а нечетность всех трех сумм — что неверно принят входящий во все суммы первый сигнал a_0. Легко видеть, что 16 кодовых обозначений длины 7 в данном случае имеют вид

0000000, 1000111, 0100110, 1100001,
0010101, 1010100, 0110011, 1110100,
0001011, 1001100, 0101101, 1101010,
0011110, 1011001, 0111000, 1111111;

использование этих кодовых обозначений обеспечивает скорость передачи, равную

$$\frac{4L}{7} \text{ бит/ед. времени} = \frac{4L}{7H} \text{ букв/ед. времени},$$

и одновременно позволяет исправить все о д и н о ч н ы е ошибки (но не ошибки большей кратности!) в «блоках» из семи элементарных сигналов.

Соответствующий код, конечно, не является «самым лучшим», но так как и кодирование и декодирование здесь осуществляются без большого труда, то он вполне может оказаться практически полезным. Рассмотрим, например, для конкретности, двоичную симметричную линию связи, в которой вероятность ошибки при приеме каждого из двух используемых элементарных сигналов равна 0,01 (так что неправильно принимается примерно одна сотая доля всех передаваемых элементарных сигналов). Пропускная способность такой линии связи равна

$$C = 0{,}92L \text{ бит/ед. времени}$$

(см. стр. 338); значит, здесь существует код, позволяющий передавать в единицу времени $0{,}92L$ бит информации и такой, что вероятность ошибки при декодировании меньше произвольного числа ε (которое можно выбрать сколь угодно малым). Однако как построить такой код мы не знаем; к тому же, если взять ε очень малым, то он, вероятно, потребует использования крайне длинных кодовых обозначений и будет очень сложным. Воспользуемся теперь описанным выше очень простым кодом с $N = 7$, в котором к каждым четырем передаваемым сигналам добавляется еще три контрольных сигнала. При этом мы

будем передавать информацию со скоростью

$$\frac{4}{7}L \approx 0{,}57L \text{ бит/ед. времени},$$

заметно меньшей предельной скорости безошибочной передачи; кроме того, вероятность ошибки при декодировании здесь, разумеется, не будет «сколь угодно малой», а будет равна вероятности того, что из семи переданных элементарных сигналов приняты с ошибкой два или больше. Исходя отсюда можно подсчитать, что при таком методе передачи в последовательности «информационных элементарных сигналов», восстановленной на приемном конце линии связи, ошибочные сигналы будут составлять несколько меньше одной тысячной части, так что вероятность ошибки при приеме одного элементарного сигнала здесь будет немного меньше чем 0,001. Мы видим, что вероятность ошибки при приеме одного элементарного сигнала в этом случае уменьшается более чем в 10 раз по сравнению с передачей без использования «контрольных сигналов»; так как и кодирование, и декодирование здесь весьма просты и могут быть очень легко автоматизированы, то с точки зрения практики использование рассматриваемого кода безусловно заслуживает внимания.

Заметим еще, что описанные здесь примеры «кодов с исправлением одной ошибки» довольно тесно связаны с содержанием разобранной на стр. 145 задачи, в которой предполагалось, что среди заданных n чисел загадано или одно число, или ни одного, и требовалось с помощью наименьшего числа вопросов (на которые отвечается только «да» или «нет») выяснить, было ли загадано число, и если да, то какое именно. Нам теперь будет удобнее вместо n чисел рассмотреть N номеров 0, 1, ..., $N-1$, входящих в кодовое обозначение $a_0a_1\ldots a_{N-1}$; такая замена, разумеется, ничего не меняет в рассуждениях. Согласно сказанному на стр. 145 для требуемого выяснения здесь надо затратить не меньше чем $\log(N+1)$ и не больше чем $\log(N+1)+1$ вопросов; но ведь наши «проверки на четность» фактически эквивалентны некоторым вопросам (поскольку каждая проверка может дать два результата: «*четное*» или «*нечетное*», подобно тому, как ответом на вопрос могло быть «да» или «нет»). В гл. III ответы на вопросы давали нам определенную информацию

о загаданном числе, так как исходили от человека, которому это число было известно; для того чтобы результат «проверки на четность» содержал информацию о возможных искажениях при передаче, надо, чтобы заранее было известно, четна или нечетна сумма передаваемых сигналов. Так как, вообще говоря, мы не можем знать, какие сигналы будут передаваться, то последнее условие может быть удовлетворено лишь в том случае, когда каждая передаваемая сумма содержит по крайней мере один «контрольный сигнал», относительно которого заранее договорено, что он подбирается так, чтобы соответствующая сумма оказалась, например, четной. Отсюда ясно, что число требуемых добавочных «контрольных сигналов» совпадает с минимальным числом необходимых «проверок на четность», т. е. равно числу тех вопросов, о которых шла речь на стр. 145. Если, например, $N = 3$, то число вопросов не может быть меньше чем $\log (3 + 1) = \log 4 = 2$; это как раз и соответствует тому, что в описанном на стр. 398—399 примере кода, исправляющего одиночные ошибки, каждый передаваемый «информационный сигнал» a_0 пришлось дополнять *двумя* добавочными «контрольными сигналами» a_1 и a_2. Заметим еще, что поскольку сигналы a_1 и a_2 подбирались так, чтобы суммы $a_0 + a_1$ и $a_0 + a_2$ были четными, проверка четности соответствующих сумм на приемном конце линии связи равносильна ответам на вопросы: «Не содержат ли ошибок сигналы a_0 и a_1?» и «Не содержат ли ошибок сигналы a_0 и a_2?»; ясно, что такие ответы позволяют однозначно определить любую одиночную ошибку. Аналогично этому, если $N = 7$, то число требуемых вопросов (т. е. «проверок на четность» и «контрольных сигналов») не может быть меньше, чем $\log (7 + 1) = \log 8 = 3$; это мы и видели на стр. 399—400. Описанная там проверка четностей сумм s_1, s_2 и s_3 равносильна ответам на вопросы: «Не содержат ли ошибок сигналы a_0, a_1, a_2 и a_4?», «Не содержат ли ошибок сигналы a_0, a_1, a_3 и a_5?» и «Не содержат ли ошибок сигналы a_0, a_2, a_3 и a_6?»; ясно, что ответы на эти вопросы также однозначно определяют искаженный сигнал.

В общем случае кодовых обозначений длины N число K «контрольных сигналов» кода, исправляющего все одиночные ошибки, должно, согласно сказанному выше,

удовлетворять неравенству
$\log(N+1) \leqslant K < \log(N+1)+1$, так что $2^{K-1}-1 < N \leqslant 2^K-1$;
число же «информационных сигналов» здесь равно $N-K$. Код, использующий кодовые обозначения длины N, состоящие из $M=N-K$ «информационных сигналов» и K не несущих информации «контрольных сигналов», используемых для «проверок на четность», мы будем называть (N, M)-кодом; отвечающая ему скорость передачи информации, очевидно, равна $L\frac{M}{N}$ бит/ед. времени. В рассматриваемом нами случае $K < \log(N+1)+1$, так что K при большом N будет гораздо меньше, чем N; поэтому скорость передачи при большом N здесь будет очень близка к максимальной скорости L бит/ед. времени. Отсюда ясно, что рассматриваемые коды при большом N будут обеспечивать очень высокую скорость передачи. Разумеется, очень большое N выбирать все же невыгодно, так как при этом сильно увеличивается вероятность наличия *нескольких* (больше одной) ошибок в блоке из N сигналов, т. е. понижается надежность кода; на практике приходится прибегать к компромиссу и выбирать какое-то промежуточное (не слишком большое, но и не слишком малое) значение N. Метод выбора «контрольных сигналов» для общего (N, M)-кода, где $M=N-K$, исправляющего все одиночные ошибки, также может быть установлен, исходя из аналогии с задачей об отгадывании задуманного числа и намеченного на стр. 145 решения этой последней задачи; мы здесь на этом не будем останавливаться, так как ниже будет указан совсем другой метод построения требуемого кода. Заметим еще, что рассмотренный на стр. 399—400 случай (7, 4)-кода, исправляющего одиночные ошибки, был рассмотрен в качестве примера еще в статье *Шеннона* [1]; общие (N, M)-коды, исправляющие одиночные ошибки, были рассмотрены в 1950 г. Р. *Хеммингом* (см. [172]) и с тех пор обычно называются *кодами Хемминга*[1]).

[1]) Впрочем, довольно часто *кодами Хэмминга* называют лишь такие исправляющие одиночные ошибки (N, M)-коды, в которых $N=2^K-1$ (т. е. является наибольшим возможным при данном числе K «контрольных сигналов»). Эти коды обладают

Аналогичным образом можно подойти и к проблеме построения кодов, позволяющих исправлять о д и у ш л и д в е ошибки. Предположим, например, что $N = 5$, причем мы пренебрегаем возможностью одновременного искажения больше чем двух сигналов из пяти, но требуем, чтобы код позволял исправить все искажения в случаях, когда их число не превосходит двух. Эта ситуация приводит нас к задаче об определении $n \leqslant 2$ загаданных чисел среди каких-то пяти чисел. В силу сказанного на стр. 145 для определения этих чисел требуется задать не менее

$$\log\,(C_5^2 + C_5^1 + 1) = \log\,(10 + 5 + 1) = \log 16 = 4$$

вопросов; поэтому здесь нам потребуются, по крайней мере, четыре проверки на четность и, значит, из каждых пяти сигналов a_0, a_1, a_2, a_3 и a_4, по крайней мере четыре должны быть «контрольными». Нетрудно видеть, что в данном случае четырех контрольных сигналов действительно достаточно для решения задачи, причем эти сигналы a_1, a_2, a_3 и a_4 можно, например, подобрать из условия, чтобы были четными суммы

$$s_1 = a_0 + a_1,\ s_2 = a_0 + a_2,\ s_3 = a_0 + a_3 \text{ и } s_4 = a_0 + a_4.$$

В таком случае четность всех рассматриваемых сумм на приемном конце линии будет означать отсутствие ошибок; нечетность одной суммы s_i — ошибку в соответствующем сигнале a_i; нечетность двух сумм s_i и s_j — ошибку в сигналах a_i и a_j; нечетность трех сумм (скажем, всех кроме s_i) — ошибку в сигналах a_0 и a_i; нечетность всех четырех сумм — единственную ошибку в сигнале a_0 [1]).

замечательным свойством, о котором еще будет сказано в конце настоящего параграфа (см. стр. 436—438). Любопытно, что такие $(2^K - 1, 2^K - K - 1)$-коды еще в 1942 г. (т. е. до появления и работы Хэмминга, и даже работы Шеннона) в совсем другом контексте (формально не связанном с теорией кодирования, но фактически ей эквивалентном) были рассмотрены известным английским статистиком Р. А. Фишером (см. Э. Берлекэмп [169], стр. 18 и 22).

[1]) Легко понять, что описанные «проверки на четность» равносильны ответам на вопросы: «*будет ли четным число ошибок при приеме сигналов* a_0 *и* a_1?»; «*при приеме сигналов* a_0 *и* a_2?»; «*при при-*

В общем случае кодов, исправляющих одну или две ошибки в «блоках» из произвольного числа N сигналов, приведенные на стр. 145 результаты точно так же показывают, что *число K «контрольных сигналов» и отвечающих им «проверок на четность» не может быть меньшим, чем*

$$\log (C_N^2 + C_N^1 + 1) = \log \frac{N^2 + N + 1}{2}.$$

Однако на вопрос о том, как именно здесь надо подбирать «контрольные сигналы» (т. е. какие «проверки на четность» наиболее быстро ведут к цели), в этом случае ответить совсем не легко и решение соответствующей задачи об отгадывании чисел еще не содержит общего метода эффективного построения соответствующего «кода, исправляющего ошибки». Аналогично этому и в еще более общем случае кодов, позволяющих обнаружить и исправить в цепочке сигналов длины N любое число ошибок, не превосходящее заданного n, приведенные на стр. 145 рассуждения позволяют утверждать, что нужное для этой цели *число K «контрольных сигналов»* (*и* отвечающих им *«проверок на четность»*) *не может быть меньшим, чем* $\log (C_N^n + C_N^{n-1} + \ldots + 1)$. Этот простой результат был указан Р. Хэммингом [172], и поэтому соответствующее неравенство для числа K часто называется неравенством Хэмминга или нижней границей Хэмминга для числа «контрольных сигналов» кода, исправляющего n ошибок. Если $n = 1$, то неравенство Хэмминга приводит к уже известному нам результату: $N \leqslant 2^K - 1$; равенство здесь достигается для кодов Хэмминга с $N = 2^K - 1$. Но и в общем случае приведенные на стр. 145—146 рассуждения не указывают, как именно следует выбирать нужные нам «проверки на четность» (т. е. как можно построить код с нужными свойствами); более того, они не позволяют

*еме сигналов a_0 и a_3?», и, наконец, «при приеме сигналов a_0 и a_1?». При этом ответ на первый вопрос выделяет из 16 различных возможных «исходов» передачи, при которых искажаются не более двух элементарных сигналов, группу из 8 допустимых исходов, т. е. содержит наибольшую возможную информацию; также и все последующие вопросы выделяют ровно половину из числа оставшихся до этого возможными «исходов».

даже утверждать, что для любого K, удовлетворяющего неравенству Хэмминга, действительно существует «код с проверками на четность», содержащий K контрольных сигналов и позволяющий исправить любое меньшее чем n число ошибок в «блоке» из N сигналов (на самом деле для некоторых K, удовлетворяющих этому неравенству, нужного нам кода построить нельзя). Оценка числа K «контрольных сигналов», заведомо достаточного для возможности обнаружить и исправить любое меньшее n число ошибок в блоках из N сигналов, была из совсем других соображений получена Р. Р. Варшамовым [173], показавшим, что *при* $K > \log (C_{N-1}^{2n-1} + C_{N-1}^{2n-2} + \ldots + 1)$ *всегда можно построить «код с проверками на четность», обладающий нужными нам свойствами.* Этот результат Варшамова (уточняющий предшествующие более грубые результаты Э. Гилберта [164]) называется неравенством Варшамова — Гилберта или верхней границей Варшамова — Гилберта для числа K контрольных сигналов кода, исправляющего n ошибок; его простое доказательство будет приведено ниже (см. стр. 421). Если $n > 1$, то верхняя граница Варшамова — Гилберта, вообще говоря, оказывается превосходящей нижнюю границу Хэмминга; таким образом, здесь существуют значения числа «контрольных сигналов» K, для которых соответствующие неравенства не исключают возможности построения кода, исправляющего n ошибок, но и не позволяют утверждать, что такой код обязательно существует. Кроме того, все доказательства неравенства Варшамова — Гилберта хотя и опираются на определенный метод построения нужных кодов, но не претендуют на то, чтобы метод этот можно было удобно применить на практике; в результате используемые при доказательстве построения оказываются совершенно неприемлемыми для реального использования (все они опираются на непосредственный перебор колоссального числа возможностей).

Даже для простейшего случая $n = 2$ реальный метод построения «кодов с проверками на четность», позволяющих исправлять любые одиночные или двойные ошибки в блоках из произвольного числа N сигналов, был найден лишь примерно через 10 лет после появления работы Хэмминга [172], в которой описывались общие

коды, исправляющие одиночные ошибки — в 1960 г. Р. Боузом и Д. Чоудхури (см. [174]) и в 1959 г. А. Хоквингемом [175], причем используемые для этой цели средства оказались удивительным образом опирающимися на тонкий и довольно сложный математический аппарат, относящийся к абстрактной алгебре. Дальнейшее обобщение того же метода, позволяющее строить коды, исправляющие любое число ошибок, меньшее заданного числа n, оказалось уже сравнительно простым и было найдено практически одновременно с нахождением кодов, исправляющих не более двух ошибок.

Для того чтобы дать представление о методе построения кодов, исправляющих не только одиночные, но и двойные (или вообще кратные не выше заданной кратности) ошибки по результатам проверок на четность, следует прежде всего строго определить само понятие «кодов с проверками на четность». С этой целью удобно начать с того, что рассмотреть все арифметические действия с числами 0 и 1 как действия, могущие иметь лишь два возможных результата: 0, символизирующий то, что в результате действия получилось *четное* число, и 1, означающий, что получилось число *нечетное*. В результате мы придем к следующей таблице, содержащей результаты всевозможных арифметических действий, производимых над числами 0 и 1:

$$0+0=0,\ 0+1=1,\ 1+0=1,\ 1+1=0;$$

$$0\cdot 0=0,\ 0\cdot 1=0,\ 1\cdot 0=0,\ 1\cdot 1=1.$$

Легко видеть, что полученные таким образом операции «сложения» и «умножения» (которые мы будем называть *сложением* и *умножением в 2-арифметике*) [1]), удовлетворяют всем обычным законам арифметики; это обстоятельство выражают, говоря, что совокупность двух чисел 0 и 1, для которых определены принятые в 2-арифметике действия сложения и умножения, образует *поле* из двух элементов (точное определение поля, знание которого, впрочем, не является строго необходимым для

[1]) Собственно говоря, существующее в 2-арифметике «умножение» можно было бы писать без всяких кавычек, так как оно не отличается от обычного; напротив, «сложение» в 2-арифметике отличается об обычного, ибо здесь $1+1=0$.

понимания всего дальнейшего, приведено в Приложении II на стр. 463—464) [1]).

Теперь мы можем легко описать общий (N, M)-код с проверками на четность. Он задается при помощи $K = N - M$ равенств вида

$$\begin{aligned} a_M &= b_{M,0}a_0 + b_{M,1}a_1 + \cdots + b_{M,M-1}a_{M-1}, \\ a_{M+1} &= b_{M+1,0}a_0 + b_{M+1,1}a_1 + \cdots + b_{M+1,M-1}a_{M-1}, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{N-1} &= b_{N-1,0}a_0 + b_{N-1,1}a_1 + \cdots + b_{N-1,M-1}a_{M-1}. \end{aligned} \quad (1)$$

Здесь все коэффициенты $b_{M,0}$, $b_{M,1}$, . . ., $b_{M,M-1}$, ... , $b_{N-1,0}$, $b_{N-1,1}$, . . ., $b_{N-1,M-1}$ — это элементы нашего поля из двух элементов (т. е. числа 0 или 1), а все входящие в эти равенства арифметические действия понимаются в смысле 2-арифметики (так что каждое равенство означает лишь, что его левая и правая части, понимаемые в обычном смысле, имеют одинаковую четность). K проверок на четность, отвечающие нашему (N, M)-коду — это проверки четности суммы контрольного сигнала a_i (где i принимает $K = N - M$ значений $M, M + 1, \ldots, M + K - 1 = N - 1$) и тех из информационных сигналов $a_0, a_1, \ldots, a_{M-1}$, которым отвечают равные единице (а не нулю!) коэффициенты $b_{i,0}, b_{i,1}, \ldots, b_{i,M-1}$ [2]). Для

[1]) То обстоятельство, что совокупность различных элементарных сигналов можно считать совокупностью всевозможных элементов некоторого конечного поля, является очень важным для всей современной алгебраической теории кодирования. Однако в алгебре доказывается, что *поле с заданным числом m различных элементов существует лишь тогда, когда m является степенью простого числа* (т. е. равно p^k, где p — простое; ср. ниже Приложение II, стр. 471). Поэтому алгебраическая теория кодирования может быть непосредственно применена к недвоичным линиям связи (которые мы здесь, впрочем, вовсе не будем рассматривать) лишь в случае, когда число m различных элементарных сигналов, которые можно передавать по линии, имеет вид p^k. Если же это не так, то приходится дополнительно прибегать к каким-то искусственным приемам (например, вовсе не использовать некоторые из допустимых сигналов).

[2]) Напомним, что в 2-арифметике $1 + 1 = 0$, и значит, $-1 = 1$. Поэтому при переносе слагаемого с одной стороны равенства в другую здесь не обязательно изменять его знак, а равенство $x = y$ можно переписать и как $x - y = 0$, и как $x + y = 0$ (оба выписанные соотношения равносильны друг другу — они означают лишь, что x и y имеют одинаковую четность).

задания кода достаточно указать все коэффициенты $b_{i,j}$, входящие в выписанные равенства. При этом удобно сперва перенести в этих равенствах все левые части a_M, a_{M+1} ... ..., a_{N-1} направо (учитывая правило, указанное в сноске [2]) на стр. 408), а затем записать все коэффициенты в получившихся равенствах в виде таблицы из $K = N - M$ строк и N столбцов, на пересечении i-й строки и j-го столбца которой стоит коэффициент при a_j в i-м из наших равенств. Легко видеть, что такая таблица будет иметь вид

$$\begin{pmatrix} b_{M,0} & b_{M,1} & \dots & b_{M,M-1} & 1 & 0 & \dots & 0 \\ b_{M+1,0} & b_{M+1,1} & \dots & b_{M+1,M-1} & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ b_{N-1,0} & b_{N-1,1} & \dots & b_{N-1,M-1} & 0 & 0 & \dots & 1 \end{pmatrix}. \qquad (2)$$

Прямоугольная таблица из m строк и n столбцов в математике называется матрицей из m строк и n столбцов или, короче, $(m \times n)$-матрицей; таким образом, общий (N,M)-код с проверками на четность задается $(K \times N)$-матрицей из нулей и единиц специального вида (2). Совокупность всевозможных кодовых обозначений такого общего (N, M)-кода с проверками на четность может быть легко описана следующим образом: информационные сигналы $a_0, a_1, \dots, a_{M-1}$ здесь могут быть любыми (т. е. каждый из них может независимо от других принимать и значение 0, и значение 1), а контрольные сигналы $a_M, a_{M+1}, \dots, a_{N-1}$ уже однозначно определяются по информационным сигналам с помощью равенств (1), понимаемых в смысле 2-арифметики. Общее число различных кодовых обозначений в этом случае, очевидно, равно $2^M = 2^{N-K}$.

Заметим еще, что иногда код с проверками на четность определяют и несколько более широко как совокупность таких N-членных цепочек $a_0, a_1, \dots, a_{N-1}$ символов 0 и 1, что числа $a_0, a_1, \dots, a_{N-1}$ удовлетворяют K соотношениям вида

$$\begin{aligned} &b_{M,0}\,a_0 + b_{M,1}\,a_1 + \dots + b_{M,N-1}\,a_{N-1} = 0, \\ &b_{M+1,0}\,a_0 + b_{M+1,1}\,a_1 + \dots + b_{M+1,N-1}\,a_{N-1} = 0, \\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ &b_{N-1,0}\,a_0 + b_{N-1,1}\,a_1 + \dots + b_{N-1,N-1}\,a_{N-1} = 0 \end{aligned} \qquad (1')$$

(где коэффициенты снова принимают лишь значения 0 и 1, а равенства понимают в смысле 2-арифметики). Отвечающая наиболее общему коду (1′) матрица будет уже произвольной $(K \times N)$-матрицей, состоящей из нулей и единиц. Имея в виду это более широкое определение, более частные коды, задаваемые равенствами вида (1) и матрицей вида (2), называют **систематическими кодами с проверками на четность**. Нетрудно показать, однако, что произвольный код с проверками на четность всегда может быть записан как систематический код, с числом «контрольных сигналов», не превосходящим числа K соотношений (1′) (см. Приложение II, стр. 482). Поэтому, как правило, в дальнейшем мы будем говорить только о систематических кодах.

В литературе по теории кодирования коды с проверками на четность часто называют также **линейными кодами** или **групповыми кодами**. Оба последних термина связаны с дополнительными свойствами рассматриваемых кодов, представляющими интерес сами по себе и весьма важными, если желать перенести теорию таких кодов на более общие недвоичные линии связи (для которых понятие проверки на четность, очевидно, не имеет прямого смысла). Для того чтобы объяснить, в чем состоят эти свойства, следует ввести в рассмотрение операции сложения и умножения на число z (принадлежащее нашему полю из двух элементов, т. е. равное или нулю, или единице) **блоков** $\boldsymbol{a} = (a_0, a_1, \ldots, a_{N-1})$ из N нулей и единиц. Эти операции могут быть естественно определены следующим образом:

$$(a_0, a_1, \ldots, a_{N-1}) + (a_0', a_1', \ldots, a_{N-1}') = \\ = (a_0 + a_0', a_1 + a_1', \ldots, a_{N-1} + a_{N-1}'),$$

$$z(a_0, a_1, \ldots, a_{N-1}) = (za_0, za_1, \ldots, za_{N-1}).$$

Заметим попутно, что так как здесь все арифметические действия понимаются в смысле 2-арифметики, то операция умножения блока на число не особенно интересна: для **любого** блока $(a_0, a_1, \ldots, a_{N-1})$

$$0 \cdot (a_0, a_1, \ldots, a_{N-1}) = (0, 0, \ldots, 0) \text{ и}$$
$$1 \cdot (a_0, a_1, \ldots, a_{N-1}) = (a_0, a_1, \ldots, a_{N-1}).$$

Нетрудно проверить, что так определенные операции сложения и умножения на число удовлетворяют всем основным законам, которым удовлетворяют обычные арифметические действия; на языке современной алгебры последнее обстоятельство выражают, говоря, что совокупность всевозможных последовательностей из N нулей и единиц $\boldsymbol{a} = (a_0, a_1, \ldots, a_{N-1})$ образует *векторное пространство* (точное определение векторного пространства, которое нам в дальнейшем не будет непосредственно нужно, можно найти в Приложении II). С другой стороны, то, что операция сложения последовательностей сама по себе (т. е. вне связи с умножением на числа) обладает большинством обычных свойств арифметических операций сложения и умножения, можно выразить, сказав, что совокупность последовательностей $\boldsymbol{a} = (a_0, a_1, \ldots, a_{N-1})$ представляет собой *группу* относительно введенной выше операции сложения (определение группы приведено на стр. 458—459 Приложения II; для понимания всего дальнейшего и оно не необходимо). Код (т. е. определенная совокупность кодовых обозначений, каждое из которых является «блоком»— цепочкой из N цифр 0 и 1) называется л и н е й н ы м, если его кодовые обозначения представляют собой *линейное подпространство* общего векторного пространства таких «блоков»— это означает, что сумма любых двух кодовых обозначений линейного кода, а также произведение кодового обозначения на число z должны быть кодовыми обозначениями [1]). Код называется г р у п п о в ы м, если его кодовые обозначения представляют собой *подгруппу* общей группы последовательностей $(a_0, a_1, \ldots, a_{N-1})$— в рассматриваемом нами здесь двоичном случае это снова означает лишь то, что сумма любых двух кодовых обозначений и «нулевой блок» (0, 0, . . ., 0) должны быть кодовыми обозначениями (смысл сделанного здесь утверждения в применении к случаю наличия более чем двух различных сигналов будет объяснен на стр. 463). Мы видим, таким образом,

---

[1]) Ясно, что в рассматриваемом нами случае наличия лишь двух сигналов условие, относящееся к умножению на число z, не очень содержательно: оно означает лишь, что последовательность (0,0, ..., 0) из N нулей должна являться кодовым обозначением. Однако в случае большего чем 2 числа элементарных сигналов указанное условие оказывается уже достаточно важным.

что в случае двоичной линии (т. е. при использовании лишь двух элементарных сигналов) термины линейный код и групповой код означают точно одно и то же [1]).

Рассмотрим теперь произвольный (не обязательно систематический) код с проверками на четность, кодовые обозначения которого совпадают с совокупностью цепочек $\boldsymbol{a} = (a_0, a_1, \ldots, a_{N-1})$ таких, что для них выполняются равенства (1′). Прежде всего ясно, что если $(a_0, a_1, \ldots, a_{N-1})$ — это блок $(0, 0, \ldots, 0)$ из одних нулей, то равенства (1′) обязательно выполняются — поэтому нулевой блок $(0, 0, \ldots, 0)$ обязательно является кодовым обозначением нашего кода. Кроме того, если блоки $\boldsymbol{a} = (a_0, a_1, \ldots, a_{N-1})$ и $\boldsymbol{a}' = (a'_0, a'_1, \ldots, a'_{N-1})$ оба являются кодовыми обозначениями (т. е. для них обоих выполняются все K соотношений (1′)), то, сложив друг с другом первые, вторые и т. д. вплоть до последних из этих соотношений для $\boldsymbol{a}$ и для $\boldsymbol{a}'$, мы убедимся, что

$$\boldsymbol{a} + \boldsymbol{a}' = (a_0 + a'_0, a_1 + a'_1, \ldots, a_{N-1} + a'_{N-1})$$

также удовлетворяет всем соотношениям (1′), т. е. также является кодовым обозначением. Отсюда вытекает, что любой код с проверками на четность является одновременно также и линейным (или групповым) кодом. С другой стороны, в алгебре доказывается, что любое линейное подпространство векторного пространства цепочек $\boldsymbol{a} = (a_0, a_1, \ldots, a_{N-1})$ может быть задано некоторым набором соотношений вида (1′) (см. Приложение II, стр. 476). Следовательно, класс линейных (или групповых) кодов для двоичной линии связи точно совпадает с классом кодов с проверкой на четность — именно это обстоятельство и дает основание называть коды с проверками на четность также линейными кодами или групповыми кодами.

Продолжим рассмотрение общих кодов с проверками на четность; поскольку, как мы уже отмечали выше,

---

[1]) В более общем случае линий связи с m элементарными сигналами эти два понятия совпадают друг с другом, если $m = p$ есть простое число, но понятие линейного кода является лишь частным случаем понятия группового кода, если $m = p^k$, где p — простое, а $k > 1$ (ср. сноску [1]) на стр. 408). Наконец, если m не равно целой степени некоторого простого числа, то ни то, ни другое понятия вообще не могут быть определены.

любой такой код может быть представлен в виде систематического кода (удовлетворяющего равенствам вида (1)), то в основном мы будем здесь говорить о кодах этого последнего вида. Такой код задается матрицей (2), называемой проверочной матрицей кода [1]); нам будет удобно обозначить ее одной буквой B. Если $a = (a_0, a_1, \ldots, a_{N-1})$ — это одно из кодовых обозначений нашего кода, то справедливость для него соотношений (1) удобно символически изображать в виде равенства

$$Ba = 0 \qquad (3)$$

(левая часть здесь служит записью $N - M$ левых частей равенств вида (1′), получаемых из (1) при перенесении всех левых частей вправо; здесь Ba есть произведение матрицы B на вектор a, понимаемое в смысле теории матриц, о котором сказано в Приложении II на стр. 480). Предположим, что по линии связи передавалось кодовое обозначение $a = (a_0, a_1, \ldots, a_{N-1})$; в результате искажений в процессе передачи на приемном конце, вообще говоря, будет принята цепочка $a' = (a'_0, a'_1, \ldots, a'_{N-1})$, отличная от той, которая передавалась. Подставим цепочку a' в левые части равенств (1′) (понимаемые, как обычно, в смысле 2-арифметики); получаемые в результате $K = N - M$ чисел 0 и 1 (представляющие собой K-членную цепочку $(s_M, s_{M+1}, \ldots, s_{N-1})$) мы будем обозначать символом Ba'. Поскольку a', вообще говоря, уже не является кодовым обозначением, цепочка $Ba' = s = (s_M, s_{M+1}, \ldots, s_{N-1})$ уже не будет нулевой (т. е. на некоторых местах она будет содержать и единицы). Наличие этих единиц, очевидно, показывает, что при передаче имели место искажения; на языке, которым мы пользовались раньше, каждая единица означает, что соответствующая «проверка на четность» привела к отрицательному результату. Пусть $e = (e_1, e_2, \ldots, e_N) = (a'_1 - a_1, a'_2 - a_2, \ldots, a'_{N-1} - a_{N-1})$ — это N-членный «блок ошибок», содержащий единицы на местах, соответствующих сигналам a_i, искаженным при передаче, и нули на всех

[1]) В случае общих (не систематических) кодов с проверками на четность проверочной матрицей, очевидно, будет произвольная $(K \times N)$-матрица из нулей и единиц (некоторые примеры таких общих проверочных матриц нам еще встретятся в дальнейшем).

остальных местах, так что

$$e = a' - a = a' + a$$

(напомним, что в 2-арифметике $a - b = a + b$). Ясно, что в силу (3)

$$Be = B(a' - a) = Ba';$$

следовательно,

$$Be = s. \tag{4}$$

К сожалению, вообще говоря, существует много цепочек $e = (e_0, e_1, \ldots, e_{N-1})$, удовлетворяющих $N - M$ равенствам (4); поэтому, исходя отсюда, нельзя еще однозначно восстановить «блок ошибок» e (а, значит, и переданную цепочку $a = a' - e = a' + e$). При декодировании кодов с проверками на четность обычно предполагается, что вероятность искажения при передаче каждого сигнала меньше вероятности правильной передачи и в соответствии с этим принимается следующее правило декодирования: *в качестве блока ошибок e принимается та из удовлетворяющих равенствам (4) цепочек, которая содержит наименьшее число единиц*, т. е. отвечает наименьшему возможному числу искажений при передаче (если среди цепочек, удовлетворяющих (4), имеется несколько, содержащих одно и то же наименьшее число единиц, то e выбирается наудачу среди них). Это правило позволяет расшифровать все принимаемые на приемном конце линии N-членные цепочки элементарных сигналов, т. е. сопоставить всем им определенные кодовые обозначения $a = a' + e$ (очевидно удовлетворяющие необходимому для кодовых обозначений условию (3)), которые и считаются переданными по линии связи.

Описанный метод декодирования кодов с проверками на четность заметно проще общего метода, описанного на стр. 377 (и опирающегося на рассмотрение групп $\mathfrak{B}$, отвечающих различным кодовым словам). Тем не менее и он не является практически пригодным: при больших значениях $K = N - M$ нахождение той из удовлетворяющих (4) цепочек, которая содержит наименьшее число единиц, оказывается настолько громоздким, что даже современные вычислительные машины не позволяют вы-

полнить его за приемлемое время. Поэтому очень важной представляется задача создания достаточно простых (т. е. реально осуществимых) методов нахождения нужного нам блока e; она пока что может считаться решенной лишь для некоторых частных случаев кодов с весьма специальной структурой проверочной матрицы B [1]). Однако даже и без этого существование указанного выше теоретически достаточно простого общего правила декодирования может быть использовано для изучения свойств произвольных кодов с проверками на четность. Такое изучение было начато Д. Слепяном [177], а П. Элайсом [159] было показано, что в случае двоичной симметричной линии связи (а также и в случае двоичной линии со стиранием, соответствующей изображенной на рис. 21 схеме со значением $p = 0$) коды с проверками на четность не уступают наилучшим из всех вообще возможных кодов в том смысле, что здесь *с помощью кодов с проверками на четность всегда можно осуществить такую передачу информации с заданной скоростью* $C_1 = Lc_1$ *бит/ед.времени, меньшей пропускной способности* $C = Lc$ *линии связи, чтобы вероятность ошибки при декодировании была меньше любого наперед заданного числа* $\varepsilon > 0$. При этом величина вероятности ошибки, достижимая при фиксированной скорости передачи $C_1 = Lc_1$ бит/ед.времени, где $c_1 < c$, и кодовых обозначениях фиксированной длины N, будет не больше чем a_1^{-N}, где a_1 — зависящее от c_1 число, большее единицы; таким образом, с ростом N вероятность ошибки здесь убывает по тому же закону, что и в случае наилучших произвольных кодов. Кроме того, Элайс также доказал, что если выбирать код с проверками на четность «наудачу» (т. е. при выборе каждого элемента $b_{i,j}$ проверочной матрицы B подбрасывать монету и полагать, что $b_{i,j} = 0$ в случае выпадения герба, но $b_{i,j} = 1$ в случае выпадения цифры), то и тогда для рассматриваемых линий связи вероятность ошибки при декодировании при $N \to \infty$

[1]) Один из таких частных случаев, специально изученный Р. Галлагером [176], касается матриц B с большими значениями N и $K = N - M$, состоящих, грубо говоря, почти из одних нулей (т. е. содержащих лишь очень небольшое число единиц). Некоторые другие частные случаи, описываемые алгебраически, будут указаны ниже.

(и $K = (1 - c_1) N$, так что $2^{N-K} = 2^{c_1 N}$) будет стремиться к нулю (и притом не медленнее, чем N-я степень некоторого меньшего единицы числа) [1]).

То обстоятельство, что для многих реально встречающихся линий связи выбранный «наудачу» код с проверками на четность при большом N оказывается «почти наверное» достаточно хорошим, делает весьма соблазнительным использование таких «случайных кодов с проверками на четность». Для того чтобы задать такой код, надо случайным образом выбрать (и запомнить) $MK = N^2 c_1 (1 - c_1)$ элементов b_{ij} (где $i = M, M + 1, \ldots, N - 1$, а $j = 0, 1, \ldots, M - 1$) соответствующей проверочной матрицы B. Так как число $N^2 c_1 (1 - c_1)$ с ростом N возрастает не слишком быстро (несравненно медленнее, чем, например, число $2^{c_1 N}$), то с подобной задачей современные вычислительные машины вполне могут справиться даже при N, имеющем порядок многих сотен. Однако процедура декодирования (т. е. нахождения по принятой цепочке a' соответствующего «блока ошибок» e), как мы уже отмечали, представляет в случае произвольно выбранного кода с проверками на четность очень большие трудности, и это существенно затруднит использование «случайных кодов». Тем не менее существуют определенные перспективные подходы к практическому построению «хороших» методов кодирования и декодирования, включающие в качестве составного элемента выбор «наудачу» некоторых величин, задающих рассматриваемый код (в качестве примера можно указать на так называемое «последовательное декодирование», с которым можно познакомиться, в частности, по книге [22] или обзорной статье [170]). Поскольку подходы эти все же являются довольно сложными, мы здесь на них не будем задерживаться, а сразу

---

[1]) В дальнейшем Р. Л. Добрушин [178] (рассматривавший произвольные групповые коды) и Г. Дригас [179] (рассматривавший несколько более частные линейные коды) обобщили результаты Элайеса, относящиеся к двоичной симметричной линии связи, на случай более общих линий связи с $m = p^k$ элементарными сигналами и таких, что $r = m$ (т. е. принимаются те же сигналы, которые передаются), а соответствующие вероятности $p_{A_j}(A_i)$ удовлетворяют определенным условиям симметрии. Однако для произвольных линий связи все эти результаты оказываются уже неверными (см. [180], [181]).

перейдем к применению «неслучайных» кодов с проверками на четность для обнаружения и исправления ошибок при передаче.

Нам будет удобно обозначить отдельные столбцы проверочной матрицы B (представляющие собой «блоки» из $K = N - M$ цифр 0 и 1) через $b_0, b_1, \ldots, b_{M-1}, b_M, \ldots, b_{N-1}$ (в случае систематического кода последние K столбцов $b_M, \ldots, b_{N-1}$ будут, очевидно, все содержать по одной единице и $N - M - 1$ нулей). Саму матрицу B при этом можно записать в виде одной строки

$$B = (b_0, b_1, \ldots, b_{M-1}, b_M, \ldots, b_{N-1}).$$

Обозначим, как и выше, через $e = (e_0, e_1, \ldots, e_{N-1})$ «блок ошибок», содержащий единицы на местах тех элементарных сигналов передаваемого кодового обозначения, которые исказились при передаче. В таком случае основное равенство (4) можно будет переписать в виде

$$e_0 b_0 + e_1 b_1 + \cdots + e_{M-1} b_{M-1} + e_M b_M + \cdots + e_{N-1} b_{N-1} = s, \tag{5}$$

где сложение понимается, как почленное сложение (в смысле 2-арифметики) соответствующих «блоков» длины K. Таким образом, «блок» s, который получается при подстановке в левую часть равенства (1′) вместо переданных сигналов $a_0, a_1, \ldots, a_{N-1}$ принятых сигналов $a'_0, a'_1, \ldots, a'_{N-1}$ и на основании которого мы должны судить о имеющихся ошибках, равен сумме столбцов проверочной матрицы B, отвечающих сигналам, искаженным при передаче (т. е. отвечающих значениям $e_i = 1$; остальным сигналам отвечают значения $e_i = 0$, и поэтому соответствующие слагаемые $e_i b_i$ обращаются в 0). Отсюда, в частности, видно, что одиночным ошибкам (т. е. блокам e, содержащим одну единицу и $N - 1$ нулей) соответствуют блоки s, совпадающие со столбцами b_i проверочной матрицы B; отсутствию же ошибок отвечает блок $s = 0$ из одних нулей. Поэтому для того, чтобы код с проверками на четность позволил различить и случай отсутствия ошибок, и все случаи одиночных ошибок при передаче, надо, чтобы все столбцы соответствующей проверочной матрицы B были различными и не один из них не был нулевым.

Общее число возможных различных K-значных блоков $b = (b_M, b_{M+1}, \ldots, b_{N-1})$ (т. е. различных последовательностей из K нулей и единиц) равно числу целых чисел, записываемых в двоичной системе счисления при помощи не более чем K цифр, т. е. равно 2^K (подобно тому, как число различных не более чем K-значных чисел в десятичной системе счисления равно 10^K). Так как нулевой блок $(0, 0, \ldots, 0)$ при этом исключается из числа возможных столбцов матрицы B, то число различных столбцов оказывается равным $2^K - 1$. Таким образом мы снова приходим к выводу, что *код с проверками на четность, исправляющий все одиночные ошибки и содержащий K «контрольных сигналов», должен состоять из кодовых обозначений, длина которых не превосходит $2^K - 1$.* Для задания такого кода надо лишь указать соответствующую проверочную матрицу B, все столбцы которой должны быть ненулевыми и различными. Получаемые коды, естественно, совпадают с кодами Хэммипга, о которых говорилось на стр. 403. В случае $N = 2^K - 1$ удобно выписать соответствующую проверочную матрицу B, выбрав в качестве ее столбцов двоичную запись (т. е. запись в двоичной системе счисления) всех целых чисел от 1 и до $2^K - 1$, перечисленных в возрастающем порядке; получаемый при этом код, разумеется, фактически будет систематическим (так как он будет содержать все возможные столбцы из $K - 1$ нулей и одной единицы), но только «контрольными сигналами» здесь будут не последние K сигналов, а какие-то сигналы с другими номерами. Так, например, в случае $K = 4$, $N = 2^4 - 1 = 15$ соответствующую (4×15)-матрицу B удобно записать в виде

$$B = \begin{pmatrix} 0\,0\,0\,0\,0\,0\,0\,1\,1\,1\,1\,1\,1\,1\,1 \\ 0\,0\,0\,1\,1\,1\,1\,0\,0\,0\,0\,1\,1\,1\,1 \\ 0\,1\,1\,0\,0\,1\,1\,0\,0\,1\,1\,0\,0\,1\,1 \\ 1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1\,0\,1 \end{pmatrix}$$

(заметим, что, пожелав здесь выписать все кодовые обозначения подобно тому, как это было сделано на стр. 403 в применении к случаю $K = 3$, $N = 7$, мы были бы вынуждены написать $2^{11} = 2048$ пятнадцатизначных чисел!). При такой матрице B роль «контрольных сигналов» будут играть первый, второй, четвертый и восьмой сигналы (так

как именно им отвечают столбцы из трех нулей и одной единицы); остальные же 11 сигналов будут информационными. Блок s будет нулевым в случае отсутствия ошибок при передаче, а в случае одной ошибки он будет равен соответствующему столбцу B, т. е. будет непосредственно задавать двоичную запись номера того сигнала, который исказился при передаче. Отсюда видно, что процедура декодирования (т. е. расшифровки принятого сигнала — исправления в нем ошибок) осуществляется в этом случае крайне просто.

Коды, исправляющие одиночные ошибки в блоках из $N < 2^K - 1$ сигналов, легко получить, вычеркнув из соответствующей проверочной матрицы B некоторое число «лишних» столбцов (которые можно выбрать произвольно из числа тех, которые содержат не меньше чем 2 единицы). Заметим еще, что свойства кода Хэмминга можно еще улучшить, добавив к каждому кодовому обозначению дополнительный $(K+1)$-й «контрольный сигнал» a_N, позволяющий уже обнаружить (но не исправить) также и все двойные ошибки. Для этого надо только выбрать этот добавочный сигнал a_N так, чтобы он давал четное число в сумме со всеми остальными сигналами, т. е. удовлетворял соотношению

$$a_0 + a_1 + \ldots + a_{N-1} + a_N = 0$$

(нетрудно понять, что это соответствует добавлению к матрице B сперва добавочного последнего столбца из одних нулей, а затем еще и добавочной последней строки из $N+1$ единиц; в результате и число строк, и число столбцов B возрастает на единицу). В таком случае отсутствию ошибок при передаче снова будет отвечать блок s из одних нулей; в случае одной ошибки первые K цифр блока s будут представлять собой двоичную запись некоторого целого числа, заключающегося в пределах от 0 до $2^K - 1$, а последняя цифра s_{K+1} будет равна единице (так как сумма всех принятых сигналов здесь обязательно будет нечетной); наконец, наличие хоть одной единицы среди первых K элементов блока s и обращение в нуль его последнего элемента будут свидетельствовать о наличии двойной ошибки. Усовершенствованный таким образом код Хэмминга был также предложен в работе [172]; его иногда называют р а с ш и р е н н ы м к о д о м Х э м м и н г а.

Перейдем теперь к кодам, исправляющим не только все о д и н о ч н ы е, но и все д в о й н ы е ошибки в блоках из N сигналов. Ясно, что при *отсутствии* ошибок при передаче блок $\boldsymbol{s} = \boldsymbol{Ba'}$ из K элементов будет состоять из одних нулей; при наличии *одной* ошибки он будет равен соответствующему столбцу проверочной матрицы B; наконец, в случае *двух* ошибок он будет равен суммо двух соответствующих столбцов B (ср. равенство (5) на стр. 417). Для того чтобы все эти случаи можно было различить на приемном конце линии связи, все столбцы B должны быть ненулевыми, отличными друг от друга и такими, что сумма любых двух из них отличается и от всех столбцов, и от всех прочих их попарных сумм. Матрицу, удовлетворяющую всем этим условиям, можно, следуя Г. С а к с у [182], попытаться построить с помощью простого перебора. С этой целью мы можем первый столбец b_0 матрицы B выбрать произвольным образом (но так, чтобы он не состоял из одних нулей); затем принять за b_1 произвольный ненулевой блок из K цифр 0 и 1, отличный от b_0; затем за b_2 принять ненулевой блок, отличный от b_0, b_1 и $b_0 + b_1$; затем за b_3 принять какой-то ненулевой блок, отличный от b_0, b_1 и b_2, а также от парных сумм $b_0 + b_1$, $b_0 + b_2$ и $b_1 + b_2$ и от тройной суммы $b_0 + b_1 + b_2$ (ибо в 2-арифметике если $b_0 + b_1 + b_2 = b_3$, то $b_0 + b_1 = b_2 + b_3$, т. е. ошибки в первых двух сигналах кодового обозначения будут неотличимы от ошибок в третьем и четвертом сигналах) и т. д. Здесь после того, как мы выберем первые i столбцов $b_0, b_1, \ldots, b_{i-1}$, при выборе $(i+1)$-го столбца b_i необходимо потребовать, чтобы этот столбец

а) не был нулевым столбцом;

б) не равнялся ни одному из $i = C_i^1$ уже выбранных столбцов $b_0, b_1, \ldots, b_{i-1}$;

в) не равнялся ни одной из C_i^2 попарных сумм уже выбранных столбцов;

г) отличался от всех C_i^3 сумм троек уже выбранных столбцов.

Разумеется, перечисленные $1 + C_i^1 + C_i^2 + C_i^3$ условий а)—г), запрещающих те или иные выборы столбца b_i, не обязательно будут все различными между собой (так, например, при $i > 5$ вполне может оказаться, что $b_0 +$

$+ b_1 + b_2 = b_3 + b_4 + b_5$ или что $b_1 + b_2 + b_3 = b_4 + b_5$); однако так как число всех различных столбцов (т. е. блоков из K цифр 0 и 1) равно 2^K, то если только $1 + C_i^1 + C_i^2 + C_i^3 < 2^K$, то условиям а)—г) наверное можно удовлетворить даже в наименее благоприятном случае, когда все фигурирующие в этих условиях столбцы и их комбинации различны. Наиболее ограничительным выписанное соотношение будет в применении к последнему столбцу b_{N-1} (так как при возрастании номера i число исключенных комбинаций, с которыми не может совпасть новый столбец, также возрастает). Поэтому *если только*

$$2^K > 1 + C_{N-1}^1 + C_{N-1}^2 + C_{N-1}^3,$$

$$\text{т. е. } K > \log (1 + C_{N-1}^1 + C_{N-1}^2 + C_{N-1}^3),$$

то наверное можно подобрать проверочную $(K \times N)$-*матрицу* $\boldsymbol{B}$, *задающую код с проверками на четность, исправляющий все одиночные и все двойные ошибки в блоках из* N *элементарных сигналов.*

Полученное здесь неравенство — это неравенство Варшамова—Гилберта, которое мы без доказательства уже приводили на стр. 406 (для случая кодов, исправляющих произвольное число n ошибок). Ясно, что в общем случае произвольного n неравенство это доказывается точно так же, как и в случае $n = 2$: здесь только надо требовать, чтобы новый столбец b_i каждый раз не был нулевым, не равнялся ни одному из старых столбцов, а также ни одной из сумм двух, трех и т. д. вплоть до $2n - 1$ старых столбцов. Отсюда и следует, что

$$K > \log (1 + C_{N-1}^1 + C_{N-1}^2 + \cdots + C_{N-1}^{2n-1}).$$

Будем теперь снова считать, что $n = 2$. Ясно, что при малых значениях K и N можно надеяться непосредственно проверить все условия, налагаемые на столбцы матрицы $\boldsymbol{B}$, — и таким образом подобрать код, исправляющий все одиночные и двойные ошибки. Именно так мы, собственно говоря, и поступили на стр. 404, где с помощью подбора для случая $K = 4$ и $N = 5$ был построен код с проверками на четность, исправляющий все одиночные и все двойные ошибки; отвечающая этому коду

проверочная матрица очевидно имеет следующий вид:

$$\begin{pmatrix} 1&1&0&0&0\\ 1&0&1&0&0\\ 1&0&0&1&0\\ 1&0&0&0&1 \end{pmatrix}.$$

[Заметим, что при $N = 5$ и $n = 2$ неравенство Хэмминга указывает, что обязательно $K \geqslant 4$; из неравенства же Варшамова — Гилберта здесь вытекает, что при $K \geqslant 4$ наверное можно построить код, исправляющий все одиночные и все двойные ошибки.] Немного более сложна, но все еще вполне доступна, проверка того, что при $K = 7$ и $N = 10$ все столбцы и попарные суммы столбцов (7×10)-матрицы

$$\begin{pmatrix} 1&0&0&0&0&0&0&1&0&1\\ 0&1&0&0&0&0&0&0&0&1\\ 0&0&1&0&0&0&0&1&0&1\\ 0&0&0&1&0&0&0&0&1&1\\ 0&0&0&0&1&0&0&1&1&0\\ 0&0&0&0&0&1&0&0&1&0\\ 0&0&0&0&0&0&1&1&1&0 \end{pmatrix}$$

различны между собой, так что соответствующий код (кодовые обозначения которого все содержат по 3 информационных сигнала и по 7 контрольных сигналов) позволяет исправить все одиночные и все двойные ошибки в блоках из 10 сигналов. [При $N = 10$ из неравенства Хэмминга вытекает, что обязательно $K \geqslant 6$, а из неравенства Варшамова — Гилберта следует, что при $K \geqslant 8$ наверное можно построить интересующий нас код.]

Однако при дальнейшем возрастании значений K и N громоздкость описанной процедуры подбора матрицы B и проверки справедливости для столбцов этой матрицы нужных условий быстро возрастают; уже в случае (8×15)-матрицы B, выписанной ниже на стр. 430, задача выполнения такой проверки вряд ли кому-нибудь покажется особенно привлекательной.

Укажем теперь вкратце некоторые основные принципы алгебраической теории кодирования, сыгравшей основную роль в нахождении общих методов построения практически используемых кодов, позволяющих обнаружить и исправить в блоке из N сигналов любое число ошибок, не превосходящее заданного числа n. До сих пор мы рассматривали код как совокупность некоторых кодовых обозначений — блоков $a = (a_0, a_1, \ldots, a_{N-1})$ из N цифр 0 и 1 (т. е. из N элементов простейшего алгебраического поля из двух элементов). Ясно, что с тем же правом мы можем сопоставить каждому кодовому обозначению многочлен степени не выше $N - 1$:

$$a(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_{N-1} x^{N-1},$$

относительно неизвестной x с коэффициентами из нашего поля и рассматривать код как некоторую совокупность «кодовых многочленов» $a(x)$. Всевозможным кодам с проверками на четность в таком случае будут соответствовать всевозможные совокупности многочленов $a(x)$ такие, что сумма любых двух многочленов, принадлежащих нашей совокупности, а также и «нулевой многочлен» $0 = 0 + 0 \cdot x + \ldots + 0 \cdot x^{N-1}$ обязательно принадлежат к той же совокупности. Существует обширный класс очень простых совокупностей многочленов, очевидным образом удовлетворяющих указанным двум условиям — это совокупности всех многочленов $a(x)$ степени не выше некоторого $N - 1$, делящихся без остатка на какой-либо фиксированный многочлен $g(x) = g_0 + g_1 x + \ldots + g_K x^K$ степени $K < N - 1$, т. е. представимых в виде

$$a(x) = c(x)\, g(x), \tag{6}$$

где $c(x)$ — произвольный многочлен, степень которого не превосходит $N - K - 1$. Каждой такой совокупности отвечает вполне определенный код с проверками на четность, который мы будем называть кодом, порожденным многочленом $g(x)$; сам же многочлен $g(x)$ в этом случае называется порождающим многочленом нашего кода. В случае кодов, порожденных многочленами, задание порождающего многочлена $g(x)$ представляет собой самый компактный способ задания соответствующего кода, однозначно определяющий все его характеристики (в частности, набор всех кодовых

обозначений $\boldsymbol{a}$ и соответствующую проверочную матрицу $\boldsymbol{B}$). Если мы запишем произвольный кодовый многочлен $a(x)$ в виде

$$a(x) = a_0 + a_1x + \cdots + a_{K-1}x^{K-1} + a_Kx^K + \\ + a_{K+1}x^{K+1} + \cdots + a_{N-1}x^{N-1},$$

то ясно, что последние $M = N - K$ коэффициентов $a_K, a_{K+1}, \ldots, a_{N-1}$ здесь можно будеть выбрать произвольным образом, а первые K коэффициентов $a_0, a_1, \ldots, a_{K-1}$ после этого уже будут однозначно определяться условием делимости $a(x)$ на $g(x)$ (а именно, поскольку в 2-арифметике $r(x) = -r(x)$, то многочлен $a_0 + a_1x + \ldots + a_{K-1}x^{K-1}$ должен равняться остатку от деления $a_Kx^K + a_{K+1}x^{K+1} + \ldots + a_{N-1}x^{N-1}$ на $g(x)$). Отсюда видно, что последние $N - K$ сигналов $a_K, a_{K+1}, \ldots, a_{N-1}$ в данном случае будут играть роль информационных сигналов, а первые K сигналов $a_0, a_1, \ldots, a_{K-1}$ будут контрольными; общее число кодовых слов здесь равно 2^{N-K}. Принятому на приемном конце линии связи блоку $\boldsymbol{a}' = (a_0', a_1', \ldots, a_{N-1}')$ будет отвечать многочлен

$$a'(x) = a_0' + a_1'x + \cdots + a_{N-1}'x^{N-1},$$

отличающийся от «переданного от многочлена» $a(x)$ на «многочлен ошибок»

$$e(x) = e_0 + e_1x + \ldots + e_{N-1}x^{N-1},$$

где, как и раньше, $e_i = a_i' - a_i$ (т. е. $e_i = 1$, если i-й сигнал исказился в процессе передачи, и $e_i = 0$, если он принят правильно). Из-за наличия добавочного «многочлена ошибок» $e(x)$ многочлен $a'(x)$, вообще говоря, уже не будет делиться без остатка на $g(x)$. Ненулевой остаток $r(x)$ от деления $a'(x)$ на $g(x)$ (равный, очевидно, остатку при делении $e(x)$ на $g(x)$) как раз и свидетельствует о наличии искажений при передаче; этот остаток содержит всю информацию об ошибках, доступную на приемном конце (в этом отношении он вполне аналогичен блоку $\boldsymbol{s} = \boldsymbol{Ba}'$, с которым мы имели дело при использовании матричной записи произвольных кодов с проверками на четность).

В алгебраической теории кодирования основное внимание уделяется не общим кодам с проверками на четность и даже не произвольным кодам, порожденным многочлена-

ми, а специальным классам таких кодов, обладающим особенно простой алгебраической структурой, позволяющей заметно облегчить общую процедуру кодирования и декодирования. Важнейшим из таких специальных классов является класс циклических кодов. Код с проверками на четность называется *циклическим*, если для каждого его кодового обозначения $a=(a_0, a_1, a_2, \ldots, a_{N-1})$ блок $(a_{N-1}, a_0, a_1, \ldots, a_{N-2})$, получаемый из a с помощью «циклического сдвига», также является кодовым обозначением. Ясно, что в таком случае блок $(a_{N-i}, a_{N-i+1}, \ldots, a_{N-i-1})$, полученный из a с помощью i-кратного применения «циклического сдвига», также будет кодовым обозначением при любом $i = 1, 2, \ldots, N-1$.

Важным свойством циклических кодов является то, что все они порождены многочленами, причем класс отвечающих им порождающих многочленов $g(x)$ может быть очень просто охарактеризован. В самом деле, допустим сперва, что мы имеем дело с кодом, порожденным многочленом $g(x)$ (т. е. с совокупностью кодовых многочленов $a(x)$ вида (6)). Пусть

$$a_1(x) = a_{N-1} + a_0 x + a_1 x^2 + \ldots + a_{N-2} x^{N-1}$$

— многочлен, отвечающий блоку $(a_{N-1}, a_0, a_1, \ldots, a_{N-2})$. Так как

$$a_1(x) = x(a_0 + a_1 x + \cdots + a_{N-1} x^{N-1}) - a_{N-1}(x^N - 1) = xa(x) - a_{N-1}(x^N - 1), \quad (7)$$

где, как обычно, $a(x) = a_0 + a_1 x + \ldots + a_{N-1} x^{N-1}$, то ясно, что в общем случае, когда $a_{N-1} \neq 0$, многочлен $a_1(x)$ будет одновременно с $a(x)$ кодовым многочленом (т. е. будет делиться без остатка на $g(x)$) тогда и только тогда, когда $g(x)$ является делителем $x^N - 1$ [1]). Таким образом, *код, порожденный многочленом $g(x)$, будет циклическим в том (и только том) случае, когда $g(x)$ — это делитель многочлена $x^N - 1$.*

---

[1]) Такие многочлены $g(x)$ в алгебре называются многочленами деления окружности; для случая, когда коэффициенты $g(x)$ — это обычные вещественные числа, они изучались еще знаменитым немецким математиком К. Ф. Гауссом в начале XIX столетия.

Рассмотрим теперь совершенно произвольный циклический код, и пусть $a(x)$ — один из отвечающих ему кодовых многочленов. В таком случае из равенства (7) сразу следует, что наряду с $a(x)$ в число кодовых многочленов нашего кода обязательно входит и остаток от деления многочлена $xa(x)$ на $x^N - 1$. Но тогда ясно, что в число кодовых многочленов будут входить и остатки от деления на $x^N - 1$ многочленов $x \cdot x\, a(x) = x^2 a(x)$, $x \cdot x^2 a(x) = x^3 a(x)$ и т. д., т. е. остатки от деления на $x^N - 1$ всевозможных произведений $x^n a(x)$, где n — какое угодно неотрицательное целое число. Так как к тому же сумма любых кодовых многочленов также всегда является кодовым многочленом, то из сказанного вытекает, что *наряду с $a(x)$ кодовыми многочленами обязательно будут и все остатки от деления на $x^N - 1$ многочленов вида $b(x)a(x)$, где $b(x) = b_0 + b_1 x + \ldots + b_n x^n$ — произвольный многочлен с коэффициентами из нашего поля с двумя элементами* (т. е. равными либо нулю, либо единице).

Совокупность всевозможных многочленов степени не выше $N - 1$ можно рассматривать как совокупность всевозможных остатков от деления многочленов любых степеней на $x^N - 1$. В таком случае выведенное выше свойство совокупности кодовых многочленов $a(x)$ произвольного циклического кода на языке общей алгебры можно будет сформулировать следующим образом: такая совокупность кодовых многочленов представляет собой *идеал* в множестве всех остатков от деления на $x^N - 1$ (см. ниже Приложение II, где на стр. 468 дано общее определение идеала, а также рассмотрен и нужный нам частный случай этого понятия). В дальнейшем общее определение идеала нами нигде не будет использоваться; единственное, что нам понадобится — это следующая простая алгебраическая теорема (которую читатель, если угодно, может принять на веру, но может и ознакомиться с ее доказательством по Приложению II): *любой идеал в множестве остатков от деления произвольных многочленов на какой-то фиксированный многочлен $f(x)$ степени N совпадает с совокупностью многочленов вида $c(x)g(x)$, где $g(x)$ — некоторый делитель многочлена $f(x)$ и степень $c(x)g(x)$ не превосходит $N - 1$.* Эта алгебраическая теорема как раз и показывает, что *любой циклический код порождается каким-то делителем $g(x)$ многочлена $x^N - 1$.*

Пусть теперь $g(x)$ — делитель $x^N - 1$, так что

$$x^N - 1 = g(x)\, h(x);$$

в таком случае легко показать, что кодовые многочлены циклического кода с порождающим многочленом $g(x)$— это такие многочлены $a(x)$ степени не выше $N - 1$, для которых $a(x)\, h(x)$ делится без остатка на $x^N - 1$. В самом деле, если $a(x) = c(x)\, g(x)$, то очевидно, что

$$a(x)h(x) = c(x)g(x)\, h(x) = c(x)(x^N - 1)$$

делится без остатка на $x^N - 1$; обратно, если $a(x)h(x) = b(x)(x^N - 1)$ делится без остатка на $x^N - 1$, то ясно, что $a(x) = b(x)\, g(x)$. Указанное свойство многочленов $a(x)$ очень облегчает проверку наличия ошибок при передаче: если $a'(x) = a(x) + e(x)$, где $e(x) \neq 0$, то $a'(x)h(x)$, вообще говоря, не будет делиться на $x^N - 1$, причем, как легко видеть, вся информация об имевших место ошибках (т. е. о многочлене $e(x)$), имеющаяся на приемном конце линии связи, будет содержаться в остатке от деления $a'(x)h(x)$ на $x^N - 1$ (заметим, что деление произвольного многочлена $d(x)$ на $x^N - 1$ осуществить крайне легко; для этого надо только заменить в $d(x)$ все степени x^M, где $M \geqslant N$, степенями x^m, где m — это остаток от деления M на N). Поэтому при декодировании циклических кодов очень большую роль играет многочлен $h(x)$, который принято называть проверочным многочленом циклического кода —полученный на приемном конце линии связи многочлен $a'(x)$ следует прежде всего умножить на проверочный многочлен $h(x)$, и тогда остаток от деления этого произведения на $x^N - 1$ будет однозначно определять расшифровку принятого сообщения (т. е. выбор «наиболее вероятного многочлена ошибок» $e(x)$).

Циклические коды представляют собой специальный класс кодов с проверками на четность, общие свойства которого пока еще мало изучены. Так, например, если ограничиться использованием лишь циклических кодов, то неизвестно, можно ли или нет добиться передачи информации по простейшей двоичной симметричной линии связи с заданной скоростью, меньшей $C = Lc$ бит/ед. времени, и сколь угодно малой вероятностью ошибки; более того, здесь даже неизвестно, можно ли или нет осуществить

передачу хоть с какой-нибудь отличной от нуля скоростью и сколь угодно малой вероятностью ошибки [1]). Однако их большим преимуществом является то, что здесь могут быть развиты сравнительно не очень сложные алгебраические методы декодирования, во многих случаях позволяющие реально осуществить это декодирование за сравнительно небольшое время (см., например, [22], [168], [169], а также довольно сложную книгу [183], специально посвященную этому вопросу).

Особенно плодотворным оказалось применение циклических кодов для исправления в блоках длины N всех ошибок, *число которых не превосходит заданного* n. Заметим, что одной ошибке при передаче отвечает «многочлен ошибок» $e(x)$, состоящий из одного слагаемого x^i. Поэтому для того, чтобы с помощью кода, порожденного многочленом $g(x)$, можно было исправить все *одиночные* ошибки, надо только, чтобы все одночлены $1, x, x^2, \ldots, x^{N-1}$ давали при делении на $g(x)$ разные остатки, т. е. чтобы ни один двучлен

$$x^j - x^i = x^i(x^{j-i} - 1), \text{ где } i < N,\ j < N \text{ и } j > i,$$

не делился на $g(x)$. В частном случае циклических кодов (т. е. многочленов $g(x)$, являющихся делителями $x^N - 1$) многочлены $g(x)$ с нужными свойствами всегда существуют и хорошо изучены для всех $N = 2^K - 1$; поэтому все коды Хэмминга с $N = 2^K - 1$ очень легко могут быть осуществлены в виде циклических кодов. В частности, легко проверить, что в случае $K = 3$, $N = 7$ (рассмотренном на стр. 399—400) порождающий многочлен $g(x)$ и проверочный многочлен $h(x)$ могут быть выбраны в виде

$$g(x) = x^3 + x + 1, \qquad h(x) = x^4 + x^2 + x + 1$$

(непосредственное умножение показывает, что $g(x)h(x) = x^7 - 1$, как и должно быть); в случае же $K = 4$, $N = 15$

---

[1]) Напомним, что как отмечалось на стр. 348—349, до появления работы Шеннона [1] невозможность такой передачи казалась правдоподобной даже в случае использования произвольных кодов. Сейчас мы знаем, что для произвольных кодов дело обстоит совсем иначе, но в применении к одним лишь циклическим кодам такая возможность пока не исключена.

(рассмотренном на стр. 418) можно положить

$$g(x) = x^4 + x + 1,\quad h(x) = x^{11} + x^8 + x^7 + x^5 + x^3 + x^2 + x + 1$$

(при этом $g(x)h(x) = x^{15} - 1$).

Аналогично этому для кодов, позволяющих исправить одиночные и двойные ошибки, все одночлены x^i и двучлены $x^i + x^j$, где $i < N$ и $j < N$, должны давать при делении на $g(x)$ различные остатки; в случае кодов, исправляющих также и тройные ошибки, сюда приходится добавить также и все трехчлены $x^i + x^j + x^k$ и т. д. Ясно, что возникающие здесь задачи являются сугубо алгебраическими по своему характеру; однако их решение оказывается довольно сложным.

Общий метод построения циклических кодов, позволяющих исправить любое, меньшее n, число ошибок в блоках длины $N = 2^K - 1$ и обладающих проверочными матрицами с nK строками и N столбцами (т. е. содержащих не более nK контрольных сигналов в блоке из $N = 2^K - 1$ сигналов [1])), был указан лишь в 1959—1960 гг. независимо А. Хоквингемом [175] и Р. Боузом и Д. Чоудхури [174] [2]). В основе конструкции Хоквингема — Боуза — Чоудхури лежит непосредственное описание порождающих коды многочленов $g(x)$ с помощью задания всех их корней, т. е. всех решений уравнения $g(x) = 0$. Основное затруднение здесь состоит в том, что подобно тому, как корни обычного многочлена с вещественными коэффициентами не обязаны быть вещественными числами, а могут принадлежать более широкому (т. е. содержащему поле действительных чисел в качестве своей части) полю *комплексных чисел*, так и корни нашего многочлена $g(x)$ с коэффициентами из поля с двумя

[1]) Так как соответствующий код не является систематическим, то из того, что проверочная матрица содержит nK строк, можно лишь заключить, что истинное число контрольных сигналов здесь не превосходит nK (см. выше, стр. 410).

[2]) Вообще говоря, кроме простейших (так называемых примитивных) кодов Боуза — Чоудхури — Хоквингема, исправляющих заданное число ошибок в блоках из $N = 2^K - 1$ сигналов, существуют и «непримитивные» коды того же типа, для которых длина блока N является нечетным числом, не представимым в виде $2^K - 1$. Об этих последних кодах мы, однако, совсем не будем говорить (за исключением, впрочем, сноски на стр. 438).

элементами 0 и 1 могут сами принадлежать более широкому полю с 2^m различными элементами (где m — некоторое целое число). Если w — элемент этого нового поля, являющийся одним из корней $g(x)$, то весь набор корней, полностью задающий $g(x)$, будет совпадать с какой-то конечной цепочкой последовательных степеней корня w. Более подробное разъяснение этого утверждения требует привлечения сложного алгебраического аппарата, явно выходящего за границы нашей книги; поэтому мы ограничимся здесь лишь ссылками на книги [22], [168], [169], [184] и на более популярную, чем они (но также более трудную, чем наша книга), обзорную статью [167]. Для того, однако, чтобы все же дать хоть некоторое представление о характере получающихся при этом результатов, мы приведем в заключение два конкретных примера кодов Боуза — Чоудхури — Хоквингема, исправляющих кратные ошибки.

Оба эти примера относятся к случаю, когда $K = 4$, $N = 2^4 - 1 = 15$. Соответствующий этим значениям K и N код Хэмминга, исправляющий все одиночные ошибки, задается проверочной матрицей, выписанной выше на стр. 418. В случае кода, исправляющего одиночные и парные ошибки, проверочная матрица будет уже (8×15)-матрицей следующего вида:

$$\begin{pmatrix} 1&1&1&1&0&1&0&1&1&0&0&1&0&0&0 \\ 0&1&1&1&1&0&1&0&1&1&0&0&1&0&0 \\ 0&0&1&1&1&1&0&1&0&1&1&0&0&1&0 \\ 1&1&1&0&1&0&1&1&0&0&1&0&0&0&1 \\ 1&1&1&1&0&1&1&1&1&0&1&1&1&1&0 \\ 1&0&1&0&0&1&0&1&0&0&1&0&1&0&0 \\ 1&1&0&0&0&1&1&0&0&0&1&1&0&0&0 \\ 1&0&0&1&1&1&0&0&0&1&1&0&0&0&1 \end{pmatrix}.$$

Эта матрица довольно громоздка; поэтому гораздо удобнее задать соответствующий код при помощи его порождающего многочлена

$$g(x) = (x^4 + x + 1)(x^4 + x^3 + x^2 + x + 1) = \\ = x^8 + x^7 + x^6 + x^4 + 1,$$

или его проверочного многочлена

$$h(x) = (x+1)(x^2+x+1)(x^4+x^3+1) = \\ = x^7 + x^6 + x^4 + 1$$

(нетрудно проверить, что действительно $g(x)\,h(x) = x^{15}-1$). Заметим, что рассматриваемый код состоит из кодовых обозначений длины 15, включающих 7 информационных и 8 контрольных сигналов. В силу неравенства Хэмминга мы можем утверждать, что при $N = 15$ код, исправляющий все одиночные и все двойные ошибки, не может содержать меньше чем 7 контрольных сигналов; неравенство Варшамова — Гилберта здесь показывает, что такой код наверное можно построить, если $K = 9$.

Если теперь пожелать построить код, исправляющий в блоках из 15 сигналов все одиночные, все двойные и все тройные ошибки, то проверочная матрица такого кода Боуза — Чоудхури — Хоквингема будет иметь $3K = 12$ строк (и, как и раньше, 15 столбцов). Порождающий многочлен интересующего нас кода имеет сравнительно простой вид:

$$g(x) = (x^2+x+1)(x^4+x+1)(x^4+x^3+x^2+x+1) = \\ = x^{10} + x^8 + x^5 + x^4 + x^2 + x + 1,$$

а его проверочный многочлен равен

$$h(x) = (x+1)(x^4+x^3+1) = x^5 + x^3 + x + 1$$

(при этом опять $g(x)\,h(x) = x^{15} - 1$). Проверочной матрицей нашего кода является следующая (12×15)-матрица:

$$\begin{pmatrix} 1&0&0&0&1&0&0&1&1&0&1&0&1&1&1 \\ 0&0&0&1&0&0&1&1&0&1&0&1&1&1&1 \\ 0&0&1&0&0&1&1&0&1&0&1&1&1&1&0 \\ 0&1&0&0&1&1&0&1&0&1&1&1&1&0&0 \\ 1&0&0&0&1&1&0&0&0&1&1&0&0&0&1 \\ 0&1&1&1&1&0&1&1&1&1&0&1&1&1&1 \\ 0&0&1&0&1&0&0&1&0&1&0&0&1&0&1 \\ 0&0&0&1&1&0&0&0&1&1&0&0&0&1&1 \\ 1&0&1&1&0&1&1&0&1&1&0&1&1&0&1 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&1&1&0&1&1&0&1&1&0&1&1&0&1&1 \\ 0&1&1&0&1&1&0&1&1&0&1&1&0&1&1 \end{pmatrix}.$$

Заметим, что хотя эта матрица имеет 12 строк, число «контрольных сигналов», отвечающих соответствующему коду, равно 10 — это сразу видно из того, что порождающий многочлен $g(x)$ здесь является многочленом десятой степени [1]). Таким образом, при использовании рассматриваемого кода каждая пятерка «информационных сигналов» дополняется десятью «контрольными сигналами» — только после этого в принятой на приемном конце линии цепочке из 15 сигналов можно обнаружить и исправить все без исключения одиночные, двойные и тройные ошибки. Нетрудно также видеть, что исправление всех таких ошибок в блоке из 15 сигналов никак не может быть достигнуто, если использовать меньше 10 «контрольных сигналов» — это обстоятельство сразу вытекает из неравенства Хэмминга (неравенство же Варшамова — Гильберта здесь показывает, что нужный нам код наверное может быть построен, если использовать 12 или больше «контрольных сигналов»).

Данные о числе «информационных» и «контрольных» сигналов для большого числа кодов Боуза — Чоудхури — Хоквингема могут быть найдены в гл. 9 книги [164] (см. также гл. 7 и 12 книги [169]). Согласно приведенным в [168] результатам все коды этого типа с $N \leqslant 15$, а также и коды с произвольным N и $n = 2$, являются о п т и м а л ь н ы м и в том смысле, что не существует кодов с той же длиной N «блоков» и тем же общим числом кодовых обозначений S (т. е. той же скоростью передачи информации $v = \frac{L}{N} \log S$ бит/ед. времени), приводящих к меньшей вероятности ошибки при их использовании для передачи по двоичной симметричной линии связи (ср. ниже 440). При $N = 1023$ $(= 2^{10} - 1)$ число «контрольных сигналов» при различных n оказывается довольно близким к соответствующей границе Варшамова — Гилберта. Однако при еще бо́льших N это число должно стать более близким не к верхней границе Вар-

[1]) Тот же вывод в рассматриваемом случае можно сделать и исходя из самого вида проверочной матрицы — так как ее третья снизу строка состоит из одних нулей, а две последние строки одинаковы, то ясно, что код не изменится, если из последних трех строк мы сохраним лишь одну (последнюю или предпоследнюю) строку.

шамова — Гилберта, а к нижней границе Хэмминга. В самом деле, воспользовавшись оценкой сверху биномиальных коэффициентов C_N^n, доставляемой неравенством (**) на стр. 221, и аналогичной же оценкой этих коэффициентов снизу (или же просто заменив в точной формуле $C_N^n = \frac{N!}{n!\,(N-n)!}$ факториалы $N!$ и $(N-n)!$ их приближенными значениями при большом N, имеющимися во многих курсах высшей математики), нетрудно показать, что при очень большом N общее неравенство Хэмминга принимает вид

$$2^K \geqslant AN^n, \text{ т. е. } K \geqslant n \log N + A_1,$$

где K — число контрольных сигналов, n — максимальное число исправляемых ошибок, а A и $A_1 = \log A$ — какие-то числа (A — положительное, а A_1 — возможно и отрицательное), зависящие от n, но не зависящие от N. Аналогичным образом неравенство Варшамова — Гилберта в случае большого N позволяет заключить, что если

$$2^K > BN^{2n}, \text{ т. е. } K > 2n \log N + B_1,$$

где B и $B_1 = \log B$ — другие зависящие от n (но не от N) числа, то наверное существует код, позволяющий исправить при передаче любое, не превосходящее n, число ошибок в блоке из N сигналов. В случае кодов Боуза — Чоудхури — Хоквингема с $N = 2^{K_1} - 1$ (так что $K_1 \approx \approx \log N$) число K контрольных сигналов, как указывалось выше, не превосходит $nK_1 \approx n \log N$; поэтому при больших значениях N число контрольных сигналов в этих кодах всегда близко к соответствующей нижней границе Хэмминга. В этом смысле указанные коды являются близкими к наилучшим возможным в отношении их использования для исправления заданного фиксированного числа ошибок в очень длинных блоках.

Разумеется, выбор очень длинных кодовых обозначений (т. е. очень большого N) невыгоден, если коды исправляют лишь фиксированное число n ошибок, так как с ростом N резко возрастает вероятность появления большего, чем n, числа ошибок в блоке длины N. Поэтому при увеличении N естественно увеличивать и значение n; однако если n увеличивать пропорционально N, то с ростом N, как оказывается, будет все время убывать достигаемая

скорость передачи информации (см. [168], гл. 9). Наиболее существен, однако, не вопрос об оптимальном выборе значений N и n, а вопрос о методах декодирования получающихся кодов при больших N; именно трудность декодирования в первую очередь ограничивает возможности подбора параметров кода, обеспечивающих и малую вероятность ошибки, и большую скорость передачи. В применении к кодам Боуза — Чоудхури — Хоквингема разработан целый ряд специальных методов декодирования, позволяющих эффективно его осуществлять вплоть до длин N кодовых обозначений, имеющих порядок многих сотен или даже нескольких тысяч. На этих методах мы здесь, однако, уже не можем задерживаться — по их поводу можно лишь отослать читателя к (довольно сложным) книгам [22], [168], [169], [183] и [184].

Будем, как и выше, рассматривать лишь случай двоичной линии связи (использующей два элементарных сигнала), а код будем понимать как некоторую совокупность кодовых обозначений — цепочек $a = (a_0, a_1, \ldots, a_{N-1})$ из N цифр 0 и 1. При изучении кодов, позволяющих исправлять ошибки при передаче, важную роль играет р а с с т о я н и е Х э м м и н г а $|b - a|_x$ между двумя цепочками $b = (b_0, b_1, \ldots, b_{N-1})$ и $a = (a_0, a_1, \ldots, a_{N-1})$, которое по определению равно числу цифр a_l таких, что $b_l \neq a_l$ (т. е. числу единиц среди разностей $b_l - a_l$, понимаемых в смысле 2-арифметики). Расстояние Хэмминга обладает многими свойствами обычного геометрического расстояния (см., например, ниже Приложение II, стр. 479); оно совпадает с числом искажений отдельных передаваемых сигналов, приводящим к тому, что переданная цепочка a воспринимается на приемном конце линии связи как цепочка b. Понятно, что чем больше будет расстояние Хэмминга между отдельными кодовыми обозначениями, тем меньшей будет вероятность перепутать эти обозначения на приемном конце, т. е. (при прочих равных условиях) тем лучше будет используемый код. Поэтому важной характеристикой кода является отвечающее ему к о д о в о е р а с с т о я н и е $D = \min |a^{(i)} - a^{(j)}|_x$ — расстояние Хэмминга между «самыми близкими» различными кодовыми обозначениями данного кода. Ясно, что в случае кода, позволяющего исправить любое не превосходящее n число ошибок, он не должен содержать двух таких кодовых обозначений $a^{(i)} = (a_0^{(i)}, a_1^{(i)}, \ldots, a_{N-1}^{(i)})$ и $a^{(j)} = (a_0^{(j)}, a_1^{(j)}, \ldots, a_{N-1}^{(j)})$, что, изменив какие-то n или менее цифр первого из них и какие-то n или менее цифр второго, мы получим одну и ту же цепочку b — иначе, приняв эту цепочку b, мы не сможем выяснить, было ли передано обозначение $a^{(i)}$ или $a^{(j)}$. Следовательно, все расстояния $|a^{(i)} - a^{(j)}|_x$ (где $i \neq j$) должны

быть больше $2n$, откуда вытекает, что $D \geqslant 2n+1$, где D — кодовое расстояние нашего кода. Обратно, если $D \geqslant 2n+1$, то, договорившись расшифровывать как кодовое обозначение $a^{(i)}$ все принимаемые цепочки b, принадлежащие шару Хэмминга радиуса n с центром $a^{(i)}$ (т. е. все такие b, что $|b - a^{(i)}|_x \leqslant n$), мы гарантированно исправим любое не превосходящее n число ошибок при передаче. Итак, *код позволяет исправить любое не превосходящее n число ошибок при передаче тогда и только тогда, когда его кодовое расстояние D не меньше чем $2n+1$.* Аналогичным образом легко показать, что *если кодовое расстояние D не меньше чем $2n$, то код позволяет исправить любое, не превосходящее $n-1$, число ошибок и, кроме того, позволяет обнаружить наличие не меньше чем n ошибок* (но в последнем случае он уже может и не позволить однозначно исправить эти n ошибок) [1].

Ясно, что «объем» V_n шара Хэмминга радиуса n, т. е. число «точек» $b = (b_0, b_1, \ldots, b_{N-1})$, принадлежащих такому шару с центром в произвольной «точке» $a = (a_0, a_1, \ldots, a_{N-1})$, задается равенством

$$V = 1 + C_N^1 + C_N^2 + \cdots + C_N^n.$$

Поскольку общее число всех N-членных цепочек равно 2^N, то отсюда немедленно следует, что число S различных кодовых обозначений длины N, входящих в код, позволяющий исправить любое, не превосходящее n, число ошибок, должно удовлетворять условию

$$S \leqslant \frac{2^N}{1 + C_N^1 + \cdots + C_N^n}. \qquad (8)$$

Это простое условие, ограничивающее сверху возможное число S кодовых обозначений (а, значит, и максимальную возможную скорость передачи информации $v = \frac{L}{N} \log S$ бит/ед. времени), называется **верхней границей Хэмминга числа кодовых обозначений**. В частном случае кодов с проверками на четность (т. е., иначе, линейных или групповых кодов) оно совпадает с рассмотренной на стр. 405 нижней границей Хэмминга числа контрольных сигналов: в самом деле, для (N,M)-кода с проверками на четность число кодовых слов S равно

$$2^M = \frac{2^N}{2^K},$$

---

[1]) Надо, впрочем, иметь в виду, что кодовое расстояние D не определяет полностью способность кода исправлять ошибки при передаче. Так, например, если $D = 2n$, то зачастую для многих (хотя и не всех) передаваемых обозначений $a^{(i)}$ код все равно позволяет исправить заметно больше чем n ошибок при передаче.

и поэтому условие (8) здесь точно совпадает с неравенством Хэмминга. Заметим, однако, что условие (8), в отличие от неравенства Хэмминга для числа K, применимо к л ю б ы м кодам, а не только к кодам с проверками на четность.

Коды, обладающие тем свойством, что для них левая и правая части (8) совпадают друг с другом, называются с о в е р ш е н н ы м и (или, реже, п л о т н о у п а к о в а н н ы м и). Совершенные коды замечательны тем, что они практически во всех отношениях являются оптимальными (т. е. самыми лучшими). Мы уже видели, что среди кодов заданной длины N, исправляющих заданное число n ошибок, совершенным кодам соответствует *наибольшее число S кодовых слов, т. е. наибольшая скорость передачи информации*; в случае совершенных кодов с проверками на четность, исправляющих заданное число ошибок, *число контрольных сигналов K является наименьшим возможным*. Предположим теперь, что наш код используется для передачи информации по двоичной симметричной линии связи; при этом очень важной характеристикой качества передачи будет с р е д н я я в е р о я т н о с т ь о ш и б к и при расшифровке

$$Q = \frac{Q_1 + Q_2 + \cdots + Q_S}{S},$$

где S — общее число кодовых обозначений кода, а Q_i — вероятность того, что переданное i-е кодовое обозначение $\boldsymbol{a}^{(i)}$ будет неправильно расшифровано на приемном конце. Пусть теперь $m_k^{(i)}$ — это число цепочек $\boldsymbol{b}$, находящихся на расстоянии Хэмминга k от i-го кодового обозначения $\boldsymbol{a}^{(i)}$ и расшифровываемых как $\boldsymbol{a}^{(i)}$ на приемном конце линии. Так как в случае передачи цепочки $\boldsymbol{a}^{(i)}$ вероятность получения на приемном конце двоичной симметричной линии связи любой такой цепочки $\boldsymbol{b}$, очевидно, равна $p^k(1-p)^{N-k}$, то вероятность правильной расшифровки переданной цепочки $\boldsymbol{a}^{(i)}$ равна сумме

$$m_0^{(i)}(1-p)^N + m_1^{(i)} p(1-p)^{N-1} + \cdots + m_k^{(i)} p^k (1-p)^{N-k} + \cdots$$

Отсюда видно, что средняя вероятность ошибки при расшифровке равна

$$Q = 1 - \frac{1}{S}\left[m_0(1-p)^N + m_1 p(1-p)^{N-1} + \cdots \right.$$
$$\left. \cdots + m_k p^k (1-p)^{N-k} + \cdots\right],$$

где $m_k = m_k^{(1)} + m_k^{(2)} + \ldots + m_k^{(S)}$ — суммарное число цепочек $\boldsymbol{b}$, находящихся на расстоянии Хэмминга k от какого-то кодового обозначения $\boldsymbol{a}^{(i)}$ и расшифровываемых как это $\boldsymbol{a}^{(i)}$ (так что $m_0 + m_1 + \ldots + m_k + \ldots = 2^N$). Но полное число цепочек длины N, находящихся на заданном расстоянии Хэмминга k от фиксированной цепочки $\boldsymbol{a}^{(i)}$, равно C_N^k; поэтому для кода, состоящего из S

кодовых обозначений длины N,

$$m_0 \leqslant S, \quad m_1 \leqslant SC_N^1, \quad \ldots, \quad m_k \leqslant SC_N^k, \ldots.$$

Пусть теперь n — это наибольшее целое число такое, что

$$S + SC_N^1 + \cdots + SC_N^n \leqslant 2^N,$$

но

$$S + SC_N^1 + \cdots + SC_N^n + SC_N^{n+1} > 2^N,$$

так что

$$2^N - (S + SC_N^1 + \cdots + SC_N^n) = T < SC_N^{n+1};$$

тогда, если $m_0 = S$, $m_1 = SC_N^1$, $\ldots$, $m_n = SC_N^n$, то $m_{n+1} \leqslant T$. Будем, как обычно, считать, что $p < \frac{1}{2}$; тогда вероятность $p^k (1-p)^{N-k}$ будет тем меньше, чем больше k, и поэтому случай, когда $m_0 = S$, $m_1 = SC_N^1$, $\ldots$, $m_n = SC_N^n$, $m_{n+1} = T$, является наиболее благоприятным, т. е. приводящим к наименьшей средней ошибке Q. Следовательно,

$$Q \geqslant 1 - \left[(1-p)^N + C_N^1 p (1-p)^{N-1} + \cdots \right.$$
$$\left. \cdots + C_N^n p^n (1-p)^{N-n} + \frac{T}{S} p^{n+1} (1-p)^{N-n-1}\right]. \qquad (9)$$

Оценка (9) наименьшей возможной средней вероятности ошибки для кода с фиксированными значениями N и S, используемого для передачи по двоичной симметричной линии связи с заданным значением вероятности p искажения сигнала, называется нижней границей Хэмминга средней вероятности ошибки. Для совершенных кодов при условии, что все принимаемые N-членные цепочки, удаленные от какого-то кодового обозначения $a^{(i)}$ на расстояние Хэмминга, не превосходящее n, расшифровываются как $a^{(i)}$, неравенство (9), очевидно, обращается в равенство (причем T здесь равняется нулю); отсюда видно, что для таких кодов *средняя вероятность ошибки меньше, чем для любых других кодов* с теми же значениями N и S.

Совершенные коды имеют очень простой геометрический смысл (в геометрии, определяемой расстоянием Хэмминга): они соответствуют случаям, когда совокупность всевозможных «точек» $b = (b_0, b_1, \ldots, b_{N-1})$ может быть разбита на конечное число «шаров Хэмминга» некоторого радиуса n, взаимно не пересекающихся, но заполняющих в своей совокупности все «пространство» (состоящее из 2^N точек), причем центры этих «шаров» и составляют код (отсюда и название «плотно упакованный код»). Их основным недостатком является то, что таких кодов имеется очень мало —

они существуют лишь для некоторых исключительных значений N и S. Простейшие совершенные коды — это тривиальные коды, состоящие всего из двух кодовых обозначений (0, 0, . . ., 0) и (1, 1, . . ., 1), каждое из которых составлено из нечетного числа $N = 2n + 1$ одинаковых цифр. Для такого кода, очевидно, $D = 2n+1$, и код позволяет исправить n или меньше ошибок; все пространство из $2^N = 2^{2n+1}$ точек здесь распадается на два шара Хэмминга радиуса n (содержащих по $2^{2n} = 2^{N-1}$ точек каждый). Кроме того имеется обширный (и весьма важный) класс совершенных кодов — это $(2^K - 1, 2^K - K - 1)$-коды Хэмминга, в отношении которых на стр. 405 уже отмечалось, что для них неравенство Хэмминга для числа «контрольных сигналов» (эквивалентное неравенству (8)) обращается в равенство. В этом случае все пространство из $2^N = 2^{2^K-1}$ точек распадается на 2^{2^K-K-1} шаров Хэмминга радиуса 1, каждый из которых содержит 2^K точек; здесь $D = 3$ и, следовательно, все одиночные ошибки могут быть исправлены. Но если только допустить, что $n > 1$, а $S > 2$, то сразу возникает прежде всего та трудность, что для существования совершенного кода сумма $1 + C_N^1 + \ldots + C_N^n$ в силу (8) должна равняться некоторой целой степени числа 2, что на самом деле выполняется весьма редко. Занимаясь поисками совершенных кодов, американский ученый Голей (M. J. E. Golay) заметил, что

$$1 + C_{23}^1 + C_{23}^2 + C_{23}^3 = 2048 = 2^{11}$$

— и это подсказало ему, что в принципе может существовать совершенный код с $N=23$ и $S = \frac{2^{23}}{2^{11}} = 2^{12} = 4096$, позволяющий исправить любую комбинацию из трех или менее ошибок. Такой код (называемый с тех пор двоичным совершенным кодом Голея) ему действительно удалось разыскать; он оказался циклическим (23, 12)-кодом с проверками на четность, задаваемым порождающим многочленом

$$g(x) = x^{11} + x^9 + x^7 + x^6 + x^5 + x + 1$$

или же проверочным многочленом

$$h(x) = \frac{x^{23} - 1}{g(x)} = x^{12} + x^{10} + x^7 + x^4 + x^3 + x^2 + x + 1$$

и имеющим кодовое расстояние $D = 7$[1]). Дальнейшие поиски

---

[1]) Код Голея оказался также совпадающим с (непримитивным) кодом Боуза — Чоудхури — Хоквингема, отвечающим значениям $N = 23$ и $n = 2$ (т. е. исправляющим все одиночные и двойные ошибки). Однако построение этого кода по методу Боуза и др. позволяет лишь утверждать, что для него $D \geqslant 5$ (именно это и означает, что код позволяет исправить одиночные и двойные ошибки), в то время как Голей установил, что на самом деле здесь $D = 7$.

новых совершенных кодов к удаче не привели: кроме перечисленных выше никаких других таких кодов до настоящего времени не было найдено, и, по-видимому, их вообще больше не существует [1]). [Это, разумеется, не означает, что больше не существует сумм $1 + C_N^1 + \ldots + C_N^n$, равных степени двойки. Так, например, нетрудно проверить, что $1 + C_{90}^1 + C_{90}^2 = 2^{12}$,—но тем не менее можно доказать, что не существует совершенного кода с $N = 90$ и $n = 2$.]

Поскольку совершенных кодов оказалось так мало, большое внимание было уделено поискам так называемых к в а з и с о в е р ш е н н ы х кодов, несколько худших, чем совершенные, но все же достаточно хороших. Квазисовершенные коды определяются как такие, что для них шары Хэмминга некоторого фиксированного радиуса n с центрами в точках, отвечающих всевозможным кодовым обозначениям, заполняют все пространство из 2^N точек b, за исключением лишь некоторого числа $T < SC_N^{n+1}$ точек (где S — число кодовых обозначений кода), находящихся на расстоянии Хэмминга $n + 1$ по крайней мере от одного (но может быть — и сразу от нескольких) кодовых обозначений. Если мы условимся в случае квазисовершенного кода расшифровывать как $a^{(i)}$ все принимаемые цепочки b, находящиеся на не большем чем n расстоянии Хэмминга от кодового обозначения $a^{(i)}$, а цепочки b, находящиеся на расстоянии $n + 1$ от самого близкого к ним кодового обозначения, расшифровывать как одно (безразлично какое) из кодовых обозначений, удаленных от b на расстояние $n + 1$, то неравенство (9) также и здесь обратится в равенство; поэтому и для квазисовершенных кодов, используемых для передачи по двоичной симметричной линии связи, *средняя вероятность ошибки при расшифровке будет меньше, чем для любых других кодов с теми же значениями N и S.* В то же время квазисовершенных кодов уже имеется значительно больше, чем кодов совершенных (хотя и их тоже все же не очень много). Так, например, коды, исправляющие все одиночные ошибки

---

[1]) П р и м е ч а н и е п р и к о р р е к т у р е. Уже после окончания работы над этой книгой факт отсутствия каких-либо совершенных двоичных кодов, отличных от тех, которые были указаны выше, был, наконец, строго доказан финскими учеными А. Т и е т я в я й н е н о м и А. П е р к о [186] и, независимо от них, В. А. З и н о в ь е в ы м и В. К. Л е о н т ь е в ы м [187] в СССР; в последней из этих работ аналогичный результат получен также и для многих недвоичных совершенных кодов, не рассматриваемых в нашей книге. В самое последнее время Зиновьев и Леонтьев и, независимо от них, Тиетявяйнен и американский ученый ван Линт получили полное решение вопроса о нахождении всех совершенных кодов, использующих p^k элементарных сигналов, где p — произвольное простое, а k — любое целое положительное число; таких кодов также оказалось крайне мало.

в блоках из $N \neq 2^K - 1$ цифр и получающиеся с помощью отбрасывания некоторого числа столбцов в проверочной матрице соответствующего совершенного кода Хэмминга с $N = 2^K - 1$, весьма часто оказываются квазисовершенными (см., например, [168], стр. 105). Квазисовершенными являются и все обсуждавшиеся на стр. 429—431 (примитивные) коды Боуза — Чоудхури—Хоквингема с $N = 2^K - 1$, исправляющие одиночные и двойные ошибки (см. [185]); именно на этом основании на стр. 432 и утверждалось, что такие коды обязательно будут оптимальными. Ряд других примеров квазисовершенных кодов описан в гл. 5 книги [168]; здесь мы, однако, не будем на этом останавливаться.

Приложение I

СВОЙСТВА ВЫПУКЛЫХ ФУНКЦИЙ

Функция $y = f(x)$ называется **выпуклой сверху** (или, короче, просто **выпуклой**) на отрезке от $x = a$ до $x = b$, если в этом интервале любая дуга MN графика функции лежит **над** соответствующей **хордой** MN [1]) (рис. 31). Примерами могут служить логарифмическая функция $y = \log x$ во всей области своего определения, т. е. от 0 до ∞; степенная функция $y = -x^m$ в той же области (здесь предполагается, что $m > 1$); показательная функция $y = -a^x$ в области от $-\infty$ до $+\infty$; функция $y = -x \log x$ в области от 0 до ∞, или функция $y = -x \log x - (1 - x) \log (1 - x)$ в области от 0 до 1 (рис. 32, *а* — *д*).

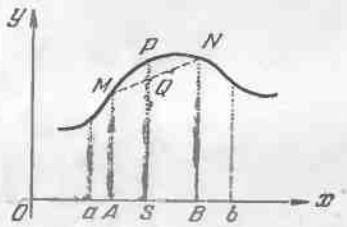


Рис. 31.

Теорема 1. *Если $y = f(x)$ — выпуклая на отрезке от a до b функция, x_1 и x_2 — два значения аргумента этой функции, взятые внутри рассматриваемого отрезка* (т. е. два произвольных числа таких, что $a \leqslant x_1 < x_2 \leqslant b$), *то*

$$\frac{f(x_1) + f(x_2)}{2} < f\left(\frac{x_1 + x_2}{2}\right). \qquad (1)$$

Доказательство (ср. выше, стр. 74). Пусть на рис. 31 $OA = x_1$, $OB = x_2$; в таком случае $AM = f(x_1)$, $BN = f(x_2)$. Далее, если S есть середина отрезка AB, то $OS = \frac{x_1 + x_2}{2}$ и, следовательно, $SP = f\left(\frac{x_1 + x_2}{2}\right)$.

---

[1]) В дифференциальном исчислении указывается **признак выпуклости функции**, применимый к достаточно широкому классу таких функций (в частности, ко всем функциям, рассматриваемым в этом Приложении); он состоит в *отрицательности второй производной y'' функции $y = f(x)$*.

С другой стороны, так как средняя линия SQ трапеции $ABNM$ равна полусумме оснований AM и BN, то

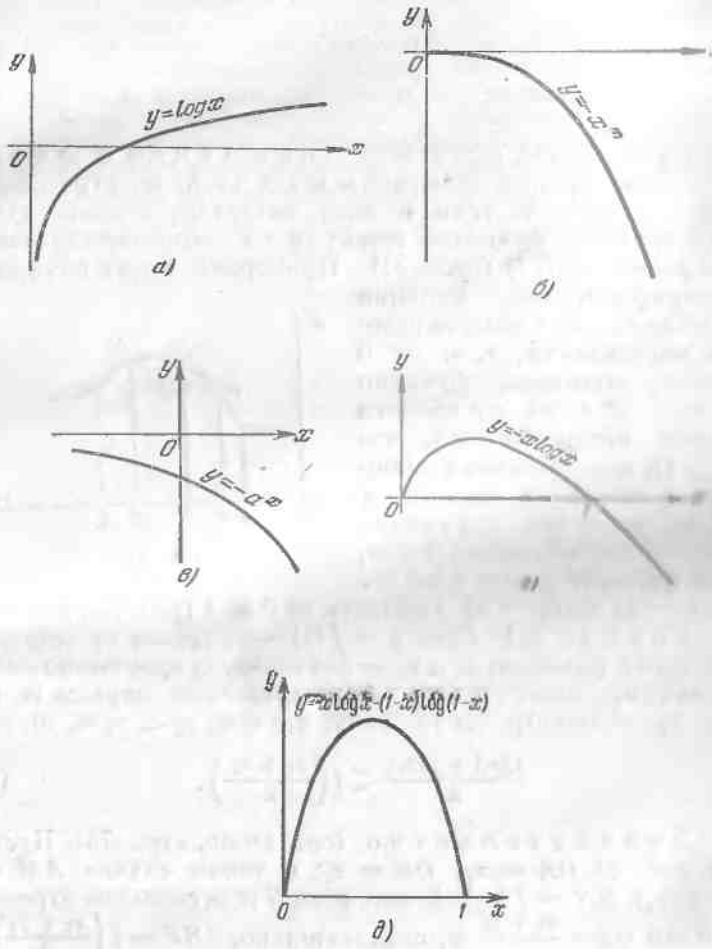


Рис. 32.

$SQ = \frac{f(x_1) + f(x_2)}{2}$. Но, согласно определению выпуклой функции, середина Q хорды MN расположена н и ж е точки

P дуги MN; следовательно,

$$\frac{f(x_1)+f(x_2)}{2}<f\left(\frac{x_1+x_2}{2}\right),$$

— что и требовалось доказать [1]).

Примеры [2]).

а) $y=\log x$. Имеем

$$\frac{\log x_1+\log x_2}{2}<\log\frac{x_1+x_2}{2},$$

т. е.

$$\log\sqrt{x_1x_2}<\log\frac{x_1+x_2}{2},$$

или, наконец,

$$\sqrt{x_1x_2}<\frac{x_1+x_2}{2}$$

— среднее геометрическое двух неравных положительных чисел меньше их среднего арифметического.

б) $y=-x^m$, $m>1$. Здесь получаем

$$-\frac{x_1^m+x_2^m}{2}<-\left(\frac{x_1+x_2}{2}\right)^m$$

или, в другой форме,

$$\frac{x_1^m+x_2^m}{2}>\left(\frac{x_1+x_2}{2}\right)^m,\quad \left(\frac{x_1^m+x_2^m}{2}\right)^{\frac{1}{m}}>\frac{x_1+x_2}{2}.$$

Выражение $\left(\frac{a_1^m+a_2^m+\cdots+a_k^m}{k}\right)^{\frac{1}{m}}$ — корень степени m из среднего арифметического m-х степеней чисел

[1]) Мы ограничиваемся при доказательстве случаем, когда $f(x_1)$ и $f(x_2)$ имеют одинаковые знаки (впрочем, единственно этот случай и будет нам нужен в дальнейшем). Предоставляем читателю самостоятельно рассмотреть случай разных знаков $f(x_1)$ и $f(x_2)$ (здесь вместо свойства средней линии трапеции придется применить следующую теорему: *отрезок средней линии трапеции, заключенный между ее диагоналями, равен полуразности оснований трапеции*).

[2]) В содержании этой книги существенно используются лишь неравенства, связанные с выпуклостью функций $y=-x\log x$ и $y=\log x$ [а также $y=-x\log x-(1-x)\log(1-x)$]; пример б) здесь и ниже имеет лишь иллюстративное значение. [Учение о выпуклых функциях является богатейшим источником всевозможных неравенств, так что число подобных примеров можно было бы значительно увеличить.]

$a_1, a_2, \ldots, a_k$ — называется степенным средним порядка m этих k чисел (в частности, выражение $\sqrt{\frac{a_1^2 + a_2^2 + \cdots + a_k^2}{k}}$, отвечающее случаю $m = 2$, называется среднim квадратичным чисел $a_1, a_2, \ldots, a_k$). Таким образом, полученный результат можно сформулировать так: *степенное среднее порядка $m > 1$ двух неравных положительных чисел всегда больше их среднего арифметического.*

в) $y = -x \log x$. Из теоремы 1 следует:

$$-\frac{x_1 \log x_1 + x_2 \log x_2}{2} < -\frac{x_1 + x_2}{2} \log \frac{x_1 + x_2}{2}$$

или

$$-\frac{1}{2} x_1 \log x_1 - \frac{1}{2} x_2 \log x_2 < -\frac{1}{2} (x_1 + x_2) \log \frac{x_1 + x_2}{2}$$

— результат, которым мы дважды пользовались в гл. II (см. стр. 74 и 94).

Неравенство теоремы 1 может быть обобщено следующим образом:

Т е о р е м а 2. *Если функция $y = f(x)$ — выпуклая в интервале от a до b, x_1 и x_2 — два произвольных числа из этого интервала ($a \leqslant x_1 < x_2 \leqslant b$) и p и q — какие угодно положительные числа, сумма которых равна единице, то*

$$pf(x_1) + qf(x_2) < f(px_1 + qx_2). \quad (2)$$

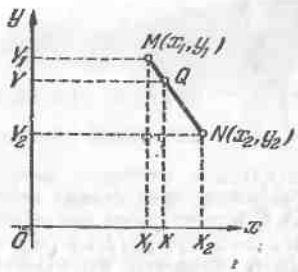


Рис. 33.

При $p = q = \frac{1}{2}$ теорема 2 переходит в теорему 1.

Д о к а з а т е л ь с т в о. Отметим прежде всего, что *если M и N — две точки, имеющие координаты (x_1, y_1) и (x_2, y_2), а Q — точка отрезка MN, делящая этот отрезок в отношении $MQ:QN = q : p$ (где $p+q=1$), то координаты точки Q равны $px_1 + qx_2$ и $py_1 + qy_2$.* Действительно, обозначим через X_1, X_2 и X; Y_1, Y_2 и Y проекции точек M, N и Q на оси координат (рис. 33); в таком случае точки

X и Y будут делить отрезки X_1X_2 и Y_1Y_2 в отношении $q:p$. Отсюда получаем[1]):

$$OX = OX_1 + X_1X = x_1 + q(x_2 - x_1) = \\ = (1 - q)x_1 + qx_2 = px_1 + qx_2$$

и

$$OY = OY_2 + Y_2Y = y_2 + p(y_1 - y_2) = \\ = (1 - p)y_2 + py_1 = py_1 + qy_2.$$

Рассмотрим теперь снова график нашей выпуклой функции $y = f(x)$ (рис. 34), и пусть $OA = x_1$, $OB = x_2$, $AM = f(x_1)$, $BN = f(x_2)$. Согласно доказанному выше координаты точки Q, делящей отрезок MN в отношении $MQ : QN = q : p$, равны $px_1 + qx_2$ и $pf(x_1) + qf(x_2)$; таким образом, на рис. 34 $SQ = pf(x_1) + qf(x_2)$ и $SP = f(px_1 + qx_2)$ (ибо $OS = px_1 + qx_2$). Но в силу выпуклости функции $y = f(x)$ точка Q расположена н и ж е точки P; значит,

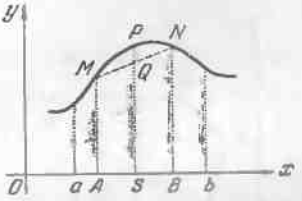


Рис. 34.

$$pf(x_1) + qf(x_2) < f(px_1 + qx_2),$$

— что нам и надо было доказать[2]).

Примеры.

а) $y = \log x$. В этом случае неравенство (2) дает

$$p \log x_1 + q \log x_2 < \log (px_1 + qx_2).$$

Отсюда следует, что

$$x_1^p x_2^q < px_1 + qx_2, \quad p + q = 1.$$

б) $y = -x^m$, $m > 1$. Имеем

$$-px_1^m - qx_2^m < -(px_1 + qx_2)^m$$

---

[1]) На рис. 33 изображен случай, когда все четыре числа x_1, x_2, y_1 и y_2 положительны (по существу только этот случай нам и будет нужен). Предоставим читателю самостоятельно рассмотреть иные случаи.

[2]) Нетрудно видеть, что координаты к а ж д о й точки отрезка MN могут быть представлены в виде $(px_1 + qx_2, py_1 + qy_2)$, где $p > 0$, $q > 0$, $p + q = 1$. Таким образом, неравенство (2) утверждает, что вся хорда MN расположена ниже кривой $y = f(x)$, т. е. оно равносильно определению выпуклости функции.

или

$$px_1^m + qx_2^m > (px_1 + qx_2)^m, \quad p + q = 1.$$

в) $y = -x\log x$. Здесь получаем

$$-px_1\log x_1 - qx_2\log x_2 < -(px_1 + qx_2)\log(px_1 + qx_2), \quad p + q = 1.$$

Теорему 1 можно обобщить еще и в другом направлении.

Т е о р е м а 3. *Если $y = f(x)$ — функция, выпуклая в интервале от a до b и $x_1, x_2, \ldots, x_k$ — какие-то k значений аргумента функции в этом интервале, не все равные между собой, то*

$$\frac{f(x_1) + f(x_2) + \cdots + f(x_k)}{k} < f\left(\frac{x_1 + x_2 + \cdots + x_k}{k}\right) \quad (3)$$

(частный случай неравенства Иенсена).

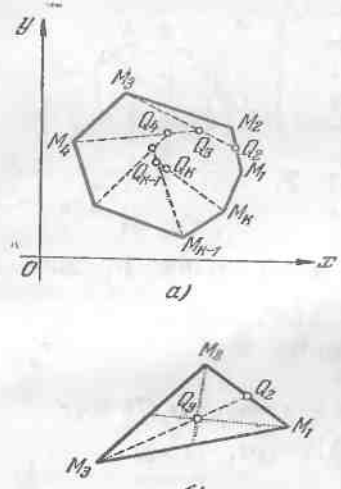


Рис. 35.

При $k = 2$ теорема 3 переходит в теорему 1.

Д о к а з а т е л ь с т в о. Начнем с определения одного понятия, часто фигурирующего в геометрических и аналитических задачах. Пусть $M_1M_2M_3 \ldots M_k$ — произвольный k-угольник (рис. 35, *а*); Q_2 — середина стороны M_1M_2 этого k-угольника ($M_1Q_2 : Q_2M_2 = \frac{1}{2} : \frac{1}{2}$); Q_3 — точка, делящая отрезок M_3Q_2 в отношении 2:1 ($M_3Q_3 : Q_3Q_2 = \frac{2}{3} : \frac{1}{3}$); Q_4 — точка, делящая отрезок M_4Q_3 в отношении 3:1 ($M_4Q_4 : Q_4Q_3 = \frac{3}{4} : \frac{1}{4}$); ...; наконец, Q_k — точка, делящая отрезок M_kQ_{k-1} в отношении $(k-1) : 1$ (т. е. такая, что $M_kQ_k : Q_kQ_{k-1} = \frac{k-1}{k} : \frac{1}{k}$).

Точка Q_k называется центроидом (или центром тяжести) k-угольника $M_1M_2 \ldots M_k$. В случае треугольника $M_1M_2M_3$ (рис. 35,*б*) центроид Q_3 совпадает с точкой пересечения медиан: действительно, в этом случае Q_2 есть середина стороны M_1M_2, отрезок M_3Q_2 является медианой и точка Q_3, делящая этот отрезок в отношении $M_3Q_3 : Q_3Q_2 = 2 : 1$— это точка пересечения медиан треугольника.

Докажем, что *если координаты вершин* $M_1, M_2, \ldots, M_k$ *k-угольника суть* $(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)$, *то координаты центроида* Q_k *будут равны* $\frac{x_1 + x_2 + \cdots + x_k}{k}$ *и* $\frac{y_1 + y_2 + \cdots + y_k}{k}$[1]). Действительно, в силу предложения, приведенного в начале доказательства теоремы 2, точки $Q_2, Q_3, Q_4, \ldots$, и, наконец, Q_k имеют следующие координаты:

$$Q_2\left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}\right),$$

$$Q_3\left(\frac{2}{3}\,\frac{x_1 + x_2}{2} + \frac{1}{3}x_3, \frac{2}{3}\,\frac{y_1 + y_2}{2} + \frac{1}{3}y_3\right)$$

или

$$\left(\frac{x_1 + x_2 + x_3}{3}, \frac{y_1 + y_2 + y_3}{3}\right),$$

$$Q_4\left(\frac{3}{4}\,\frac{x_1 + x_2 + x_3}{3} + \frac{1}{4}x_4, \frac{3}{4}\,\frac{y_1 + y_2 + y_3}{3} + \frac{1}{4}y_1\right)$$

или

$$\left(\frac{x_1 + x_2 + x_3 + x_4}{4}, \frac{y_1 + y_2 + y_3 + y_4}{4}\right),$$

. . . . . .

$$Q_k\left(\frac{(k-1)}{k}\,\frac{x_1 + x_2 + \cdots + x_{k-1}}{k-1} + \frac{1}{k}x_k,\right.$$

$$\left.\frac{(k-1)}{k}\,\frac{y_1 + y_2 + \cdots + y_{k-1}}{k-1} + \frac{1}{k}y_k\right)$$

---

[1]) Отсюда, в частности, следует, что центроид k-угольника полностью определяется этим k-угольником и не зависит от порядка перечисления его вершин (как можно было бы думать, исходя из определения центроида); в случае треугольника это обстоятельство вытекает также из совпадения центроида с точкой пересечения медиан.

или

$$\left(\frac{x_1+x_2+\cdots+x_{k-1}+x_k}{k},\ \frac{y_1+y_2+\cdots+y_{k-1}+y_k}{k}\right).$$

Вернемся теперь к нашей выпуклой функции $y=f(x)$. Пусть M_1, M_2, ..., M_k — это k последовательных точек графика этой функции, взятых в рассматриваемом интервале (рис. 36). В силу выпуклости функции k-угольник $M_1M_2\ldots M_k$ будет выпуклым и будет лежать целиком под кривой $y=f(x)$. Если абсциссы точек M_1, M_2, ..., M_k равны $x_1, x_2, \ldots, x_k$, то ординаты их, очевидно, будут равны $f(x_1)$, $f(x_2)$, ..., $f(x_k)$. Поэтому координаты центроида Q k-угольника $M_1M_2\ldots M_k$ будут равны

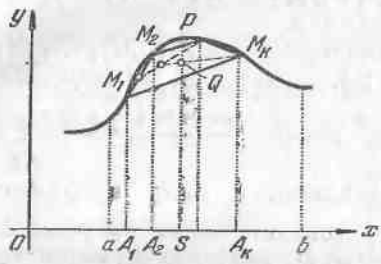


Рис. 36.

$\frac{x_1+x_2+\cdots+x_k}{k}$ и $\frac{f(x_1)+f(x_2)+\cdots+f(x_k)}{k}$, следовательно,

$$OS=\frac{x_1+x_2+\cdots+x_k}{k},\quad SQ=\frac{f(x_1)+f(x_2)+\cdots+f(x_k)}{k}$$

и

$$SP=f\left(\frac{x_1+x_2+\cdots+x_k}{k}\right)$$

(см. рис. 36). Но центроид выпуклого k-угольника лежит в н у т р и k-угольника (это вытекает из самого определения центроида); следовательно, точка Q расположена н и ж е точки P и, значит,

$$\frac{f(x_1)+f(x_2)+\cdots+f(x_k)}{k}<f\left(\frac{x_1+x_2+\cdots+x_k}{k}\right),$$

—что и требовалось доказать.

Это рассуждение сохраняет свою силу и в том случае, когда некоторые (но не все!) из точек M_1, M_2, ..., M_k совпадают (некоторые из чисел $x_1, x_2, \ldots, x_k$ равны между собой) и k-угольник $M_1M_2\ldots M_k$ вырождается в многоугольник с меньшим числом вершин.

Примеры.

а) $y = \log x$. Из теоремы 3 следует, что

$$\frac{\log x_1 + \log x_2 + \cdots + \log x_k}{k} < \log \frac{x_1 + x_2 + \cdots + x_k}{k}$$

или

$$\sqrt[k]{x_1 x_2 \ldots x_k} < \frac{x_1 + x_2 + \cdots + x_k}{k}$$

— *среднее геометрическое k положительных чисел, которые не все равны между собой, меньше их среднего арифметического* (теорема о среднем геометрическом и среднем арифметическом).

б) $y = -x^m$, $m > 1$. В таком случае получаем

$$-\frac{x_1^m + x_2^m + \cdots + x_k^m}{k} < -\left(\frac{x_1 + x_2 + \cdots + x_m}{k}\right)^m$$

или

$$\left(\frac{x_1^m + x_2^m + \cdots + x_k^m}{k}\right)^{\frac{1}{m}} > \frac{x_1 + x_2 + \cdots + x_k}{k}$$

— *степенное среднее порядка $m > 1$ произвольных k положительных чисел, которые не все равны между собой, больше их среднего арифметического.*

в) $y = -x \log x$. В этом случае теорема 3 дает

$$-\frac{x_1 \log x_1 + x_2 \log x_2 + \cdots + x_k \log x_k}{k} <$$

$$< -\frac{x_1 + x_2 + \cdots + x_k}{k} \log\left(\frac{x_1 + x_2 + \cdots + x_k}{k}\right). \quad (4)$$

Наконец, докажем еще следующую теорему, обобщающую как теорему 2, так и теорему 3:

Теорема 4. *Пусть $y = f(x)$ — функция, выпуклая в интервале от a до b, а $x_1, x_2, \ldots, x_k$ — какие-то k значений аргумента этой функции, не все равные между собой, взятые в рассматриваемом интервале, и $p_1, p_2, \ldots, p_k$ — k положительных чисел, сумма которых равна единице. В таком случае*

$$p_1 f(x_1) + p_2 f(x_2) + \ldots + p_k f(x_k) <$$

$$< f(p_1 x_1 + p_2 x_2 + \ldots + p_k x_k) \quad (5)$$

(общее неравенство Иенсена).

При $k = 2$ теорема 4 переходит в теорему 2, а при $p_1 = p_2 = \ldots = p_k = \frac{1}{k}$ — в теорему 3.

Доказательство. Рассмотрим снова график выпуклой функции $y = f(x)$ и вписанный в этот график выпуклый k-угольник $M_1M_2 \ldots M_k$, вершины которого

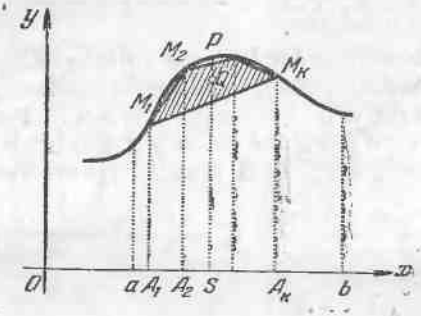


Рис. 37.

имеют координаты (x_1, y_1), (x_2, y_2), ..., (x_k, y_k) (рис. 37). Пусть теперь Q_2 — такая точка стороны M_1M_2 этого k-угольника, что $M_1Q_2 : Q_2M_2 = \frac{p_2}{p_1 + p_2} : \frac{p_1}{p_1 + p_2}$; Q_3 — такая точка отрезка M_3Q_2, что

$$M_3Q_3 : Q_3Q_2 = \frac{p_3}{p_1 + p_2 + p_3} : \frac{p_1 + p_2}{p_1 + p_2 + p_3};$$

Q_4 — такая точка отрезка M_4Q_3, что

$$M_4Q_4 : Q_4Q_3 = \frac{p_4}{p_1 + p_2 + p_3 + p_4} : \frac{p_1 + p_2 + p_3}{p_1 + p_2 + p_3 + p_4}; \ldots;$$

наконец, Q — такая точка отрезка M_kQ_{k-1}, что $M_kQ : QQ_{k-1} = p_k : (p_1 + p_2 + \ldots + p_{k-1})$ (если $p_1 = p_2 = \ldots = p_k = 1/k$ то Q — центроид k-угольника $M_1M_2 \ldots M_k$). Воспользовавшись предложением, с которого мы начали доказательство теоремы 2, найдем координаты точек Q_2, Q_3, Q_4, ..., Q:

$$Q_2\left(\frac{p_1x_1 + p_2x_2}{p_1 + p_2}, \frac{p_1 f(x_1) + p_2 f(x_2)}{p_1 + p_2}\right),$$

$$Q_3\left(\frac{p_1 + p_2}{p_1 + p_2 + p_3} \frac{p_1x_1 + p_2x_2}{p_1 + p_2} + \frac{p_3}{p_1 + p_2 + p_3} x_3,\right.$$

$$\left.\frac{p_1 + p_2}{p_1 + p_2 + p_3} \frac{p_1 f(x_1) + p_2 f(x_2)}{p_1 + p_2} + \frac{p_3}{p_1 + p_2 + p_3} f(x_3)\right)$$

или

$$\left(\frac{p_1x_1+p_2x_2+p_3x_3}{p_1+p_2+p_3},\ \frac{p_1f(x_1)+p_2f(x_2)+p_3f(x_3)}{p_1+p_2+p_3}\right),$$

$$Q_4\left(\frac{p_1+p_2+p_3}{p_1+p_2+p_3+p_4}\,\frac{p_1x_1+p_2x_2+p_3x_3}{p_1+p_2+p_3}+\frac{p_4}{p_1+p_2+p_3+p_4}x_4,\right.$$
$$\left.\frac{p_1+p_2+p_3}{p_1+p_2+p_3+p_4}\,\frac{p_1f(x_1)+p_2f(x_2)+p_3f(x_3)}{p_1+p_2+p_3}+\right.$$
$$\left.+\frac{p_4}{p_1+p_2+p_3+p_4}f(x_4)\right)$$

или

$$\left(\frac{p_1x_1+p_2x_2+p_3x_3+p_4x_4}{p_1+p_2+p_3+p_4},\ \frac{p_1f(x_1)+p_2f(x_2)+p_3f(x_3)+p_4f(x_4)}{p_1+p_2+p_3+p_4}\right),$$

. . . . . .

$$Q\left(\frac{p_1x_1+p_2x_2+\cdots+p_{k-1}x_{k-1}+p_kx_k}{p_1+p_2+\cdots+p_{k-1}+p_k},\right.$$
$$\left.\frac{p_1f(x_1)+p_2f(x_2)+\cdots+p_{k-1}f(x_{k-1})+p_kf(x_k)}{p_1+p_2+\cdots+p_{k-1}+p_k}\right)$$

или иначе,

$$(p_1x_1+p_2x_2+\ldots+p_kx_k,\ p_1f(x_1)+p_2f(x_2)+\ldots$$
$$\ldots+p_kf(x_k))$$

(так как $p_1+p_2+\ldots+p_k=1$).

Таким образом на рис. 37

$$SQ=p_1f(x_1)+p_2f(x_2)+\ldots+p_kf(x_k),$$
$$OS=p_1x_1+p_2x_2+\ldots+p_kx_k,$$
$$SP=f(p_1x_1+p_2x_2+\ldots+p_kx_k).$$

А так как точка Q расположена н и ж е точки P (ибо весь k-угольник $M_1M_2\ldots M_k$ лежит под кривой $y=f(x)$, а Q — в н у т р е н н я я точка этого k-угольника), то

$$p_1f(x_1)+p_2f(x_2)+\ldots+p_kf(x_k)<$$
$$<f(p_1x_1+p_2x_2+\ldots+p_kx_k),$$

— что и требовалось доказать [1]).

[1]) Нетрудно видеть, что координаты к а ж д о й внутренней точки k-угольника $M_1M_2\ldots M_k$ можно представить в виде $(p_1x_1+p_2x_2+\ldots+p_kx_k,\ p_1f(x_1)+p_2f(x_2)+\ldots+p_kf(x_k)))$, где $p_1>0,\ p_2>0,\ \ldots,\ p_k>0$, и $p_1+p_2+\ldots+p_k=1$. Таким образом, неравенство (5) выражает то обстоятельство, что вписанный в график выпуклой функции многоугольник весь лежит н и ж е этого графика.

Примеры.

а) $y = \log x$. В таком случае получаем:

$$p_1 \log x_1 + p_2 \log x_2 + \ldots + p_k \log x_k < \log (p_1 x_1 + p_2 x_2 + \ldots + p_k x_k),$$

откуда следует, что

$$x_1^{p_1} x_2^{p_2} \ldots x_k^{p_k} < p_1 x_1 + p_2 x_2 + \ldots + p_k x_k,$$

где $p_1 + p_2 + \ldots + p_k = 1$

(обобщенная теорема о среднем геометрическом и среднем арифметическом).

б) $y = -x^m$, $m > 1$. Имеем

$$-p_1 x_1^m - p_2 x_2^m - \ldots - p_k x_k^m < -(p_1 x_1 + p_2 x_2 + \ldots + p_k x_k)^m$$

или

$$p_1 x_1^m + p_2 x_2^m + \ldots + p_k x_k^m > (p_1 x_1 + p_2 x_2 + \ldots + p_k x_k)^m,$$

где $p_1 + p_2 + \ldots + p_k = 1$.

в) $y = -x \log x$. Теорема 4 дает

$$-p_1 x_1 \log x_1 - p_2 x_2 \log x_2 - \ldots - p_k x_k \log x_k < -(p_1 x_1 + p_2 x_2 + \ldots + p_k x_k) \log (p_1 x_1 + p_2 x_2 + \ldots + p_k x_k),$$

где $p_1 + p_2 + \ldots + p_k = 1$. (6)

Вывод неравенств (4) (стр. 449) и (6) и составлял основную цель настоящего Приложения. Из неравенства (4) сразу следует, что энтропия опыта α, имеющего k исходов, не превосходит энтропии $\log k$ опыта α_0, имеющего k равновероятных исходов; при этом $H(\alpha) = \log k$ лишь в том случае, когда все исходы α равновероятны, т. е. когда опыт α не отличается от α_0. Действительно, умножим обе части этого неравенства на k и затем положим в нем $x_1 = p(A_1)$, $x_2 = p(A_2)$, ..., $x_k = p(A_k)$, где $A_1, A_2, \ldots, A_k$ — исходы опыта α (так что $p(A_1) + p(A_2) + \ldots + p(A_k) = 1$; вероятности $p(A_1), p(A_2), \ldots, p(A_k)$ не все равны между собой). В таком случае

будем иметь:

$$-p(A_1)\log p(A_1) - p(A_2)\log(A_2) - \cdots - p(A_k)\log p(A_k) <$$

$$< -[p(A_1) + p(A_2) + \cdots + p(A_k)] \times$$

$$\times \log \frac{p(A_1) + p(A_2) + \cdots + p(A_k)}{k} = -1 \cdot \log \frac{1}{k} = \log k$$

или

$$H(\alpha) < H(\alpha_0).$$

Неравенство (6) может быть использовано для доказательства того, что условная энтропия $H_\alpha(\beta)$ опыта β при условии α не превосходит безусловной энтропии $H(\beta)$ того же опыта. В самом деле, полагая в неравенстве (6) $p_1 = p(A_1)$, $p_2 = p(A_2)$, ..., $p_k = p(A_k)$, $x_1 = p_{A_1}(B_1)$, $x_2 = p_{A_2}(B_1)$, ... $x_k = p_{A_k}(B_1)$ (где $A_1, A_2, \ldots, A_k; B_1, B_2, \ldots$ ..., B_l — исходы опытов α и β; $p(A_1) + p(A_2) + \ldots$ ... $+ p(A_k) = 1$), мы получим

$$-p(A_1)p_{A_1}(B_1)\log p_{A_1}(B_1) - p(A_2)p_{A_2}(B_1)\log p_{A_2}(B_1) - \ldots$$

$$\ldots - p(A_k)p_{A_k}(B_1)\log p_{A_k}(B_1) <$$

$$< -[p(A_1)p_{A_1}(B_1) + p(A_2)p_{A_2}(B_1) + \ldots + p(A_k)p_{A_k}(B_1)] \times$$
$$\times \log [p(A_1)p_{A_1}(B_1) + p(A_2)p_{A_2}(B_1) + \ldots + p(A_k)p_{A_k}(B_1)].$$

Так как в силу формулы полной вероятности (см. выше, стр. 44)

$$p(A_1)p_{A_1}(B_1) + p(A_2)p_{A_2}(B_1) + \ldots + p(A_k)p_{A_k}(B_1) = p(B_1),$$

то последнее неравенство можно переписать так:

$$-p(A_1)p_{A_1}(B_1)\log p_{A_1}(B_1) - p(A_2)p_{A_2}(B_1)\log p_{A_2}(B_1) - \ldots$$

$$\ldots - p(A_k)p_{A_k}(B_1)\log p_{A_k}(B_1) < -p(B_1)\log p(B_1).$$

Заметим, что если $p_{A_1}(B_1) = p_{A_2}(B_1) = \ldots = p_{A_k}(B_1) = p(B_1)$ (последнее равенство здесь следует из формулы полной вероятности), то наше неравенство обращается в

равенство. Точно так же получаем

$$-p(A_1)\,p_{A_1}(B_2)\log p_{A_1}(B_2) - p(A_2)\,p_{A_2}(B_2)\log p_{A_2}(B_2) - \ldots$$
$$\ldots - p(A_k)p_{A_k}(B_2)\log p_{A_k}(B_2) < -p(B_2)\log p(B_2),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$-p(A_1)\,p_{A_1}(B_l)\log p_{A_1}(B_l) - p(A_2)\,p_{A_2}(B_l)\log p_{A_2}(B_l) - \ldots$$
$$\ldots - p(A_k)p_{A_k}(B_l)\log p_{A_k}(B_l) < -p(B_l)\log p(B_l).$$

Теперь, складывая все эти неравенства, имеем

$$p(A_1)H_{A_1}(\beta) + p(A_2)\,H_{A_2}(\beta) + \ldots + p(A_k)H_{A_k}(\beta) < H(\beta)$$

или

$$H_\alpha(\beta) < H(\beta).$$

Это неравенство имеет место, если опыты α и β не независимы, т. е. если существуют такие i и j $(1 \leqslant i \leqslant k,\ 1 \leqslant j \leqslant l)$, что $p_{A_i}(B_j) \neq p(B_j)$. Если же опыты α и β независимы, то, очевидно, $H_\alpha(\beta) = H(\beta)$.

Заметим еще, что если подставить в неравенство

$$p_1\log x_1 + p_2\log x_2 + \ldots + p_k\log x_k \leqslant \log(p_1x_1 + p_2x_2 + \ldots + p_kx_k),$$
$$p_1 + p_2 + \ldots + p_k = 1$$

(ср. пример а) на стр. 452), $x_1 = \dfrac{q_1}{p_1}$, $x_2 = \dfrac{q_2}{p_2}$, $\ldots$, $x_k = \dfrac{q_k}{p_k}$, где $q_1 + q_2 + \ldots + q_k \leqslant 1$, то мы получим

$$p_1\log\frac{q_1}{p_1} + p_2\log\frac{q_2}{p_2} + \ldots + p_k\log\frac{q_k}{p_k} \leqslant$$
$$\leqslant \log(q_1 + q_2 + \ldots + q_k) \leqslant \log 1 = 0,$$

откуда, учитывая, что $\log\dfrac{q_1}{p_1} = \log q_1 - \log p_1$, $\log\dfrac{q_2}{p_2} = \log q_2 - \log p_2$, $\ldots$, $\log\dfrac{q_k}{p_k} = \log q_k - \log p_k$, имеем

$$-p_1\log p_1 - p_2\log p_2 - \ldots - p_k\log p_k \leqslant$$
$$\leqslant -p_1\log q_1 - p_2\log q_2 - \ldots - p_k\log q_k,$$

т. е. неравенство (*), стр. 180.

Наконец, упомянем еще об обобщающем неравенство $H_\alpha(\beta) \leqslant H(\beta)$ неравенстве

$$H_{\alpha\gamma}(\beta) \leqslant H_\gamma(\beta),$$

о котором шла речь в конце § 3 гл. II (это неравенство переходит в $H_\alpha(\beta) \leqslant H(\beta)$, если предположить, что опыт γ имеет единственный исход, реализующийся с вероятностью 1). Его легко вывести из неравенства $H_\alpha(\beta) \leqslant H(\beta)$. Действительно, обозначим исходы опыта γ через $C_1, C_2, \ldots, C_m$; пусть $\alpha^{(1)}$ и $\beta^{(1)}$ — опыты, исходы $A_1^{(1)}, A_2^{(1)}, \ldots, A_k^{(1)}$ и $B_1^{(1)}, B_2^{(1)}, \ldots, B_l^{(1)}$, которых осуществляются с вероятностями $p(A_1^{(1)}) = p_{C_1}(A_1)$, $p(A_2^{(1)}) = p_{C_1}(A_2)$, ..., $p(A_k^{(1)}) = p_{C_1}(A_k)$, соответственно $p(B_1^{(1)}) = p_{C_1}(B_1)$, $p(B_2^{(1)}) = p_{C_1}(B_2)$, $\ldots$, $p(B_l^{(1)}) = p_{C_1}(B_l)$. В силу доказанного выше имеем

$$H_{\alpha^{(1)}}(\beta^{(1)}) \leqslant H(\beta^{(1)}).$$

Но

$$H(\beta^{(1)}) = -p(B_1^{(1)}) \log p(B_1^{(1)}) - p(B_2^{(1)}) \log p(B_2^{(1)}) - \ldots - p(B_l^{(1)}) \log p(B_l^{(1)}) =$$
$$= -p_{C_1}(B_1) \log p_{C_1}(B_1) - p_{C_1}(B_2) \log p_{C_1}(B_2) - \ldots$$
$$\ldots - p_{C_1}(B_l) \log p_{C_1}(B_l) = H_{C_1}(\beta)$$

и

$$H_{\alpha^{(1)}}(\beta^{(1)}) = p(A_1^{(1)}) H_{A_1^{(1)}}(\beta^{(1)}) + p(A_2^{(1)}) H_{A_2^{(1)}}(\beta^{(1)}) + \ldots$$
$$\ldots + p(A_k^{(1)}) H_{A_k^{(1)}}(\beta^{(1)}),$$

где

$$H_{A_1^{(1)}}(\beta^{(1)}) = -p_{A_1^{(1)}}(B_1^{(1)}) \log p_{A_1^{(1)}}(B_1^{(1)}) -$$
$$- p_{A_1^{(1)}}(B_2^{(1)}) \log p_{A_1^{(1)}}(B_2^{(1)}) - \ldots - p_{A_1^{(1)}}(B_l^{(1)}) \log p_{A_1^{(1)}}(B_l^{(1)}),$$

$$H_{A_2^{(1)}}(\beta^{(1)}) = -p_{A_2^{(1)}}(B_1^{(1)}) \log p_{A_2^{(1)}}(B_1^{(1)}) -$$
$$- p_{A_2^{(1)}}(B_2^{(1)}) \log p_{A_2^{(1)}}(B_2^{(1)}) - \ldots - p_{A_2^{(1)}}(B_l^{(1)}) \log p_{A_2^{(1)}}(B_l^{(1)}),$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$H_{A_k^{(1)}}(\beta^{(1)}) = -p_{A_k^{(1)}}(B_1^{(1)}) \log p_{A_k^{(1)}}(B_1^{(1)}) -$$
$$- p_{A_k^{(1)}}(B_2^{(1)}) \log p_{A_k^{(1)}}(B_2^{(1)}) - \ldots - p_{A_k^{(1)}}(B_l^{(1)}) \log p_{A_k^{(1)}}(B_l^{(1)}).$$

Найдем теперь условные вероятности $p_{A_1^{(1)}}(B_1^{(1)})$, $p_{A_1^{(1)}}(B_2^{(1)})$ и т. д. В силу правила умножения вероятностей (см. § 3 гл. I, стр. 42) $p_{A_1^{(1)}}(B_1^{(1)})$ равно отношению вероятностей событий $A_1^{(1)}B_1^{(1)}$ и $A_1^{(1)}$. Но $p(A_1^{(1)}) = p_{C_1}(A_1)$; что же касается вероятности события $A_1^{(1)}B_1^{(1)}$, то она, очевидно, равна условной вероятности $p_{C_1}(A_1B_1)$ ($A_1^{(1)}$ — осуществление события A_1 при том условии, что произошло событие C_1, $B_1^{(1)}$ — осуществление события B_1 при том же самом условии; поэтому $A_1^{(1)}B_1^{(1)}$ это есть осуществление A_1B_1 при том же условии). Но в силу правила умножения вероятностей $p_{C_1}(A_1B_1) = p_{C_1}(A_1)p_{C_1A_1}(B_1)$; следовательно,

$$p_{A_1^{(1)}}(B_1^{(1)}) = \frac{p(A_1^{(1)}B_1^{(1)})}{p(A_1^{(1)})} = \frac{p_{C_1}(A_1)\,p_{C_1A_1}(B_1)}{p_{C_1}(A_1)} = p_{C_1A_1}(B_1).$$

Точно так же показывается, что

$$p_{A_1^{(1)}}(B_2^{(1)}) = p_{C_1A_1}(B_2),\ p_{A_1^{(1)}}(B_3^{(1)}) = p_{C_1A_1}(B_3)\ldots$$

$$\ldots,\ p_{A_k^{(1)}}(B_l^{(1)}) = p_{C_1A_k}(B_l).$$

Отсюда получаем

$$H_{A_1^{(1)}}(\beta^{(1)}) = -\,p_{C_1A_1}(B_1)\log p_{C_1A_1}(B_1) - p_{C_1A_1}(B_2)\log p_{C_1A_1}(B_2) - \ldots$$

$$\ldots - p_{C_1A_1}(B_l)\log p_{C_1A_1}(B_l) = H_{C_1A_1}(\beta)$$

и аналогично

$$H_{A_2^{(1}}(\beta^{(1)}) = H_{C_1A_2}(\beta),\ H_{A_3^{(1)}}(\beta^{(1)}) = H_{C_1A_3}(\beta),\ \ldots$$

$$\ldots,\ H_{A_k^{(1)}}(\beta^{(1)}) = H_{C_1A_k}(\beta).$$

Таким образом, имеем (напоминаем, что $p(A_1^{(1)}) = p_{C_1}(A_1)$, $p(A_2^{(1)}) = p_{C_1}(A_2)$, $\ldots$, $p(A_k^{(1)}) = p_{C_1}(A_k)$)

$$H_\alpha^{(1)}(\beta^{(1)}) = p_{C_1}(A_1)\,H_{C_1A_1}(\beta) + p_{C_1}(A_2)\,H_{C_1A_2}(\beta) + \ldots$$

$$\ldots + p_{C_1}(A_k)\,H_{C_1A_k}(\beta).$$

Поэтому неравенство $H_{\alpha^{(1)}}(\beta^{(1)}) \leqslant H(\beta^{(1)})$ можно записать в виде

$$p_{C_1}(A_1)\,H_{C_1A_1}(\beta) + p_{C_1}(A_2)H_{C_1A_2}(\beta) + \ldots + p_{C_1}(A_k)\,H_{C_1A_k}(\beta) \leqslant H_{C_1}(\beta).$$

Умножая обе его части на $p(C_1)$ и учитывая, что $p(C_1)\,p_{C_1}(A_1) = p(C_1A_1)$, $p(C_1)\,p_{C_1}(A_2) = p(C_1A_2)$, . . ., $p(C_1)\,p_{C_1}(A_k) = p(C_1A_k)$, будем иметь

$$p(C_1A_1)\,H_{C_1A_1}(\beta) + p(C_1A_2)\,H_{C_1A_2}(\beta) + \ldots \\ \ldots + p(C_1A_k)\,H_{C_1A_k}(\beta) \leqslant p(C_1)\,H_{C_1}(\beta).$$

Точно так же доказываются неравенства

$$p(C_2A_1)\,H_{C_2A_1}(\beta) + p(C_2A_2)\,H_{C_2A_2}(\beta) + \ldots \\ \ldots + p(C_2A_k)\,H_{C_2A_k}(\beta) \leqslant p(C_2)\,H_{C_2}(\beta),$$

. . . . . .

$$p(C_mA_1)\,H_{C_mA_1}(\beta) + p(C_mA_2)\,H_{C_mA_2}(\beta) + \ldots \\ \ldots + p(C_mA_k)\,H_{C_mA_k}(\beta) \leqslant p(C_m)\,H_{C_m}(\beta).$$

Складывая почленно все эти неравенства, получим

$$H_{\gamma\alpha}(\beta) \leqslant H_\gamma(\beta),$$

что и требовалось доказать (события $\gamma\alpha$ и $\alpha\gamma$ не различаются).

Приложение II

НЕКОТОРЫЕ АЛГЕБРАИЧЕСКИЕ ПОНЯТИЯ

Основным предметом изучения в алгебре являются те или иные алгебраические системы, т. е. множества элементов, для которых определены некоторые *алгебраические операции*, подобные известным из арифметики операциям сложения и умножения чисел. При этом характер элементов системы и конкретный смысл рассматриваемых операций обычно никак не оговариваются, так что одна и та же алгебраическая схема может описывать весьма разнородные примеры. Напротив, свойства алгебраических операций подробно описываются — и это описание является определением соответствующей алгебраической системы.

1. Первым алгебраическим понятием, широко используемым в самых разных разделах математики, является понятие (коммутативной[1]) группы.

Множество G элементов a, b, c и т. д. называется (коммутативной) группой, если в этом множестве определена операция $\circ$, сопоставляющая каждым двум элементам a и b нашего множества единственный третий элемент, обозначаемый символом $a \circ b$, причем

1° *операция $\circ$ коммутативна:*

$$a \circ b = b \circ a \text{ для любых } a \text{ и } b \text{ из } G;$$

2° *операция $\circ$ ассоциативна:*

$$(a \circ b) \circ c = a \circ (b \circ c) \text{ для любых } a,\ b \text{ и } c \text{ из } G;$$

3° *в множестве G существует такой элемент e, что*

$$a \circ e = a \text{ для всех } a \text{ из } G;$$

---

[1]) В алгебре рассматриваются также и некоммутативные группы, для которых сформулированное ниже условие 1° не имеет места; однако так как в этой книге встречаются лишь коммутативные группы, то мы позволили себе, в отступление от традиции, включить условие 1° в определение группы.

4° для каждого элемента a из G существует такой элемент a^*, что

$$a \circ a^* = e.$$

Групповую операцию $\circ$ иногда обозначают знаком $+$; при этом элемент $a + b$ называют с у м м о й элементов a и b; элемент e такой, что

$$a + e = a \text{ для всех } a,$$

называют н у л е в ы м э л е м е н т о м или просто н у л е м группы и зачастую обозначают символом 0; элемент a^* такой, что

$$a + a^* = 0,$$

называют п р о т и в о п о л о ж н ы м к a и обозначают через $-a$. Можно также результат $a \circ b$ применения к элементам a и b групповой операции $\circ$ обозначить через $a \cdot b$ или через ab; в таком случае $ae = a$ для всех a, и поэтому e называют е д и н и ч н ы м э л е м е н т о м или е д и н и ц е й группы и иногда обозначают символом 1; далее, $aa^* = 1$, и поэтому a^* называется о б р а т н ы м к a и обозначается через a^{-1}. Мы в дальнейшем всегда будем обозначать групповую операцию знаком $+$; при этом через $a - b$ обозначается такой элемент x (р а з н о с т ь элементов a и b), что $x + b = a$ (нетрудно видеть, что такой элемент x всегда существует: он равен $a + (-b)$).

П р и м е р ы.

А. *Множество целых чисел* (или *рациональных чисел*, или *вещественных чисел*) *образует г р у п п у п о с л о ж е н и ю*; другими словами, соответствующее множество, где за групповую операцию принято (обыкновенное!) сложение, образует группу с нулем 0 и противоположным a элементом $-a$.

Б. Примем за групповую операцию (которую мы теперь будем обозначать знаком «$+$», чтобы подчеркнуть, что это н е е с т ь обыкновенное сложение) *умножение* чисел. При этом множество целых чисел уже не будет образовывать группы, поскольку здесь, очевидно, не выполнено условие 4°: ведь целое число a^* такое, что a «$+$» $a^* = aa^* = 1$, существует, только если $a = 1$ или $a = -1$.

Также и множество всех рациональных чисел не образует группы по умножению, поскольку здесь условие 4° нарушается при $a = 0$. Однако *множество всех отличных от* 0 (или *всех положительных*) *рациональных чисел* (или *вещественных чисел*) *уже образует группу по умножению.*

В. Рассмотрим снова множество всех целых чисел и определенную в этом множестве операцию сложения чисел. Выберем теперь какое-то целое положительное число q и условимся *заменять каждое число A остатком a от деления A на q*; так, например, если $q = 10$, то мы условимся оставлять у каждого целого положительного числа A только его последнюю цифру a (это и есть остаток от деления A на 10). Множество всевозможных *остатков от деления целых чисел на q*, состоящее из q чисел $0, 1, 2, \ldots, q-1$, мы назовем «*q-арифметикой*»; суммой же элементов a и b q-арифметики мы назовем остаток от деления обычной суммы $a + b$ на q (равный $a + b$, если $a + b < q$). Вот как выглядят, например, «таблицы сложения» в 2-арифметике, 5-арифметике и 6-арифметике:

+	0	1
0	0	1
1	1	0

+	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

+	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

Легко видеть, что *q-арифметика по отношению к определенному в ней сложению представляет собой группу из q элементов* (или, как говорят, группу порядка q); нулевым элементом этой группы является число 0, а противоположным числу $a \neq 0$ будет число $q - a$ (ибо сумма $a + (q - a)$ при делении на q дает остаток 0). Для 2-арифметики, очевидно, противоположным для каждого числа a (т. е. и для $a = 0$ и для $a = 1$) будет оно само: здесь всегда $-a = a$.

Г. Пусть G — *какая-угодно* группа, например, группа целых чисел по сложению или группа сложения чисел в q-арифметике. Рассмотрим теперь произвольную

таблицу из m строк и n столбцов или $(m \times n)$-*матрицу*

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix},$$

составленную из элементов группы G, которые мы далее будем называть *числами*. Ясно, что, условившись складывать матрицы «поэлементно» (т. е. считая, что число, стоящее на некотором месте матрицы-суммы, равно сумме чисел, стоящих на тех же местах в матрицах-слагаемых), мы придем к *группе* $(m \times n)$-*матриц по сложению*; нулевым элементом этой группы служит нулевая матрица O, составленная из одних нулей.

$(1 \times n)$-матрицы называют также *векторами* (или *векторами-строками*); аналогично этому $(m \times 1)$-матрицы называют *векторами-столбцами*. Разумеется, векторы с одним и тем же числом элементов строки (или столбца) также можно складывать между собой; если элементы векторов принадлежат какой-то группе («группе чисел»), то и *векторы образуют группу по сложению*. Векторы чаще всего обозначают малыми латинскими буквами жирного шрифта; «нулевой вектор» (т. е. строку или столбец из одних нулей) иногда обозначают жирной цифрой 0.

Если группа G «чисел» является бесконечной, то бесконечной будет и группа $(m \times n)$-матриц (и частности, векторов), которые строятся из этих «чисел». Если же группа G имеет конечный порядок q, то группа $(m \times n)$-матриц будет иметь порядок q^{mn}: ведь матрица имеет mn элементов, вместо каждого из которых можно подставить любой из q элементов группы G. Аналогично группа векторов-строк из n элементов и группа векторов-столбцов из m элементов будет иметь конечный порядок q^n, соответственно, q^m, если основная группа G имеет порядок q.

Д. Рассмотрим произвольный многочлен

$$f(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_{n-1} x^{n-1},$$

коэффициенты $a_0, a_1, \dots, a_{n-1}$ которого являются элементами произвольно выбранной группы G. Если $g(x)$ —

другой многочлен

$$g(x) = b_0 + b_1x + b_2x^2 + \ldots + b_{n-1}x^{n-1}$$

(мы считаем, что $f(x)$ и $g(x)$ имеют одну и ту же степень, ибо в противном случае всегда можно дополнить тот из них, степень которого ниже, несколькими «старшими» членами с коэффициентами 0 при них), то можно определить с у м м у многочленов

$$f(x) + g(x) = (a_0 + b_0) + (a_1 + b_1)x + (a_2 + b_2)x^2 + \ldots \\ \ldots + (a_{n-1} + b_{n-1})\,x^{n-1}.$$

Легко видеть, что *многочлены* с определенной таким образом операцией сложения *образуют группу*; эта группа всегда будет бесконечной, ибо степень многочлена может быть сколь угодно большой. Роль нулевого элемента этой группы играет, очевидно, «нулевой» многочлен 0, все коэффициенты которого равны нулю; противоположным $f(x)$ будет многочлен $-f(x)$, все коэффициенты которого противоположны коэффициентам $f(x)$.

Если мы ограничимся *многочленами степени ниже n*, где n — какое-то фиксированное число, то мы также получим *группу*; она, как легко видеть, отличается от группы векторов

$$f = (a_0, a_1, a_2, \ldots, a_{n-1})$$

лишь формой записи элементов группы. Эта группа уже будет конечной, если конечной является группа G; если порядок группы G равен q, то порядок группы многочленов степени $< n$ равен q^n. Так, например, имеется всего $2^2 = 4$ многочлена степени < 2 с коэффициентами из 2-арифметики: 0,1, x и $x + 1$; «таблица сложения» этих многочленов такова

+	0	1	x	$x+1$
0	0	1	x	$x+1$
1	1	0	$x+1$	x
x	x	$x+1$	0	1
$x+1$	$x+1$	x	1	0

Пусть теперь G — произвольная группа и H — часть элементов этой группы. Если множество H элементов группы таково, что

1° *если a принадлежит H и b принадлежит H, то a + b тоже принадлежит H*;

2° *если* a *принадлежит* H, *то и* $-a$ *принадлежит* H;

3° *нулевой элемент* 0 *группы* G *принадлежит* H,

то множество H само образует группу относительно определенной в G операции сложения. В таком случае говорят, что H представляет собой подгруппу группы G.

В частности, если G — группа целых чисел по сложению, то совокупность H всех чисел, кратных фиксированному числу l, образует подгруппу группы G. Точно так же, если G — группа сложения чисел в q-арифметике и $q = kl$ — составное число, то совокупность H всех принадлежащих G чисел, делящихся на l (т. е. чисел l, $2l$, $3l$, ..., $(k-1)l$), образует подгруппу группы G (несущественно отличающуюся, как легко понять, от группы сложения чисел в k-арифметике).

Подгруппой группы $(m \times n)$-матриц по сложению является, например, группа всевозможных матриц, у которых все строки, кроме первой, состоят из одних лишь нулей (эта подгруппа, очевидно, лишь записью отличается от группы по сложению векторов-строк), а также группа матриц, у которых равны 0 все элементы, кроме какого-то одного фиксированного, — например, элемента a_{11}, стоящего в правом верхнем углу (эта подгруппа сводится к группе G, поскольку каждый ее элемент задается одним числом a_{11}). Укажем еще, что если G есть группа $(m \times n)$-матриц с элементами из «2-арифметики», то для того чтобы убедиться, что некоторая ее часть составляет подгруппу, достаточно проверить выполнение одного условия 1° (ибо в 2-арифметике каждое число обратно самому себе, а потому здесь $\boldsymbol{A} + \boldsymbol{A} = \boldsymbol{O}$ для каждой матрицы $\boldsymbol{A}$ и, значит, $-\boldsymbol{A} = \boldsymbol{A}$).

Подгруппой группы всех многочленов является группа многочленов степени $< n$; для этой же последней группы подгруппой является группа многочленов степени $< k$, где $k < n$, или группа многочленов, обращающихся в 0 при $x = 0$ (эти многочлены характеризуются равенством нулю «свободного члена» a_0).

2. Следующими по важности алгебраическими системами являются поля и кольца.

Множество F *элементов* a, b, *с и т. д. называется поле м, если в нем определены две операции, сопоставляющие двум элементам* a *и* b *поля третий элемент*; эти операции

можно назвать «с л о ж е н и е м» (и обозначать «сумму» элементов a и b поля через $a + b$) и «у м н о ж е н и е м» («произведение» элементов a и b естественно обозначить через ab). При этом:

I) *элементы поля должны образовывать группу по сложению*;

II) *отличные от нуля элементы поля должны образовывать группу по умножению*;

III) *сложение и умножение должны подчиняться дистрибутивному закону*:

$$(a + b)c = ac + bc \text{ для всех } a,\ b \text{ и } c.$$

Легко понять, что для любых элементов a и b поля F, где b отлично от нуля, существует их «частное» $\frac{a}{b}$, т. е. такое число y, что $by = a$: это y можно определить формулой $y = ab^{-1}$.

П р и м е р ы.

А. Ясно, что *множество всех рациональных* (или *вещественных*, или *комплексных*) *чисел образует поле* относительно обычных операций сложения и умножения.

Б. «Произведение чисел a и b q-арифметики» определим как *остаток от деления на q обычного произведения ab*; так, например, «произведение чисел a и b 10-арифметики» — это просто последняя цифра числа ab. Вот, например, как выглядят «таблицы умножения чисел» в 2-арифметике, в 5-арифметике и в 6-арифметике:

·	0	1
0	0	0
1	0	1

·	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

·	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

Сравнение этих таблиц позволяет усмотреть существенную разницу между ними: в то время как для 2-арифметики и 5-арифметики каждая строка таблицы, кроме первой строки, состоящей из одних нулей, содержит единицу, для 6-арифметики это будет уже не так (здесь не содержит единицы 3-я, 4-я и 5-я из шести строк таблицы). Таким образом, в 2-арифметике и в 5-арифметике *любое*

отличное от 0 *число имеет обратное* (в 2-арифметике имеем $1^{-1} = 1$; для 5-арифметики верны равенства $1^{-1} = 1$, $2^{-1} = 3$, $3^{-1} = 2$ и $4^{-1} = 4$); напротив, в 6-арифметике числа 2, 3 и 4 не имеют обратных. Отсюда легко следует, что 2-*арифметика* и 5-*арифметика* по отношению к определенным в них сложению и умножению *являются полями*, а 6-арифметика полем н е я в л я е т с я.

Нетрудно понять, что для любого с о с т а в н о г о $q = kl$ (где $k > 1$, $l > 1$) q-арифметика н е м о ж е т явиться полем: это вытекает, например, из того, что здесь $kl = 0$ (где умножение понимается в указанном выше смысле). Если же p — п р о с т о е число, то в p-арифметике каждое число имеет обратное (см. ниже, стр. 467); поэтому p-*арифметика* с определенными в ней действиями сложения и умножения чисел *представляет собой конечное поле* F_p из p элементов (или поле п о р я д к а p).

Обратимся теперь к случаю произвольной q-арифметики, где q, вообще говоря,— с о с т а в н о е число. В таком случае мы не получаем поля, поскольку не каждый элемент q-арифметики будет иметь обратный; все же остальные определяющие поле условия сохраняют силу и для этого случая.

Множество K *элементов* a, b, *с и т. д., в котором определены операции сложения и умножения, причем*

I) *элементы нашего множества образуют группу по сложению;*

II) *умножение элементов множества таково, что* $ab = ba$ *для всех* a *и* b; $(ab)c = a\,(bc)$ *для всех* a, b *и* c; *существует такой элемент* 1, *что* $a \cdot 1 = a$ *для всех* a;

III) *сложение и умножение подчиняются дистрибутивному закону:*

$$(a + b)\,c = ac + bc \quad \text{для всех } a,\ b \text{ и } c,$$

называется (*коммутативным*) к о л ь ц о м [1]).

---

[1]) Здесь мы также отклоняемся от традиций, согласно которым при определении кольца всегда требуют коммутативности сложения (выполнимости равенства $a + b = b + a$ при всех a и b), но не настаивают на коммутативности умножения, т. е. на обязательности равенства $ab = ba$. (Отметим также, что иногда в определение кольца не в к л ю ч а ю т и требование существования единичного элемента 1.)

Примеры.

а) Ясно, что поле — это частный случай кольца (поле — это кольцо с делением); поэтому все *примеры полей одновременно являются и примерами колец.*

б) *Совокупность всех целых чисел составляет кольцо* (относительно обыкновенных операций сложения и умножения чисел).

в) *Совокупность всех многочленов с коэффициентами из некоторого поля F составляет кольцо* относительно почленного сложения многочленов и почленного их умножения: если $a(x) = a_0 + a_1x + a_2x^2 + \ldots + a_{n-1}x^{n-1}$ и $b(x) = b_0 + b_1x + b_2x^2 + \ldots + b_{m-1}x^{m-1}$, то

$$a(x)\,b(x) = a_0b_0 + (a_0b_1 + a_1b_0)\,x + (a_0b_2 + a_1b_1 + a_2b_0)x^2 + \ldots + a_{n-1}b_{m-1}x^{n+m-2}.$$

Нулевым элементом этого кольца является многочлен 0, а единичным — многочлен 1 (оба они — многочлены нулевой степени).

Примеры б) и в) имеют между собой много общего; это общее проявляется, например, в существовании в обоих рассматриваемых кольцах деления с остатком числа a на b или многочлена $a(x)$ на $b(x)$ (где $|a| \geqslant |b|$, соответственно, ст. $a(x) \geqslant$ ст. $b(x)$; через ст. $f(x)$ мы обозначаем степень многочлена $f(x)$): $a = ub + r$, где $|r| < |b|$; $a(x) = u(x)\,b(x) + r(x)$, где ст.$r(x) <$ ст.$b(x)$. Здесь число u (многочлен $u(x)$) называется частным от деления a на b (или $u(x)$ на $b(x)$)), а число r (многочлен $r(x)$) — остатком (остаток от деления может оказаться равным 0).

Процедуру деления с остатком можно использовать для нахождения наибольшего общего делителя (НОД) двух чисел или двух многочленов. Так, например, ограничиваясь случаем (целых) чисел a и b и обозначая НОД этих чисел через (a, b), последовательно находим (числа мы считаем положительными):

$$a = ub + r, \quad \text{где } r < b \text{ и } (a, b) = (b, r);$$
$$b = u_1r + r_1, \quad \text{где } r_1 < r \text{ и } (b, r) = (r, r_1);$$
$$r = u_2r_1 + r_2, \quad \text{где } r_2 < r_1 \text{ и } (r, r_1) = (r_1, r_2);$$

. . . . . .

$$r_{k-2} = u_kr_{k-1} + r_k, \text{ где } r_k < r_{k-1} \text{ и } (r_{k-2}, r_{k-1}) = (r_{k-1}, r_k);$$
$$r_{k-1} = u_{k+1}r_k, \text{ и, значит, } (r_{k-1}, r_k) = r_k.$$

Таким образом, число r_k — это и есть

$$d = (a, b).$$

Важно заметить, что найденное описанным способом [1]) число $d = (a, b)$ можно выразить через исходные числа a и b в виде

$$d = Ma + Nb, \qquad (*)$$

где M и N — какие-то ц е л ы е числа. В самом деле, из выписанных выше формул последовательно находим

$$r = 1 \cdot a + (-u) \cdot b \quad (= m \cdot a + n \cdot b),$$
$$r_1 = 1 \cdot b + (-u_1) \cdot r = m_1 \cdot a + n_1 \cdot b,$$

$$r_2 = 1 \cdot r + (-u_2) \cdot r_1 = m_2 \cdot a + n_2 \cdot b, \ \ldots,$$
$$r_k = 1 \cdot r_{k-2} + (-u_k) \cdot r_{k-1} = M \cdot a + N \cdot b,$$

где все числа m и n (т. е. 1 и $-u$), m_1 и n_1 (они равны $-u_1$ и $1 + uu_1$), m_2 и n_2, ..., M и N — целые.

Из формулы (*), в частности, вытекает, что *в p-арифметике* (где p — п р о с т о е) *каждое число $a \neq 0$ имеет обратное*. В самом деле, если $0 < a < p$, то, очевидно, $(a, p) = 1$, и поэтому

$$1 = (a, p) = Ma + Np;$$

таким образом, произведение Ma $(= (-N) \cdot p + 1)$ при делении на p дает остаток 1. Но это и значит, что отвечающее M число m p-арифметики (остаток от деления M на p) в p-арифметике является обратным a: при перемножении чисел по правилам p-арифметики мы имеем $ma = 1$ и, значит, $m = a^{-1}$.

Совершенно та же процедура позволяет найти НОД $(a(x), b(x))$ двух м н о г о ч л е н о в $a(x)$ и $b(x)$

---

[1]) Описанная процедура нахождения наибольшего общего делителя a и b носит название а л г о р и т м а Е в к л и д а; кольца, в которых эта процедура применима (в частности — кольцо целых чисел или кольцо многочленов) иногда называют е в к л и д о в ы м и к о л ь ц а м и.

и доказать, что если $(a(x), b(x)) = d(x)$, то

$$d(x) = M(x) \cdot a(x) + N(x) \cdot b(x), \qquad (**)$$

где $M(x)$ и $N(x)$ — какие-то многочлены.

Аналогию между кольцом целых чисел и кольцом многочленов (с коэффициентами из какого-либо поля F) можно охарактеризовать еще и иначе. Подмножество J элементов произвольного кольца K называется идеалом этого кольца, если

(I) *множество J представляет собой подгруппу по отношению к определенной в K операции сложения*;

II) *для каждого a из J также и все произведения ak, где k — какой-угодно элемент K, принадлежат J.*

Типичным примером идеала кольца целых чисел является множество всех чисел, кратных произвольно выбранному целому числу i (т. е. чисел вида ai, где a пробегает все целые значения); аналогично этому примером идеала в множестве многочленов является множество многочленов, кратных произвольному наперед заданному многочлену $i(x)$ (т. е. множество многочленов вида $a(x)i(x)$, где $a(x)$ — произвольный многочлен). Идеалы описанного строения называются главными идеалами кольца целых чисел, соответственно кольца многочленов, порожденными числом i и многочленом $i(x)$.

Имеет место следующее утверждение, раскрывающее глубокие общие свойства колец целых чисел и многочленов:

В кольце целых чисел и в кольце многочленов каждый идеал J является главным, т. е. состоит из всевозможных кратных фиксированного целого числа i, соответственно — фиксированного многочлена $i(x)$.

Доказательство высказанного утверждения не представляет никакого труда. В самом деле, конечно, возможно, что идеал кольца *целых чисел* состоит из одного лишь числа 0 (для этого множества из одного элемента очевидно выполняются все определяющие идеал условия),—но в таком случае это есть главный идеал, порожденный числом 0. Если же это не так, то обозначим через i наименьшее по абсолютной величине отличное от нуля число, входящее в состав идеала J (для простоты можно условиться считать, например, что $i > 0$).

Докажем теперь, что любое другое принадлежащее J отличное от нуля число b обязательно будет кратно i. Так как $|b| \geqslant i$, то b можно разделить на i:

$$b = ai + r, \text{ где } 0 \leqslant r < i.$$

Но так как J — идеал, то наряду с b и i ему принадлежат и числа $ai, -ai$ и $r = b + (-ai)$. Поэтому $r = 0$ (ибо i — наименьшее по абсолютной величине из принадлежащих J и отличных от нуля чисел) и, значит, $b = ai$.

Относящееся к *кольцу многочленов* утверждение доказывается точно так же; здесь только за $i(x)$ надо принять отличный от 0 многочлен наинизшей степени, входящий в состав идеала J.

Обратимся теперь к дальнейшим примерам колец.

г) Мы уже видели, что *q-арифметика* с определенными в ней сложением и умножением *представляет собой кольцо* из q элементов (кольцо конечного порядка q); если при этом число q — простое, то наше кольцо является *полем*.

д) Выше отмечалось, что совокупность многочленов степени $< n$, где n — фиксированное число, представляет собой группу по сложению (конечную группу, если коэффициентами многочлена являются элементы конечного поля). Однако кольца такие многочлены не образуют, ибо степень произведения двух многочленов, вообще говоря, выше степени каждого из сомножителей. Для того, чтобы обратить совокупность многочленов степени $< n$ в кольцо, можно поступить так.

Выберем фиксированный (какой угодно!) многочлен $Q(x)$ степени n и условимся *заменять каждый многочлен остатком от его деления на* $Q(x)$; степень этого остатка будет уже $< n$. Так мы приходим к «*$Q(x)$-арифметике*» *многочленов*, в которой невозможны никакие многочлены степени $\geqslant n$; в частности, «произведение» двух многочленов, понимаемое в смысле «$Q(x)$-арифметики», всегда имеет степень $< n$. *$Q(x)$-арифметика* всегда (т. е. при любом выборе многочлена $Q(x)$) является кольцом; оно будет конечным, если поле коэффициентов многочленов конечно. Если порядок поля F коэффициентов равен p и ст. $Q(x) = n$, то порядок рассматриваемого кольца будет равен p^n.

Вот как выглядят «таблицы умножения» четырех многочленов степени < 2 с коэффициентами из 2-арифметики в $(x^2 + x)$-арифметике и в $(x^2 + x + 1)$-арифметике:

$\cdot$	0	1	x	$x+1$
0	0	0	0	0
1	0	1	x	$x+1$
x	0	x	x	0
$x+1$	0	$x+1$	0	$x+1$

и

$\cdot$	0	1	x	$x+1$
0	0	0	0	0
1	0	1	x	$x+1$
x	0	x	$x+1$	1
$x+1$	0	$x+1$	1	x

Поучительно сравнить эти две таблицы. Две последние строки первой из них *не содержат* числа 1 — это означает, что в $(x^2 + x)$-арифметике многочлены x и $x + 1$ не имеют обратных. Напротив, во второй таблице все строчки, кроме одной лишь первой, состоящей только из нулей, число 1 содержат; это значит, что в $(x^2 + x + 1)$-арифметике в *с е* многочлены, отличные от нулевого, имеют обратный: здесь

$$1^{-1} = 1,\ x^{-1} = x + 1 \text{ и } (x + 1)^{-1} = x.$$

Таким образом, в то время как $(x^2 + x)$-арифметика многочленов с коэффициентами из 2-арифметики представляет собой лишь кольцо, $(x^2 + x + 1)$-арифметика многочленов с коэффициентами из того же поля образует *п о л е*.

Нетрудно понять, с чем связано такое различие. Многочлен $Q(x) = x^2 + x$ является *составным*: он разлагается на множители степени $\geqslant 1$:

$$x^2 + x = x(x + 1).$$

Отсюда уже следует, что $(x^2 + x)$-арифметика полем являться не может — это вытекает, хотя бы из того, что здесь $x(x + 1) = 0$. Напротив, многочлен $P(x) = x^2 + x + 1$ — *простой* (или, как чаще говорят в алгебре, *неприводимый*): его нельзя разложить на множители степени $\geqslant 1$. А отсюда, в свою очередь, сразу вытекает, что в $P(x)$-арифметике каждый многочлен $a(x) \neq 0$

имеет обратный; доказательство этого факта, опирающееся на формулу (**), стр. 468, во всем аналогично доказательству того, что в p-арифметике, где p — простое, каждое число a имеет обратное ему.

Таким образом, мы приходим к еще одному примеру поля:

В. Если $P(x)$ — неприводимый многочлен с коэффициентами из некоторого поля F, то $P(x)$-*арифметика* с коэффициентами из F *образует поле*. Если F — это описанное выше конечное поле F_p порядка p (где p — произвольное простое число) и ст. $P(x) = n$, то порядок полученного поля равен p^n.

Можно показать, что при любом простом p для каждого $k > 1$ существует неприводимый многочлен степени k с коэффициентами из поля F_p; отсюда следует, что *при любом целом* $k \geqslant 1$ *и любом простом* p *существует конечное поле порядка* p^k (полем порядка $p^1 = p$ является сама p-арифметика). При этом хотя неприводимых многочленов $P(x)$ данной степени k с коэффициентами из поля F_p может существовать много, все отвечающие им $P(x)$-арифметики устроены одинаково: для каждого простого p и каждого $k \geqslant 1$ существует лишь одно (с точностью до переименования элементов) поле порядка p^k. Если же целое число m не имеет вида p^k (т. е. если m содержит хотя бы два различных простых множителя), то поля порядка m не существует вовсе [1]).

В заключение заметим еще, что поскольку $Q(x)$-арифметика получается из полного кольца многочленов (с коэффициентами из какого-то выбранного поля F) «склеиванием» всех многочленов, дающих один и тот же остаток при делении на $Q(x)$, то и идеалы $Q(x)$-арифметики получаются из идеалов кольца всех многочленов таким же отождествлением всех многочленов, дающих один остаток при делении на $Q(x)$. А отсюда, в свою очередь, следует, что идеалы $Q(x)$-арифметики устроены аналогично идеалам кольца всех многочленов: здесь также *каждый идеал*

---

[1]) Таким образом, поле конечного порядка m существует, если $m = p^k$, где p — некоторое простое число, и но существует для всех других чисел m, причем для каждого простого p и целого положительного k имеется лишь одно поле порядка p^k. Все эти поля были найдены замечательным французским математиком Э. Галуа; поэтому они называются полями Галуа.

является главным (т. е. состоит из всех многочленов, кратных в смысле $Q(x)$-арифметики некоторому фиксированному многочлену $i(x)$). При этом, однако, необходимо иметь в виду, что, как нетрудно усмотреть из формулы (**), стр. 468, понимаемое в смысле $Q(x)$-арифметики множество всех многочленов, кратных данному многочлену $i(x)$, совпадает с множеством всех многочленов, кратных многочлену $d(x)$, где $d(x) = (Q(x), i(x))$ есть НОД многочленов $i(x)$ и $Q(x)$. Отсюда следует, что при н е п р и в о д и м о м (простом) многочлене $Q(x)$ $Q(x)$-арифметика не содержит никаких идеалов, отличных от 0 и от всего кольца (от всей $Q(x)$-арифметики) — ведь здесь НОД $Q(x)$ и $i(x)$ совпадает с 1 или с $Q(x)$. Если же многочлен $Q(x)$ п р и в о д и м, т. е. раэлагается на множители, степени которых меньше ст. $Q(x)$, то множество всех многочленов, кратных каждому из этих множителей многочлена $Q(x)$, образует идеал $Q(x)$-арифметики — так, например, в случае $(x^2 + x)$-арифметики над 2-арифметикой множество всех идеалов состоит из «нулевого идеала» $\{0\}$; всей $(x^2 + x)$-арифметики; множества $\{x, 0\}$ многочленов, кратных x, и множества $\{x+1, 0\}$ многочленов, кратных $x + 1$ (см. верхнюю таблицу на стр. 470).

3. Перейдем теперь к следующему из используемых в теории кодирования алгебраических понятий.

Множество V элементов a, b, c и. т. д. (называемых в е к т о р а м и) *образует в е к т о р н о е п р о с т р а н с т в о над полем F* (элементы поля мы будем называть ч и с л а м и; нулевой и единичный элементы поля ниже обозначаются символами 0 и 1), *если*

I) *в множестве векторов определена операция сложения, относительно которой векторы образуют группу* (нулевой элемент этой группы обозначается символом 0);

II) *определена операция умножения вектора на число*; *при этом произведение $a\boldsymbol{a}$* (где a — число, а $\boldsymbol{a}$ — вектор) *есть вектор и*

1° *умножение вектора на число ассоциативно*: $a(b\boldsymbol{a}) = (ab)\boldsymbol{a}$; для всех чисел a, b и всех векторов $\boldsymbol{a}$;

2° *умножение вектора на число дистрибутивно относительно сложения чисел*:

$(a + b)\boldsymbol{a} = a\boldsymbol{a} + b\boldsymbol{a}$ для всех чисел a, b и всех векторов $\boldsymbol{a}$;

3° *умножение вектора на число дистрибутивно относительно сложения векторов*:
$a(\boldsymbol{a} + \boldsymbol{b}) = a\boldsymbol{a} + a\boldsymbol{b}$ для всех чисел a и всех векторов $\boldsymbol{a}, \boldsymbol{b}$;

4° $1\boldsymbol{a} = \boldsymbol{a}$ *для всех векторов* $\boldsymbol{a}$.

Из свойств (аксиом) умножения вектора на число легко выводится также, что

$0\boldsymbol{a} = \boldsymbol{0}$ для всех векторов $\boldsymbol{a}$; $a\boldsymbol{0} = \boldsymbol{0}$ для всех чисел a; $(-1)\,\boldsymbol{a} = -\boldsymbol{a}$ для всех векторов $\boldsymbol{a}$.

Примеры.

А. «*Блоки*» (*векторы*) $\boldsymbol{a} = (a_0, a_1, \ldots, a_{N-1})$, где N — фиксированное натуральное число и $a_0, a_1, \ldots, a_{N-1}$ — произвольные числа из поля F, *образуют векторное пространство* относительно следующим образом определенных операций сложения векторов и умножения вектора на число:
если $\boldsymbol{a} = (a_0, a_1, \ldots, a_{N-1})$ и $\boldsymbol{b} = (b_0, b_1, \ldots, b_{N-1})$, то

$$\boldsymbol{a} + \boldsymbol{b} = (a_0 + b_0, a_1 + b_1, \ldots, a_{N-1} + b_{N-1});$$

если $\boldsymbol{a} = (a_0, a_1, \ldots, a_{N-1})$, то $a\boldsymbol{a} = (aa_0, aa_1, \ldots, aa_{N-1})$. При этом поле F называется полем скаляров или основным полем, над которым строится векторное пространство V; числа $a_0, a_1, \ldots, a_{N-1}$ называются координатами вектора $\boldsymbol{a}$, а число N — размерностью нашего векторного пространства.

Если поле F — бесконечно, то число возможных векторов также является бесконечным; если же F — поле порядка m, то векторное пространство V размерности N (N-мерное векторное пространство) содержит всего m^N векторов.

Этот пример является основным; другие примеры всегда стараются свести к нему.

Б. *Векторы* (*направленные отрезки*) *плоскости или пространства образуют векторное пространство* относительно следующим образом определенных операций сложения векторов и умножения вектора на (вещественное!) число:

$\overline{OA} + \overline{OB} = \overline{OC}$, если OC — диагональ параллелограмма $OACB$, построенного на отрезках OA и OB;

$\overline{OD} = a \cdot \overline{OA}$, если $\overline{OD}$ и $\overline{OA}$ принадлежат одной прямой; $OD = |a| \cdot OA$; $\overline{OD}$ и $\overline{OA}$ направлены в одну сторону, если $a > 0$, и в противоположные стороны, если $a < 0$.

Пример Б сводится к основному примеру А, если обычным образом ввести координаты (x, y) вектора $\overline{OA}$ плоскости (рис. 38, *а*) и координаты (x, y, z) вектора $\overline{OA}$ пространства (рис. 38, *б*). При этом оказывается, что в случае векторов плоскости

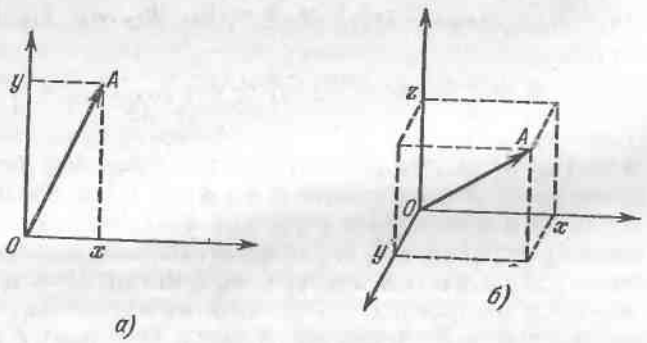
если $a = (x, y)$ и $b = (x_1, y_1)$, то

$$a + b = (x + x_1, y + y_1) \text{ и } aa = (ax, ay);$$

в случае векторов пространства если $a = (x, y, z)$ и $b = (x_1, y_1, z_1)$, то

$$a + b = (x + x_1, y + y_1, z + z_1) \text{ и } aa = (ax, ay, az).$$

Таким образом, *векторы плоскости образуют двумерное векторное пространство, а векторы пространства —*



а) *б)*

Рис. 38.

трехмерное векторное пространство над полем вещественных чисел.

В. Ясно, что произвольные $(m \times n)$-*матрицы с элементами из поля* F *образуют* $(m\,n)$-*мерное пространство над* F, если сложение матриц определять как выше, а умножение матрицы на число a — как умножение всех элементов матрицы на это число: ведь различие этого примера с основным примером А заключается лишь в том, что mn координат вектора здесь записывается не в одну строку, а в виде прямоугольной таблицы.

Г. *Многочлены степени* $< n$

$$a_0 + a_1 x + a_2 x^2 + \ldots + a_{n-1} x^{n-1}$$

с коэффициентами из поля F образуют n-мерное векторное пространство над F: ведь каждый многочлен можно характеризовать его коэффициентами $a_0, a_1, \ldots, a_{n-1}$ (которые, если угодно, можно выписывать, заключая их в круглые скобки), а (обычное) сложение многочленов и умножение многочлена на число сводится к сложению коэффициентов двух многочленов и к умножению коэффициентов многочлена на число.

Д. *Всевозможные многочлены*

$$a_0 + a_1 x + a_2 x^2 + \ldots + a_k x^k,$$

степень которых заранее никак не ограничивается, тоже *образуют векторное пространство* относительно обычных операций сложения многочленов и умножения многочлена на число. Этот пример, однако, не сводится к примеру А, поскольку число коэффициентов многочлена может быть сколь угодно велико; поэтому говорят, что пространство всех многочленов размерности не имеет (иногда вместо этого говорят, что оно имеет бесконечную размерность).

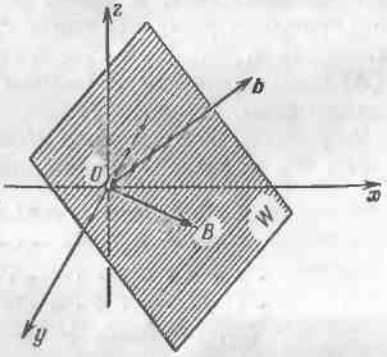


Рис. 39.

Пусть теперь W, — некоторая часть векторов векторного пространства v. Если множество W таково что

1° *если вектор a принадлежит W и вектор b принадлежит W, то также и вектор $a + b$ принадлежит W*;

2° *если вектор a принадлежит W, то W принадлежат также и все векторы αa, где α — всевозможные числа*,

то множество векторов W само представляет векторное пространство относительно определенных в V операций сложения векторов и умножения вектора на число. В этом случае говорят, что W представляет собой (линейное или векторное) подпространство векторного пространства V.

В частности, если V — множество векторов $\overline{OA}$ обыкновенного пространства, а W — проходящая через точку O

плоскость (рис. 39), то принадлежащие W векторы $\overline{OB}$ образуют подпространство всего векторного пространства. Если V — множество всех n-мерных векторов

$$a = (a_1, a_2, \ldots, a_n),$$

то множество W векторов a, координаты которых удовлетворяют фиксированному соотношению вида

$$b_1a_1 + b_2a_2 + \ldots + b_na_n = 0, \qquad (А)$$

где $b_1, b_2, \ldots, b_n$ — произвольные «числа», т. е. элементы того поля, которому принадлежат координаты векторов, образует подпространство. В самом деле, легко видеть, что если числа $a_1, a_2, \ldots, a_n$ и $a_1', a_2', \ldots, a_n'$ удовлетворяют соотношению (А), то и числа $a_1 + a_1', a_2 + a_2', \ldots, a_n + a_n'$ удовлетворяют тому же соотношению; аналогично этому, если числа $a_1, a_2, \ldots, a_n$ удовлетворяют соотношению (А), то (А) удовлетворяют и числам $aa_1, aa_2, \ldots, aa_n$, где a — произвольное число. Также просто устанавливается, что множество W векторов a, координаты которых удовлетворяют с и с т е м е соотношений

$$\left.\begin{aligned} b_{11}a_1 + b_{12}a_2 + \cdots + b_{1n}a_n &= 0, \\ b_{21}a_1 + b_{22}a_2 + \cdots + b_{2n}a_n &= 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ b_{m1}a_1 + b_{m2}a_2 + \cdots + b_{mn}a_n &= 0, \end{aligned}\right\} \qquad (Б)$$

образует векторное пространство. Более того, во всех учебниках линейной алгебры доказывается, что к а ж д о е подпространство W n-мерного векторного пространства V может быть задано системой соотношений (Б) (быть может, единственным соотношением (А)), связывающих координаты принадлежащих этому подпространству векторов. В частности, принадлежащие фиксированной плоскости W векторы трехмерного пространства V выделяются тем, что их координаты x, y, z удовлетворяют соотношению

$$b_1x + b_2y + b_3z = 0,$$

где (b_1, b_2, b_3) — координаты произвольного вектора b, п е р п е н д и к у л я р н о г о плоскости W (рис. 39).

Вот еще примеры векторных подпространств:

Множество всех многочленов степени $< n$ составляет линейное подпространство пространства всех вообще много-

членов; если $k < n$, то множество всех многочленов степени $< k$ составляет подпространство пространства многочленов степени $< n$; множество всех многочленов вида

$$a(x) = g(x)\, b(x),$$

где $g(x)$ — фиксированный многочлен, а $b(x)$ — какой угодно, составляет подпространство множества всех многочленов (а если $g(x)$ имеет степень k, а $b(x)$ — произвольный многочлен степени $< n - k$, то множество рассматриваемых многочленов представляет собой подпространство множества всех многочленов степени $< n$).

Заметим еще, что в случае векторного пространства над полем из чисел 0 и 1 (над 2-арифметикой) проверка того, что некоторое множество векторов образуют подпространство исходного пространства, сводится к проверке свойства 1^0 (ибо здесь нет отличных от 0 и 1 чисел, а вектор $0 \cdot a$ всегда можно представить в виде суммы $a + a$). Таким образом, здесь *все подпространства векторного пространства совпадают с подгруппами группы векторов по сложению*. Нетрудно показать, что точно так же обстоит дело и в случае векторного пространства, построенного над любой p-арифметикой, где p — простое число; однако в случае отличного от p-арифметики основного поля (например, когда в качестве основного поля фигурирует $P(x)$-арифметика, где $P(x)$ — неприводимый многочлен) существуют и подгруппы векторного пространства, не являющиеся его подпространствами.

От понятия векторного пространства легко перейти к (основному для геометрии!) понятию евклидова пространства. А именно, *N-мерное векторное пространство E называется евклидовым, если в нем определена длина* $|a|_e$ (или просто $|a|$) *вектора a с координатами* $(a_0, a_1, \ldots, a_{N-1})$:

$$|a|_e = \sqrt{a_0^2 + a_1^2 + \ldots + a_{N-1}^2} \qquad (^1)$$

(разумеется, основное поле здесь должно быть таково, чтобы в нем существовал корень квадратный из суммы квадратов любых двух элементов поля). Далее, если условиться называть векторы евклидова пространства «точками», сопоставив нулевой вектор 0 некоторой точке O, а вектор a — точке A с теми же координатами и условившись писать $a = \overrightarrow{OA}$, то *расстояние* $|AB|_e$ *или просто*

$|AB|$ *между точками* A и B определится так:

$$|AB| = |\overline{OB} - \overline{OA}| = \sqrt{(b_0 - a_0)^2 + (b_1 - a_1)^2 + \ldots + (b_{N-1} - a_{N-1})^2}, \qquad (^{**})$$

где $(a_0, a_1, \ldots, a_{N-1})$ и $(b_0, b_1, \ldots, b_{N-1})$ — координаты точек A и B (т. е. векторов $\overline{OA}$ и $\overline{OB}$). После этого содержание евклидовой геометрии можно охарактеризовать как *описание тех свойств фигур* (т. е. множеств точек) *евклидова пространства* E, *которые будут одинаковыми для любых двух равных фигур* (где равенство фигур определяется условием равенства расстояний между парами соответствующих друг другу точек этих фигур).

Евклидово пространство с вещественным и координатами точек и векторов является примером метрического векторного пространства. Множество M точек называется *метрическим пространством*, если для каждых двух точек A и B определено (вещественное) число ρ_{AB}, называемое *расстоянием* между A и B, причем

1° $\rho_{AB} > 0$ при $A \neq B$; $\rho_{AA} = 0$ (положительность расстояния);

2° $\rho_{AB} = \rho_{BA}$ (симметричность расстояния);

3° $\rho_{AB} + \rho_{BC} \geqslant \rho_{AC}$ при любых A, B и C (неравенство треугольника).

Если число $\rho_{AB} = |AB|_e$ определяется по формуле $(^{++})$, то условия 1° и 2°, очевидно, выполняются. Несколько сложнее установить выполнимость условия 3°, т. е. справедливость неравенства

$$\sqrt{(b_0 - a_0)^2 + (b_1 - a_1)^2 + \ldots + (b_{N-1} - a_{N-1})^2} + \sqrt{(c_0 - b_0)^2 + (c_1 - b_1)^2 + \ldots + (c_{N-1} + b_{N-1})^2} \geqslant \sqrt{(c_0 - a_0)^2 + (c_1 - a_1)^2 + \ldots + (c_{N-1} - a_{N-1})^2}.$$

—но и она может быть доказана без особого труда [1]).

Существуют и много других способов введения «метрики» в N-мерном векторном пространстве. Так, например, во многих отношениях более простой, чем евклидова метрика $(^{+})$—$(^{++})$, является так называемая «метрика Минковского» [2]):

$$|a|_M = |a_0| + |a_1| + \ldots + |a_{N-1}| \qquad (^{*})$$

и

$$|AB|_M = |b_0 - a_0| + |b_1 - a_1| + \ldots + |b_{N-1} - a_{N-1}|, \qquad (^{**})$$

---

[1]) См., например, А. Н. Колмогоров, С. В. Фомин, Элементы теории функций и функционального анализа, М., «Наука», 1972, стр. 45.

[2]) Немецкий математик Г. Минковский в своих исследованиях по теории чисел рассмотрел более общий метод введения метрики в N-мерном векторном пространстве, охватывающий обе формулы $(^{+})$ и $(^{**})$.

где $|a|$ — абсолютная величина (вещественного) числа a; из формулы (**) непосредственно следует, что и расстояние $\rho_{AB} = |AB|_м$ удовлетворяет условиям 1°—3°.

Метрику (*)—(**) можно определить для векторного пространства, построенного над любым основным полем F, для которого существует *абсолютная величина* элемента a поля — такое вещественное число $|a|$, что [1])

1) $|a| > 0$ при $a \neq 0$; $|0| = 0$;

2) $|ab| = |a| \cdot |b|$;

3) $|a + b| \leqslant |a| + |b|$.

В частности, если основное поле представляет собой 2-арифметику, а абсолютная величина элементов поля определяется привычными равенствами

$$|0| = 0, \quad |1| = 1$$

(где справа стоят *вещественные числа* 0 и 1), то введенная выше метрика называется «метрикой Хэмминга»:

$$|a|_х = |a_0| + |a_1| + \ldots + |a_{N-1}|; \qquad (*)$$

$$|AB|_х = |b_0 - a_0| + |b_1 - a_1| + \ldots + |b_{N-1} - a_{N-1}|. \qquad (**)$$

Ясно, что если точки $A\ (a_0, a_1, \ldots, a_{N-1})$ и $B\ (b_0, b_1, \ldots, b_{N-1})$ N-мерного пространства с координатами из 2-арифметики отвечают двум последовательностям сигналов, то расстояние $|AB|_х$ равно *числу несовпадающих сигналов в последовательностях* A *и* B, чем и объясняется широкое использование метрики Хэмминга в теории кодирования [2]). При этом из неравенства треугольника следует, что два «шара Хэмминга» радиуса n с центрами Q_1 и Q_2 (т. е. множества точек A таких, что $|Q_1A|_х \leqslant n$, соответственно, $|Q_2A|_х \leqslant n$;

---

[1]) При этом в равенстве $|0| = 0$ стоящие слева и справа символы 0 имеют несколько разный смысл: нуль слева является *элементом* рассматриваемого *поля*, в то время как справа стоит просто *вещественное число*. Аналогичное замечание можно сделать и по поводу некоторых других равенств ниже.

[2]) В том случае, когда основное поле F содержит более двух элементов, метрика Хэмминга определяется теми же формулами (*)—(**), что и выше, где, однако, теперь уже надо положить

$$|a| = \begin{cases} 0, & \text{если } a = 0, \\ 1, & \text{если } a \neq 0. \end{cases}$$

При этом расстояние Хэмминга $|AB|_х$ по-прежнему будет равно числу несовпадающих между собой сигналов в последовательностях A и B.

Заметим еще, что наряду с «расстоянием Хэмминга» в теории кодирования используются и некоторые другие метрики в пространствах последовательностей сигналов (например, так называемая «метрика Ли», совпадающая с «метрикой Хэмминга» в случае поля F из двух элементов, но в других случаях учитывающая уже не только сам факт несовпадения каких-то координат точек A и B между собой, но также и то, насколько сильно эти координаты отличаются друг от друга; см. [169], раздел 8.2).

ср. выше, стр. 435) не могут пересечься, если $Q_1Q_2 > 2n$ (это обстоятельство уже использовалось выше на стр. 435).

Заметим еще, что если последовательности A (a_0, a_1, . . ., a_{N-1}), где все a_i принимают значения 0 и 1, изображать точками обычного («вещественного») N-мерного пространства (эти точки будут являться вершинами «единичного куба» N-мерного евклидова пространства), то, очевидно,

$$|AB|_e = \sqrt{|AB|_x}.$$

Поэтому евклидово расстояние $|AB|_e$ между точками A и B, определяемое по формуле (++), может служить вполне удовлетворительной характеристикой различия между последовательностями A (a_0, a_1, . . ., a_{N-1}) и B(b_0, b_1, . . ., b_{N-1}) элементарных сигналов. Это обстоятельство позволяет использовать в теории связи результаты, относящиеся к (N-мерной) евклидовой геометрии (в первую очередь — результаты так называемой дискретной геометрии, специально занимающейся проблемами *«плотнейших укладок» непересекающихся равных шаров* в многомерных пространствах и задачами отыскания таких конфигураций из конечного числа точек, расположенных в данной области пространства, для которых *наименьшее из попарных расстояний между этими точками является наибольшим возможным*). В частности, задача отыскания всех двоичных кодов, где кодовые обозначения представляют собой последовательности из N элементарных сигналов, *исправляющих любое не превосходящее n число ошибок*, сводится к задаче отыскания всех возможных *заполнений «единичного куба» N-мерного евклидова пространства непересекающимися шарами радиуса $\sqrt{n}$ с центрами в вершинах куба*. В силу сказанного задача нахождения такого заполнения N-мерного куба шарами заданного радиуса, где число участвующих шаров — самое большое из возможных (или, по крайней мере, достаточно велико), представляет значительный интерес для теории кодирования; однако сегодня, мы, к сожалению, не знаем никаких перспективных геометрических подходов к решению этой задачи.

4. В линейной алгебре важную роль играет операция *умножения матриц*, частным случаем которой является *умножение ($m \times n$)-матрицы на ($n \times 1$)-матрицу* (*на вектор-столбец*):

$$B\,a = \begin{pmatrix} b_{11}\,b_{12}\,\ldots\,b_{1n} \\ b_{21}\,b_{22}\,\ldots\,b_{2n} \\ \ldots\ldots\ldots \\ \ldots\ldots\ldots \\ b_{m1}b_{m2}\ldots b_{mn} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} =$$

$$= \begin{pmatrix} b_{11}a_1 + b_{12}a_2 + \cdots + b_{1n}a_n \\ b_{21}a_1 + b_{22}a_2 + \cdots + b_{2n}a_n \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ b_{m1}a_1 + b_{m2}a_2 + \cdots + b_{mn}a_n \end{pmatrix}.$$

Разумеется, в последнем произведении можно также писать вектор a с координатами $a_1, a_2, \ldots, a_n$ в виде вектора-строки: $a = (a_1, a_2, \ldots, a_n)$, хотя это и не соответствует принятым в линейной алгебре соглашениям. В таком случае соотношениям (Б) (см. стр. 476) оказывается возможным придать вид

$$Ba = 0,$$

где 0 — нулевой вектор-столбец из n нулей.

Для некоторых разделов линейной алгебры оказывается также существенным понятие *элементарных преобразований* матриц, под которыми мы здесь будем понимать следующие преобразования:

1° перестановку местами любых двух строк матрицы;

2° перестановку местами любых двух ее столбцов;

3° замену любой строки матрицы ее суммой с любой другой строкой (где сумма строк понимается как сумма векторов-строк).

Матрицы, получающиеся одна из другой с помощью конечной последовательности элементарных преобразований, называются эквивалентными.

Указанные элементарные преобразования [1]) являются особенно естественными в случае проверочных матриц кодов с проверками на четность. В самом деле, в этом случае перестановка столбцов матрицы сводится лишь к перенумерации сигналов, а перестановка строк — к перенумерации используемых проверок. Замена же некоторой строки ее суммой с другой строкой означает, что вместо двух проверок на четность мы проверяем четность одного из ранее использовавшихся выражений и суммы этого выражения со вторым из них — ясно, что такие две проверки полностью равносильны первоначальным. Далее легко установить, что с помощью последовательности элементарных преобразований каждая проверочная матрица может быть приведена к виду (2), указанному на стр. 409 (или, что ничего не меняет, к виду, отличающемуся от формы (2) лишь наличием у матрицы еще нескольких строк, составленных из одних нулей — этим строкам, очевидно, не отвечают никакие новые проверки и потому их можно

[1]) В разных задачах линейной алгебры оказываются удобными разные наборы элементарных преобразований.

просто отбросить). В самом деле, не представляющие интереса нулевые строки матрицы, если только такие строки у нее имеются, мы с помощью операции 1° сделаем самыми верхними — и так же будем поступать далее, если в процессе преобразования матрицы у нее будут появляться новые «нулевые» строки. Рассмотрим теперь самую нижнюю строку; ясно, что с помощью операции 2°, имеющийся в ней элемент 1 можно переместить в крайний правый столбец. Прибавляя затем эту строку ко всем, у которых в последнем столбце стоит 1, и учитывая, что в 2-арифметике $1 + 1 = 0$, мы можем превратить в нули все элементы последнего столбца, кроме одной лишь единицы, стоящей в последней строке. Если после этого 2-я снизу строка окажется состоящей из одних нулей, мы ее переместим вверх; если же она содержит хоть одну единицу, то с помощью операции 2° мы переместим эту единицу на предпоследнее место, а затем с помощью операции 3° обратим в нуль все прочие элементы предпоследнего столбца. Далее перейдем к третьей от конца строки — и с помощью тех же операций придадим требуемый вид третьему от конца столбцу — и т. д. В результате мы получим матрицу вида (2), быть может, только дополненную сверху несколькими строками из одних нулей.

В применении к проверочным матрицам кодов с проверками на четность этот результат доказывает, что *любой код с проверками на четность может быть записан в виде систематического кода*, число проверок на четность в котором, однако, может оказаться меньшим, чем в исходном «несистематическом» коде (ср. стр. 410 и пример на стр. 431).

ПРИЛОЖЕНИЕ III

ТАБЛИЦА ВЕЛИЧИН — $p \log_2 p$

p	0	1	2	3	4	5	6	7	8	9
0,00	—	0,0100	0,0179	0,0251	0,0319	0,0382	0,0443	0,0501	0,0557	0,0612
0,01	0,0664	0,0716	0,0766	0,0815	0,0862	0,0909	0,0955	0,0999	0,1043	0,1086
0,02	0,1129	0,1170	0,1211	0,1252	0,1291	0,0330	0,1369	0,1407	0,1444	0,1481
0,03	0,1518	0,1554	0,1589	0,1624	0,1659	0,1693	0,1727	0,1760	0,1793	0,1825
0,04	0,1858	0,1889	0,1921	0,1952	0,1983	0,2013	0,2043	0,2073	0,2103	0,2132
0,05	0,2161	0,2190	0,2218	0,2246	0,2274	0,2301	0,2329	0,2356	0,2383	0,2409
0,06	0,2435	0,2461	0,2487	0,2513	0,2538	0,2563	0,2588	0,2613	0,2637	0,2661
0,07	0,2686	0,2709	0,2733	0,2756	0,2780	0,2803	0,2826	0,2848	0,2871	0,2893
0,08	0,2915	0,2937	0,2959	0,2980	0,3002	0,3023	0,3044	0,3065	0,3086	0,3106
0,09	0,3127	0,3147	0,3167	0,3187	0,3207	0,3226	0,3246	0,3265	0,3284	0,3303
0,10	0,3322	0,3341	0,3359	0,3378	0,3398	0,3414	0,3432	0,3450	0,3468	0,3485
0,11	0,3503	0,3520	0,3537	0,3555	0,3571	0,3588	0,3605	0,3622	0,3638	0,3654
0,12	0,3671	0,3687	0,3703	0,3719	0,3734	0,3750	0,3766	0,3781	0,3796	0,3811
0,13	0,3826	0,3841	0,3856	0,3871	0,3886	0,3900	0,3915	0,3929	0,3943	0,3957
0,14	0,3971	0,3985	0,3999	0,4012	0,4026	0,4040	0,4053	0,4066	0,4079	0,4092
0,15	0,4105	0,4118	0,4131	0,4144	0,4156	0,4169	0,4181	0,4194	0,4206	0,4218
0,16	0,4230	0,4242	0,4254	0,4266	0,4277	0,4289	0,4301	0,4312	0,4323	0,4335
0,17	0,4346	0,4357	0,4368	0,4379	0,4390	0,4400	0,4411	0,4422	0,4432	0,4443
0,18	0,4453	0,4463	0,4474	0,4484	0,4494	0,4504	0,4514	0,4523	0,4533	0,4543
0,19	0,4552	0,4562	0,4571	0,4581	0,4590	0,4599	0,4608	0,4617	0,4626	0,4635
0,20	0,4644	0,4653	0,4661	0,4670	0,4678	0,4687	0,4695	0,4704	0,4712	0,4720

Продолжение

p	0	1	2	3	4	5	6	7	8	9
0,21	0,4728	0,4736	0,4744	0,4752	0,4760	0,4768	0,4776	0,4783	0,4791	0,4798
0,22	0,4806	0,4813	0,4820	0,4828	0,4835	0,4842	0,4849	0,4856	0,4863	0,4870
0,23	0,4877	0,4883	0,4890	0,4897	0,4903	0,4910	0,4916	0,4923	0,4949	0,4935
0,24	0,4941	0,4947	0,4954	0,4960	0,4966	0,4971	0,4977	0,4983	0,4989	0,4994
0,25	0,5000	0,5006	0,5011	0,5016	0,5022	0,5027	0,5032	0,5038	0,5043	0,5048
0,26	0,5053	0,5058	0,5063	0,5068	0,5072	0,5077	0,5082	0,5087	0,5091	0,5096
0,27	0,5100	0,5105	0,5109	0,5113	0,5118	0,5122	0,5126	0,5130	0,5134	0,5138
0,28	0,5142	0,5146	0,5150	0,5154	0,5158	0,5161	0,5165	0,5169	0,5172	0,5176
0,29	0,5179	0,5182	0,5186	0,5189	0,5192	0,5196	0,5199	0,5202	0,5205	0,5208
0,30	0,5211	0,5214	0,5217	0,5220	0,5222	0,5225	0,5228	0,5230	0,5233	0,5235
0,31	0,5238	0,5240	0,5243	0,5245	0,5247	0,5250	0,5252	0,5254	0,5256	0,5258
0,32	0,5260	0,5262	0,5264	0,5266	0,5268	0,5270	0,5272	0,5273	0,5275	0,5277
0,33	0,5278	0,5280	0,5281	0,5283	0,5284	0,5286	0,5287	0,5288	0,5289	0,5290
0,34	0,5292	0,5293	0,5294	0,5295	0,5296	0,5297	0,5298	0,5299	0,5299	0,5300
0,35	0,5301	0,5302	0,5302	0,5303	0,5304	0,5304	0,5305	0,5305	0,5305	0,5306
0,36	0,5306	0,5306	0,5307	0,5307	0,5307	0,5307	0,5307	0,5307	0,5307	0,5307
0,37	0,5307	0,5307	0,5307	0,5307	0,5307	0,5306	0,5306	0,5306	0,5305	0,5305
0,38	0,5304	0,5304	0,5303	0,5303	0,5302	0,5302	0,5301	0,5300	0,5300	0,5299
0,39	0,5298	0,5297	0,5296	0,5295	0,5294	0,5293	0,5292	0,5291	0,5290	0,5289
0,40	0,5288	0,5286	0,5285	0,5284	0,5283	0,5281	0,5280	0,5278	0,5277	0,5275
0,41	0,5274	0,5272	0,5271	0,5269	0,5267	0,5266	0,5264	0,5262	0,5260	0,5258
0,42	0,5256	0,5255	0,5253	0,5251	0,5249	0,5246	0,5244	0,5242	0,5240	0,5238
0,43	0,5236	0,5233	0,5231	0,5229	0,5226	0,5224	0,5222	0,5219	0,5217	0,5214
0,44	0,5211	0,5209	0,5206	0,5204	0,5201	0,5198	0,5195	0,5193	0,5190	0,5187
0,45	0,5184	0,5181	0,5178	0,5175	0,5172	0,5169	0,5166	0,5163	0,5160	0,5157
0,46	0,5153	0,5150	0,5147	0,5144	0,5140	0,5137	0,5133	0,5130	0,5127	0,5123

Продолжение

p	0	1	2	3	4	5	6	7	8	9
0,47	0,5120	0,5116	0,5112	0,5109	0,5105	0,5102	0,5098	0,5094	0,5090	0,5087
0,48	0,5083	0,5079	0,5075	0,5071	0,5067	0,5063	0,5059	0,5055	0,5051	0,5047
0,49	0,5043	0,5039	0,5034	0,5030	0,5026	0,5022	0,5017	0,5013	0,5009	0,5004
0,50	0,5000	0,4996	0,4991	0,4987	0,4982	0,4978	0,4973	0,4968	0,4964	0,4959
0,51	0,4954	0,4950	0,4945	0,4940	0,4935	0,4930	0,4926	0,4921	0,4916	0,4911
0,52	0,4906	0,4901	0,4896	0,4891	0,4886	0,4880	0,4875	0,4870	0,4865	0,4860
0,53	0,4854	0,4849	0,4844	0,4839	0,4833	0,4828	0,4822	0,4817	0,4811	0,4806
0,54	0,4800	0,4795	0,4789	0,4784	0,4778	0,4772	0,4767	0,4761	0,4755	0,4750
0,55	0,4744	0,4738	0,4732	0,4726	0,4720	0,4714	0,4708	0,4702	0,4697	0,4691
0,56	0,4684	0,4678	0,4672	0,4666	0,4660	0,4654	0,4648	0,4641	0,4635	0,4629
0,57	0,4623	0,4616	0,4610	0,4603	0,4597	0,4591	0,4584	0,4578	0,4571	0,4565
0,58	0,4558	0,4551	0,4545	0,4538	0,4532	0,4525	0,4518	0,4512	0,4505	0,4498
0,59	0,4491	0,4484	0,4477	0,4471	0,4464	0,4457	0,4450	0,4443	0,4436	0,4429
0,60	0,4422	0,4415	0,4408	0,4401	0,4393	0,4386	0,4379	0,4372	0,4365	0,4357
0,61	0,4350	0,4343	0,4335	0,4328	0,4321	0,4313	0,4306	0,4298	0,4291	0,4383
0,62	0,4276	0,4268	0,4261	0,4253	0,4246	0,4238	0,4230	0,4223	0,4215	0,4207
0,63	0,4199	0,4192	0,4184	0,4176	0,4168	0,4160	0,4153	0,4145	0,4137	0,4129
0,64	0,4121	0,4113	0,4105	0,4097	0,4089	0,4080	0,4072	0,4064	0,4056	0,4048
0,65	0,4040	0,4032	0,4023	0,4015	0,4007	0,3998	0,3990	0,3982	0,3973	0,3965
0,66	0,3957	0,3948	0,3940	0,3931	0,3922	0,3914	0,3905	0,3897	0,3888	0,3880
0,67	0,3871	0,3862	0,3854	0,3845	0,3836	0,3828	0,3819	0,3810	0,3801	0,3792
0,68	0,3784	0,3775	0,3766	0,3757	0,3748	0,3739	0,3730	0,3721	0,3712	0,3703
0,69	0,3694	0,3685	0,3676	0,3666	0,3657	0,3648	0,3639	0,3430	0,3621	0,3611
0,70	0,3602	0,3593	0,3583	0,3574	0,3565	0,3555	0,3546	0,3536	0,3527	0,3518
0,71	0,3508	0,3499	0,3489	0,3480	0,3470	0,3461	0,3451	0,3441	0,3432	0,3422
0,72	0,3412	0,3403	0,3393	0,3383	0,3373	0,3364	0,3354	0,3344	0,3334	0,3324
0,73	0,3314	0,3304	0,3295	0,3285	0,3275	0,3265	0,3255	0,3245	0,3235	0,3225

Продолжение

p	0	1	2	3	4	5	6	7	8	9
0,74	0,3215	0,3204	0,3194	0,3184	0,3174	0,3164	0,3154	0,3144	0,3133	0,3123
0,75	0,3113	0,3103	0,3092	0,3082	0,3071	0,3061	0,3051	0,3040	0,3030	0,3019
0,76	0,3009	0,2999	0,2988	0,2978	0,2967	0,2956	0,2946	0,2935	0,2925	0,2914
0,77	0,2903	0,2893	0,2882	0,2871	0,2861	0,2850	0,2839	0,2828	0,2818	0,2807
0,78	0,2796	0,2785	0,2774	0,2763	0,2753	0,2741	0,2731	0,2720	0,2709	0,2698
0,79	0,2687	0,2676	0,2664	0,2653	0,2642	0,2631	0,2620	0,2609	0,2598	0,2587
0,80	0,2575	0,2564	0,2553	0,2542	0,2531	0,2519	0,2508	0,2497	0,2485	0,2474
0,81	0,2462	0,2451	0,2440	0,2428	0,2417	0,2405	0,2394	0,2382	0,2371	0,2359
0,82	0,2348	0,2336	0,2324	0,2313	0,2301	0,2290	0,2278	0,2266	0,2255	0,2243
0,83	0,2231	0,2220	0,2208	0,2196	0,2184	0,2172	0,2160	0,2149	0,2137	0,2125
0,84	0,2113	0,2101	0,2089	0,2077	0,2065	0,2053	0,2041	0,2029	0,2017	0,2005
0,85	0,1993	0,1981	0,1969	0,1957	0,1944	0,1932	0,1920	0,1908	0,1896	0,1884
0,86	0,1871	0,1859	0,1847	0,1834	0,1822	0,1810	0,1797	0,1785	0,1773	0,1760
0,87	0,1748	0,1735	0,1723	0,1711	0,1698	0,1686	0,1673	0,1661	0,1648	0,1635
0,88	0,1623	0,1610	0,1598	0,1585	0,1572	0,1560	0,1547	0,1534	0,1522	0,1509
0,89	0,1496	0,1484	0,1471	0,1458	0,1445	0,1432	0,1419	0,1407	0,1394	0,1381
0,90	0,1368	0,1355	0,1342	0,1329	0,1316	0,1303	0,1290	0,1277	0,1264	0,1251
0,91	0,1238	0,1225	0,1212	0,1199	0,1186	0,1173	0,1159	0,1146	0,1133	0,1120
0,92	0,1107	0,1094	0,1080	0,1067	0,1054	0,1040	0,1027	0,1014	0,1000	0,0987
0,93	0,0974	0,0960	0,0947	0,0933	0,0920	0,0907	0,0893	0,0880	0,0866	0,0853
0,94	0,0839	0,0826	0,0812	0,0798	0,0785	0,0771	0,0758	0,0744	0,0730	0,1717
0,95	0,0703	0,0689	0,0676	0,0662	0,0648	0,0634	0,0621	0,0607	0,0593	0,0579
0,96	0,0565	0,0552	0,0538	0,0524	0,0510	0,0496	0,0482	0,0468	0,0454	0,0440
0,97	0,0426	0,0412	0,0398	0,0384	0,0370	0,0356	0,0342	0,0328	0,0314	0,0300
0,98	0,0286	0,0271	0,0257	0,0243	0,0230	0,0214	0,0201	0,0186	0,0172	0,0158
0,99	0,0140	0,0129	0,0115	0,0101	0,0086	0,0072	0,0058	0,0043	0,0029	0,0014

ЛИТЕРАТУРА

Общие сочинения по теории информации и по кибернетике

1. К. Шеннон, Математическая теория связи, в книге: «Работы по теории информации и кибернетике», М., ИЛ, 1963, стр. 243—332. [В этой книге, рассчитанной, в первую очередь, на специалистов, собраны все основные работы К. Шеннона по теории информации и теории кодирования.]
2. Л. Бриллюэн, Наука и теория информации, М., Физматгиз, 1959.
3. Дж. Пирс, Символы, сигналы, шумы, М., «Мир», 1967.

За. К. Черри, Человек и информация, М., «Связь», 1972.

4. Ф. М. Вудворд, Теория вероятностей и теория информации с применениями к радиолокации, М., «Советское радио», 1955.
5. А. Файнстейн, Основы теории информации, М., ИЛ, 1960.
6. Р. Фано, Передача информации. Статистическая теория связи, М., «Мир», 1965.
7. Дж. Вольфовиц, Теоремы кодирования теории информации, М., «Мир», 1967.
8. А. Н. Колмогоров, Теория передачи информации, в книге: «Сессия Академии наук СССР по научным проблемам автоматизации производства 15—20 октября 1956 г.; пленарные заседания», М., Изд-во АН СССР, 1957.
9. Н. Винер, Кибернетика, М., «Советское радио», 1968.
10. И. А. Полетаев, Сигнал, М., «Советское радио», 1958.
11. У. Р. Эшби, Введение в кибернетику, М., ИЛ, 1958.
12. А. Моль, Теория информации и эстетическое восприятие, М., «Мир», 1966.
13. Дж. Возенкрафт, И. Джекобс, Теоретические основы техники связи, М., «Мир», 1966.
14. Теория информации в биологии (сборник переводов), М., ИЛ, 1960.
15. А. Н. Колмогоров, Три подхода к определению понятия „количество информации", Пробл. передачи информ. 1, № 1, 1965, стр. 3—11; К логическим основам теории информации и теории вероятностей, Пробл. передачи информ. 5, № 3, 1969, стр. 3—7.
16. А. К. Звонкин, Л. А. Левин, Сложность конечных объектов и обоснование понятия информации и случайности с помощью теории алгоритмов, Успехи матем. наук 25, вып. 6, 1970, стр. 85—127.

17. И. М. Гельфанд, А. Н. Колмогоров, А. М. Яглом, К общему определению количества информации, Докл. Акад. Наук СССР **111**, № 4, 1956, стр. 745—748; Количество информации и энтропия для непрерывных распределений, Труды 3-го Всесоюзного математического съезда, т. 3, М., Изд-во АН СССР, 1958, стр. 300—320.
18. Ю. А. Шрейдер, Об одной модели семантической теории информации, Проблемы кибернетики, вып. 13, М., 1965, стр. 233—240; О семантических аспектах теории информации, в сборнике: «Информация и кибернетика», М., «Советское радио», 1967, стр. 15—47.
19. И. Бар-Хиллел, Р. Карнап, Семантическая информация (Y. Bar-Hillel, R. Carnap, Semantic information), Brit. Journ. Phil. of Sci. 4, № 14, 1953, стр. 147—157 и в сборнике: «Теория связи» (Communication Theory; сост.— W. Jackson), New York, Academic Press, 1953, стр. 503—512.
19а. Сборник «Информация и умозаключение» (Information and Inference; ред.— J. Hintikka, P. Suppes), Dordrecht, Reidel, 1970.
20. Н. Эбрамсон, Теория информации и кодирование (N. Abramson, Information theory and coding), New York, McGraw-Hill, 1963.
21. Р. Эш, Теория информации (R. B. Ash, Information theory), New York, Interscience, 1965.
22. Р. Г. Галлагер, Теория информации и надежная связь (R. G. Gallager, Information theory and reliable communication), New York, Wiley, 1968. (Русский перевод этой книги готовится к печати издательством «Советское радио».)
23. П. Фей, Теория информации (P. Fey, Informationstheorie), Berlin (DDR), Akademie-Verlag, 1968.
24. Ж. Кульман, М. Дени-Папен, Задачи по теории информации с решениями (G. Cullman, M. Denis-Papin, Exercices de calcul informationnel avec leurs solutions), Paris, Michel, 1966.

Литература к гл. I

25. Б. В. Гнеденко, А. Я. Хинчин, Элементарное введение в теорию вероятностей, М., «Наука», 1970.
26. Ф. Мостеллер, Р. Рурке, Дж. Томас, Вероятность, М., «Мир», 1969.
27. С. Дайменд, Мир вероятностей, М., «Статистика», 1970.
28. Ю. Нейман, Вводный курс теории вероятностей и математической статистики, М., «Наука», 1968.
29. Е. С. Венцель, Теория вероятностей. М., «Наука», 1964.
30. Ю. А. Розанов, Теория вероятностей и ее приложения, в сборнике: «О некоторых вопросах современной математики и кибернетики», М., «Просвещение», 1965, стр. 78—141.
31. Дж. Т. Кальбертсон, Математика и логика цифровых устройств, М., «Просвещение», 1965, гл. III.
32. Дж. Кемени, Дж. Снелл, Дж. Томпсон, Введение в конечную математику, М., «Мир», 1964, гл. IV.

33. А. Н. Колмогоров, Теория вероятностей, в сборнике: «Математика, ее содержание, методы и значение», т. II, М., Изд-во АН СССР, 1956, стр. 252—284.
34. М. Кац, Теория вероятностей, в сборнике: «Математика в современном мире», М., «Мир», 1967, стр. 78—93.
35. Ф. Мостеллер, Пятьдесят занимательных вероятностных задач, М., «Наука», 1971.
36. Л. Д. Мешалкин, Сборник задач по теории вероятностей, М., Изд-во МГУ, 1963.
37. А. М. Яглом, И. М. Яглом, Неэлементарные задачи в элементарном изложении, М., Гостехиздат, 1954, п. 6 раздела I.

Литература к гл. II

38. Р. Хайман, Информация, содержащаяся в раздражении, как величина, определяющая время реакции (R. Hyman, Stimulus information as a determinant of reaction times), Journ. of Experemental Psychology 45, № 3, 1953, стр. 188—196.
39. У. Е. Хик, О скорости получения информации (W. E. Hick, On the rate of gain of information), Quart. Journ. Experemental Psychology 4, № 1, 1952, стр. 11—26.
40. В. И. Николаев, Определение времени, затрачиваемого оператором на решение задач по управлению судовой энергетической установкой, Изв. Акад. наук СССР (энергетика и транспорт), № 4, 1965, стр. 130—145.
41. Б. Ф. Ломов, Человек и техника (очерки инженерной психологии), М., «Советское радио», 1966.
42. Дж. А. Леонард, Опыты по определению времени реакции выбора и теория информации (J. A. Leonard, Choise reaction time experiments and information theory), в сборнике: «Теория информации» (Information Theory; ред.— C. Cherry), London, Butterworths, 1961, стр. 137—146.
43. Р. Д. Льюс, Теория селективной информации и некоторые ее применения к изучению поведения (R. D. Luce, The theory of selective information and some of its behavioral applications), в сборнике: «Developments in Mathematical Psychology» ред.) —R. D. Luce), Glencoe (USA), The Free Press, 1960, стр.5—119.
44. А. Н. Леонтьев, Е. П. Кричик, О применении теории информации в конкретно-психологических исследованиях, Вопросы психологии, № 5, 1961, стр. 25—46.
45. Ф. Аттнив, Применение теории информации к психологии: обзор основных понятий, методов и результатов (F. Attneave, Applications of information theory to psychology: a summary of basic concepts, methods and results), New York, Holt — Dryden, 1959.
46. Сборник «Теория информации в психологии» (Information theory in psychology; ред.— H. Quastler), Glencoe (USA), The Free Press, 1955.
47. Р. Л. Добрушин, Передача информации по каналу с обратной связью, Теор. вероятн. и ее примен. 3, № 4, 1958, стр. 395—412.
48. Д. К. Фаддеев, К понятию энтропии конечной вероятностной схемы, Успехи матем. наук 11, № 1, 1956, стр. 227—231.

49. З. Дароци, Обобщенные информационные функции (Z. Daróczy, Generalized information functions), Information and Control 16, № 1, 1970, стр. 36—51.

Литература к гл. III

50. Б. А. Кордемский, Математическая смекалка, М., «Наука», 1965.
51. Д. О. Шклярский, Н. Н. Ченцов, И. М. Яглом, Избранные задачи и теоремы элементарной математики (арифметика и алгебра), М., «Наука», 1965.
52. В. Дэвиде, Одна задача о взвешиваниях (V. Devidé. Ein Problem über Wagen), Elemente der Math. 10, № 1, 1959, стр. 11—15.
53. П. Дж. Келлог, Д. Дж. Келлог, Информационная энтропия и задача о фальшивой монете (P. J. Kellogg, D. J. Kellogg, Entropy of information and the odd ball problem), Journ. of Appl. Phys. 25, № 11, 1954. стр. 1438—1439.
54. С. С. Кислицин, Современное состояние теории поиска, Успехи матем. наук 17, № 1, 1962, стр. 243—244.
55. Р. Беллман, Б. Глас, О разных вариантах задачи о фальшивой монете (R. Bellman, B. Gluss, On various versions of the defective coin problem), Information and Control 4, № 2—3, 1961, стр. 118—131; исправление — там же, № 4, стр. 391.
56. Г. Штейнгауз, Сто задач, М., Физматгиз, 1959.
57. С. С. Кислицин, Уточнение оценки наименьшего среднего числа сравнений, необходимых для полного упорядочивания конечной совокупности, Вестник ЛГУ, № 19, вып. 4, 1963, стр. 143—145.
58. Л. Р. Форд, С. М. Джонсон, Проблема соревнований (L. R. Ford, S. M. Johnson, A tournament problem), American Math. Montly 66, № 5, 1959, стр. 387—389.
59. К.-Ф. Пикар, Теория вопросников (C.-F. Picard, Théorie des questionnaires), Paris, Gauthier — Villars, 1965; Графы и вопросники, т. II. Вопросники (Graphes et questionnaires, tome II, Questionnaires), Paris, Gauthier—Villars, 1972.
60. П. П. Пархоменко. Теория вопросников (обзор), Автоматика и телемеханика, № 4, 1970, стр. 140—159.

Литература к гл. IV

§ 1

61. А. А. Сардинас, Дж. У. Паттерсон, Необходимое и достаточное условие однозначного разложения закодированных сообщений, Кибернетический сборник, вып. 3. М., ИЛ, 1961, стр. 93—102.
62. Э. Н. Гилберт, Э. Ф. Мур, Двоичные кодовые системы переменной длины, там же, стр. 103—141.

§ 2

63. Д. А. Хаффмен (или Хафман), Метод построения кодов с минимальной избыточностью, Кибернетический сборник, вып. 3, М., ИЛ, 1961, стр. 79—87.
64. Б. Макмиллан, Основные теоремы теории информации (B. McMillan, The basic theorems of information theory), Annals Math. Statist. 24, № 2, 1953, стр. 196—219.
65. Б. Макмиллан, Два неравенства, обусловленных однозначностью расшифровывания, Кибернетический сборник, вып. 3, М., ИЛ, 1961, стр. 88—92.
66. Дж. Каруш, Простое доказательство неравенства Макмиллана (J. Karush, A simple proof of an inequality of McMillan), IRE Trans. on Inform. Theory IT-7, № 2, 1961, стр. 118.

§ 3

67. Р. Г. Пиотровский, Информационные измерения языка, Л., «Наука», 1968.
68. И. М. Яглом, Р. Л. Добрушин, А. М. Яглом, Теория информации и лингвистика, Вопросы языкознания, 1960, № 1, стр. 100—110.
69. А. А. Харкевич, Очерки общей теории связи, М., Гостехиздат, 1955.
70. Д. С. Лебедев, В. А. Гармаш, О возможности увеличения скорости передачи телеграфных сообщений, Электросвязь, 1958, № 1, стр. 68—69.
71. Г. П. Башарин. О статистической оценке энтропии последовательности независимых случайных величин, Теор. вероятн. и ее примен. 4, № 3, 1959, стр. 361—364.
72. Р. Л. Добрушин, Математические методы в лингвистике, Математическое просвещение (новая серия), вып. 6, М., Физматгиз, 1961, стр. 37—60.
73. В. Белевич, Теория информации и лингвистическая статистика (V. Belevitch, Théorie de l'information et statistique linguistique), Bulletin Acad. Royale Belgique (classe de sciences), 1956, стр. 419—436.
74. Г. А. Барнард, Статистическое определение энтропии слов для четырех западных языков (G. A. Barnard, Statistical calculation of world entropies for four western languages), IRE Trans. on Inform. Theory IT-1, № 1, 1955, стр. 49—53.
75. К. Шеннон, Предсказание и энтропия английского печатного текста, в книге: «Работы по теории информации и кибернетике» (см. [1]), стр. 669—686.
76. В. Ю. Урбах, К учету корреляций между буквами алфавита при вычислении количества информации в сообщении, Проблемы кибернетики, вып. 10, 1963, стр. 111—117.
77. Н. Г. Бертон, Дж. Ликлайдер, Длительные связи в статистической структуре печатного английского текста (N. G. Burton, J. C. R. Licklider, Longrange constraints in the statistical structure of printed English), Amer. Journ. of Psychology 68, № 4, 1955, стр. 650—653.

78. Г. Сиромони, Теоретико-информационная проверка знакомства с иностранным языком (G. Siromoney, An information-theoretical test for familiarity with a foreign language), Journ. Psychol. Researches 8, 1964, стр. 1—6.
79. Д. Джемисон, К. Джемисон, Заметка об энтропии частично-знакомых языков (D. Jamison, K. Jamison, A note on the entropy of partially-known languages), Information and Control 12, № 2, 1968, стр. 164—167.
80. П. Б. Невельский, М. Д. Розенбаум, Угадывание профессионального текста специалистами и неспециалистами, в сборнике: «Статистика речи и автоматический анализ текста», Л., «Наука», 1971, стр. 134—148.
81. А. П. Савчук, Об оценках энтропии языка по Шеннону, Теор. вероятн. и ее примен. 9, № 1, 1964, стр. 154—157.
82. К. Кюпфмюллер, Энтропия немецкого языка (K. Küpfmüller, Die Entropie der deutschen Sprache), Fernmeldetechnische Zeitschrift (FTZ), № 6, 1954, стр. 265—272.
83. Н. В. Петрова, Энтропия французского печатного текста, Изв. Акад. наук СССР (серия литературы и языка) 24, № 1, 1965, стр. 63—67; Н. Петрова, Р. Пиотровский, Р. Жиро, Энтропия французской письменной речи (N. Petrova, R. Piotrovski, R. Giraud, L'entropie du français écrit), Bull. Soc. de linguistique de Paris 58, № 1, 1964, стр. 130—152.
84. Р. Манфрионо, Энтропия итальянского языка и ее вычисление (R. Manfrionо, L'entropia della lingua italiana ed il sua calcolo), Alta frequenza 29, № 1, 1960, стр. 4—29; Х. Хансон, Энтропия шведского языка (H. Hansson, The entropy of the Swedish language), Trans. of the Second Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, Prague, 1960, стр. 215—217; Л. Долежель, Предсказания энтропии и избыточности чешских текстов (L. Doležel, Predbezny odhad entropie e redundance psané cestiny), Slovo a Sbovesnost 24, № 3, 1963, стр. 165—175; Ф. Зитек, Несколько замечаний по поводу энтропии чешского языка (F. Zitek, Quelques remarques au sujet de l'entropie du tchèque), Trans. of the Third Prague Conference on Inforamation Theory, Statistical Decision Functions, Random Processes, Prague, 1964, стр. 841—846; Е. Николау, К. Сала, А. Роченик, Наблюдения над энтропией румынского языка (E. Nicolau, C. Sala, A. Roceric, Observaţii asupra entropiei limbii române), Studişi cercetâi lingvist 10, № 1, 1959, стр. 35—54.
85. Р. А. Казарян, Оценка энтропии армянского текста, Изв. Акад. наук Арм. ССР (физико-математические науки) 14, № 4i, 1961, стр. 161—173; Д. Н. Ленский, К оценке энтропии адыгейских печатных текстов, Учен. записки Кабардино-балкарского университета (серия физико-математическая), вып. 16, Нальчик, 1962, стр. 165—166; Т. И. Ибрагимов, Оценка взаимосвязи бука в татарском литературном языке, Учен. записки Казанского гос. ун-та, 124, кн. 2 (Вероятностные методы и кибернетика, вып. III), Казань, 1964, стр. 141—145.

86. Е. Б. Ньюман, Н. Во, Избыточность текстов на трех языках (E. B. Newman, N. C. Waugh, The redundancy of texts in three languages), Information and Control 3, № 2, 1960, стр. 141—153.
87. Е. Б. Ньюман, Л. Дж. Герстман, Новый метод анализа письменного английского текста (E. B. Newman, L. J. Gerstman, A new method for analyzing printed English), Journ. of Experemental Psychology 44, № 2, 1952, стр. 114—125.
88. Г. Блюме, Трехмерные кроссворды на древне-еврейском языке (H. Bluhme, Three-dimensional crossword puzzles in Hebrew), Information and Control 6, № 3, 1963, стр. 306—309.
89. Г. Сиромони, Энтропия прозы на языке тамили (G. Siromoney, Entropy of Tamil prose), Information and Control 6, № 3, 1963, стр. 297—300; К. Р. Раджагопалан, Заметка об энтропии прозы на языке каннада (K. R. Rajagopalan, A note on entropy of Kannada prose), Information and Control 8, № 6, 1965, стр. 640—644; П. Баласубраманьям, Г. Сиромони, Заметка об энтропии прозы на языке телугу (P. Balasubrahmanyam, G. Siromoney, A note on entropy of Telugu prose), Information and Control 13, № 4, 1968, стр. 281—285; Б. С. Рамакришна, К. К. Наир, В. Н. Чиплункар, В. С. Атал, В. Рамачандран, Р. Субраманиан, Сравнительные эффективности индийских языков (B. S. Ramakrishna, K. K. Nair, V. N. Chiplunkar, B. S. Atal, V. Ramachandran, R. Subramanian, Relative efficiencies of Indian languages), Nature 189, № 4768, 1961, стр. 614—617.
90. Б. С. Рамакришна, Р. Субраманиан, Сравнительная эффективность английского и немецкого языков для передачи смыслового содержания (B. S. Ramakrishna, R. Subramanian, Relative efficiency of English and German languages for communication of semantic content), IRE Trans. on Inform. Theory IT-4, № 3, 1958, стр. 127—129.
91. Н. Рычкова, Лингвистика и математика, Наука и жизнь, № 9, 1961, стр. 76—77.
92. П. М. Алексеев, Частотные словари английского языка и их практические применения, в сборнике: «Статистика речи и автоматический анализ текста» (см. [80]), стр. 160—178.
93. Дж. Ципф, Поведение человека и принцип наименьшего усилия (G. K. Zipf, Human behavior and the principle of least effort), Cambridge (USA), Addison — Wesley, 1963.
94. Л. Апостель, Б. Мадельброт, А. Морф, Логика, речь и теория информации (L. Apostel, B. Mandelbrot, A. Morf, Logique, langage et théorie de l'information), Paris, Presses Universitaires de France, 1957.
95. Дж. А. Миллер, Речь и язык, в сборнике «Экспериментальная психология» (сост.— С. С. Стивенс), т. II, М., ИЛ, 1963, стр. 348—374.
96. Б. Мандельброт, Информационная теория статистической структуры языка (B. Mandelbrot, An informational

theory of the statistical structure of language), в сборнике: «Теория связи» (Communication Theory; см. [19]), стр. 486—502.

97. М. Григнетти, Заметка об энтропии слов в письменном английском тексте (M. C. Grignetti, A note on the entropy of words in printed English), Information and Control 7, № 3, 1964, стр. 304—306.
98. А. А. Пиотровская, Р. Г. Пиотровский, К. А. Разживин, Энтропия русского языка, Вопросы языкознания, № 6, 1962, стр. 115—130.
99. О. Л. Смирнов, А. В. Екимов, Энтропия русского телеграфного текста, Труды Ленинградск. ин-та авиацион. приборостроения, вып. 54 (системы обработки и передачи информации), Л., 1967, стр. 76—84.
100. Ф. Фрик, У. Самби, Язык наземного управления самолетом (F. C. Frick, W. H. Sumby, Control tower language), Journ. Acoust. Soc. Amer. 24, 1952, стр. 595—596.
101. Э. Л. Фриц, Дж. У. Грайер, Практическая связь: изучение потока информации в управлении воздушным движением (E. L. Fritz, G. W. Grier, Pragmatic communication: a study of information flow in air traffic control), статья в сборнике [40], стр. 232—243.
102. Т. Тарноци, О факторах, влияющих на различия значений энтропии языка (T. Tarnóczy, A jeloszlás es a hírtartalom nyelveket meghatározó tulajdonságáiról), Nyelvtudományi Közlemenyék 63, 1961, стр. 161—178.
103. А. М. Кондратов, Теория информации и поэтика (энтропия ритма русской речи), Проблемы кибернетики, вып. 9, М., 1963, стр. 279—286.
104. С. Маркус, Энтропия и поэтическая энергия (S. Marcus, Entropie et énergie poétique), Cahiers de linguistique théoretique et appliquée 4, 1967, стр. 171—180.
105. Сборник «Математика и поэзия» (Mathematik und Dichtung, сост.— H. Kreuzer, R. Gunzenhäuser), München, Nympfenbürger Verlaghandlung, 1965.
106. У. Дж. Пейсли, Влияние авторства, темы, структуры и времени написания на избыточность букв в английских текстах (W. J. Paisley, The effects of authorship, topic, structure and time of composition on letter redunancy in English texts), Journ. Verbal Learning and Verbal Behavior 5, № 1, 1966, стр. 28—34.
107. Дж. Берри, Некоторые статистические аспекты разговорной речи (J. Berry, Some statistical aspects of conversational speech), в сборнике: «Теория связи» (Communication Theory; см. [19]), стр. 392—401.
108. Б. Мандельброт, Закон Берри и определение „ударения", в сборнике: «Теория передачи сообщений», М., ИЛ, 1957, стр. 248—254.
109. В. А. Успенский, Одна модель для понятия фонемы, Вопросы языкознания, № 6, 1964, стр. 39—53.
110. Е. К. Черри, М. Халле, Р. Якобсон, К логическому описанию языков с точки зрения фонем (E. C. Cherry,

M. Halle, R. Jakobson, Toward the logical description of languages in their phonemic aspect), Language 29, № 1, 1953, стр. 34—46.

111. А. М. Пешковский, Десять тысяч звуков, Сборник статей. Л.— М., ГИЗ, 1925, стр. 167—191.
112. Л. Р. Зиндер, О лингвистической вероятности, в сборнике: «Вопросы статистики речи», Л., Изд-во ЛГУ, 1958, стр. 58—61.
113. Дж. У. Блэк, Информация звуков и фонетические диаграммы в одно- и двусложных словах (J. W. Black, The information of sounds and phonetic digrams of one- and two-syllable words), Journ. Speech Hearing Disorders 19, 1954, стр. 397—411; П. Дьенеш, О статистике устной английской речи (P. Denes, On the statistics of spoken English), Journ. Acoust. Soc. Amer. 35, № 6, 1963, стр. 892—904.
114. Ж. П. Гатон, М. Ламотт, Изучение статистики фонем и дифонем в устной французской речи (J. P. Haton, M. Lamotte, Étude statistique des phonèmes et diphonèmes dans le français parlé), Revue d'acoustique 4, № 16, 1971, 258—262.
115. В. Эндрес, Сравнение избыточности устной и письменной речи (W. Endres, A comparison of the redundancy in the written and spoken language), Пробл. управл. и теория информ., Приложен., 1973 (Труды 2-й Междунарооди. конфер. по теории информации; Цахкадзор, Арм. ССР, 2—8 сентября 1971 г.).
116. А. Фрадис, Л. Михайлеску, И. Воинеску, Энтропия и информационная энергия устной румынской речи (A. Fradis, L. Mihăilescu, I. Voinescu, L'entropie et l'énergie informationnelle de la langue roumaine parlée), Revue roumaine de linguistique 12, № 4, 1967, стр. 331—339.
117. Т. И. Ибрагимов, Исследование слоговой организации слов татарского языка, Ученые записки Казанского гос. ун-та 129, кн. 4 (Вероятностные методы и кибернетика, вып. VII), Казань, 1969, стр. 101—108.
118. И. Воинеску, А. Фрадис, Л. Михайлеску, Энтропия первого порядка фонем в речи больных афазией (I. Voinescu, A. Fradis, L. Mihăilescu, The first degree entropy of phonemes in aphasics), Revue roumaine de neurologie, 4, № 1, 1967, стр. 67—79; Энтропия второго порядка фонем и соотношение между порядковым номером и частотой пар фонем в речи больных афазией (Second order entropy of phonemes and rank-freguency relation of biphonematic groups in aphasics), Revue roumaine de neurologie, 5, № 2, 1968, стр. 111—120; Энтропия первого порядка слов в речи больных афазией (First order entropy of words in aphasics), Cybernetica 12, № 1, 1969 стр. 39—49; см. также А. Крейндлер, А. Фрадис, Афазия, гл. IX «Теория информации, речь и афазия» (A. Kreindler, A. Fradis, Afazia, Cap. IX «Theoria informatiei, limbajul si afazia»), Bucuresti, Ed. Acad. Rep. Social. Romania, 1970.
119. Р. Пинкертон, Теория информации и мелодия (R. C. Pinkerton, Information theory and melody), Scient. Amer. 194, № 2, 1956, стр. 77—86.

120. Ф. П. Брукс, А. Л. Хопкинс, П. Г. Нейман, У. В. Райт, Опыт по сочинению музыки (F. P. Brooks, A. L. Hopkins, P. G. Neumann, W. V. Wright, An experiment in musical composition), IRE Trans. on Electron. Comput. EC-6, № 3, 1957, стр. 175—182.
121. Г. Олсон, Г. Белар, Использование случайной вероятностной системы для помощи в музыкальных композициях (H. Olson, H. Belar, Aid to music composition employing a random probability system), Journ. Acoust. Soc. America 33, № 9, 1961, стр. 1163—1170.
122. Р. Х. Зарипов, Кибернетика и музыка, М., «Наука», 1971.
123. Дж. Е. Юнгблад, Стиль как информация (J. E. Youngblood, Style as information), Journ. Music Theory 2, № 1, 1958, стр. 24 и след.
124. Дж. Е. Коэн, Теория информации и музыка (J. E. Cohen, Information theory and music), Behav. Sci., 7, № 2, 1962, стр. 137—163.
125. Г. Сиромони, К. Р. Раджагопалан, Стиль как информация в карнатической музыке (G. Siromoney, K. R. Rajagopalan, Style as information in Karnatic music), Journ. Music Theory 8, № 2, 1964, стр. 267—272.
126. Л. Хиллер, Дж. Бишем, Исследования в области музыки с использованием электроники (L. Hiller, J. Beauchamp, Research in music with electronics), Science 150, № 3693, 1965, стр. 161—169.
127. М. Роланд, Уменьшение информации из-за зависимости между несколькими одновременными источниками информации и из-за перехода к марковским цепям высокого порядка, исследованное на примерах музыкальных произведений (M. Roland, Die Entropieabnahme bei Abhängigkeit zwischen mehreren simultanen Informationsquellen und bei Übergang zu Markoff-Ketten höherer Ordnung, untersucht an musikalischen Beispielen), Forschungsber. Landes Nordheim-Westfalen, 1967, № 1768, стр. 39, 41, 43—44, 79—80.
128. Д. С. Лебедев, И. И. Цуккерман, Телевидение и теория информации, М., «Энергия», 1965.
129. У. Ф. Шрейбер, Измерение трехмерных распределений вероятностей для телевизионных изображений (W. F. Schreiber, The measurement of third order probability distributions of television signals), IRE Trans. on Inform. Theory IT-2, № 3, 1956, стр. 94—105.
130. Д. С. Лебедев, Е. И. Пийль, Экспериментальные исследования статистики телевизионных сообщений, Техника кино и телевидения, № 3, 1959, стр. 37—39.
131. Дж. О. Лимб, Энтропия квантованных телевизионных сигналов, (J. O. Limb, Entropy of quantised television signals), Proc. Inst. Elec. Eng. (Proc. IEE) 115, № 1, 1968, стр. 16—20.
132. П. Нейдгардт, Введение в теорию информации (P. Neidhardt, Einführung in die Informationstheorie), Berlin, VEB Verlag Technik, 1957.
133. Н. С. Цаннес, Р. В. Спенсер, А. Дж. Каплан, Об оценке энтропии случайных полей (N. S. Tsannes,

R. V. Spencer, A. J. Kaplan, On estimating the entropy of random fields), Information and Control 16, № 1, 1970, стр. 1—6.

134. С. Дейч, Заметка о некоторых статистических характеристиках машинописного или печатного текста (S. Deutsch, A note on some statistics consering typewritten or printed material), IRE Trans. on Inform. Theory IT-3, № 2, 1957, стр.136—143.
135. Г. А. Кайзер, К вопросу об энтропии текстов, напечатанных на пишущей машинке (G. A. Kayser, Zur Entropie schreibmaschinengeschriebener Textvorlagen), Nachrichtentechn. Zeitschr. (NTZ) 13, № 5, 1960, стр. 219—224.
136. У. С. Майчел, Статистическое кодирование для передачи текста и рисунков (W. S. Michel, Statistical encoding for text and picture communications), Commun. and Electr., № 35, 1958, стр. 33—36.
137. В. А. Гармаш, Н. Е. Кириллов, Экспериментальное исследование статистики фототелеграфных сообщений, Научн. доклады высш. школы (радиотехника и электроника), № 1, 1959, стр. 37—42.
138. Р. Р. Васильев, О статистических методах передачи фототелеграмм, Радиотехника и электроника 2, № 2, 1957, стр. 136—143.
139. В. Г. Фролушкин, Анализ статистической структуры текстовых фототелеграмм, Электросвязь, № 5, 1959, стр. 63—70.
140. У. Х. Фой, Энтропия простых линейных чертежей (W. H. Foy, Entropy of simple line drawings), IEEE Trans. on Inform. Theory IT-10, № 2, 1964, стр. 165—167.
141. Ф. Е. Темников, В. А. Афонин, В. И. Дмитриев, Теоретические основы информационной техники, М., «Энергия», 1971.
142. Быстрая связь (Fast data communication), Sci. News Letters 83, № 1, 1963, стр. 5.
143. Г. Якобсон, Информационная пропускная способность человеческого глаза (H. Jacobson, The informational capacity of the human eye), Science 113, № 2933, 1951, стр. 292—293.
144. Г. Якобсон, Информация и ухо человека (H. Jacobson, Information and the human ear), Journ. Acoust. Soc. Amer. 23, № 4, 1951, стр. 463—471.
145. Г. Шобер. Основополагающие замечания о применимости теории информации к оптике (H. Schober, Grundlegende Bemerkungen dur Anwendbarkeit der Informationstheorie auf die Optik), Wiss. Zeitschr. Hochschule Elektrotechn. Ilmenau 3, № 3—4, 1957, стр. 273—276.
146. Д. Г. Келли, Информационная пропускная способность единичного зрительного канала (D. H. Kelly, Information capacity of a single retinal channel), IRE Trans. on Inform. Theory IT-8, № 3, 1962, стр. 221—226.
147. К. Кюпфмюллер, Переработка информации человеком (K. Küpfmüller, Informationsverarbeitung durch den

Menschen), Nachrichtentechnische Zeitschr. (NTZ), № 2, 1959, стр. 68—74.
148. Е. Б. Ньюман, Люди и информация: точка зрения психолога (E. B. Newman, Men and information: a psychologist's view), Nuovo Cimento Suppl. 13, № 2, 1959, стр. 539—559.
149. Г. Сиклаи, Изучение скорости зрительного восприятия (G. C. Sziklai, Some studies in the speed of visual perception), IRE Trans. on Inform. Theory IT-2, № 3, 1956, стр. 125—128.
150. Г. Квастлер, Изучение пропускной способности человеческого канала (H. Quastler, Studies of human channel capacity), в сборнике: «Information Theory, Third London Symposium» (ред.— C. Cherry), London, Buttorworths, 1956, стр. 361—371.
151. Г. Гамов, Возможное отношение между дезоксирибонуклеиновой кислотой и белковыми структурами (G. Gamow, Possible relation between deoxyribonucleic acid and protein structures), Nature 173, 1954, стр. 318.
152. Г. Гамов, М. Ичас, Статистическая связь между составом белка и рибонуклеиновой кислоты (G. Gamow, M. Yčas, Statistical correlation of protein and ribonucleic acid composition), Proc. Nat. Acad. Sci. USA 41, 1955, стр. 1011—1019.
153. Ф. Крик, Дж. Гриффит, Л. Оргел, Коды без запятых (F. H. C. Crick, J. S. Griffith, L. E. Orgel, Codes without commas), Pros. Nat. Acad. Sci. USA 43, 1957, стр. 416—421.
154. С. В. Голомб, Л. Р. Велч, М. Дельбрюк, Строение и свойства кодов без запятой, журнал переводов «Математика» 4, № 5, 1960, стр. 137—160.
155. Г. Гамов, А. Рич, М. Ичас (или Икас), Проблема передачи информации от нуклеиновых кислот к белкам, в сборнике: «Вопросы биофизики», М., ИЛ, 1957, стр. 205—263; Г. Гамов, М. Ичас, Криптографический подход к проблеме синтеза белка, в сборнике [14], стр. 66—71; М. Ичас, Белковый текст, там же, стр. 72—103.
156. Ф. Крик, К расшифровке генетического кода, в сборнике: «Живая клетка», М., ИЛ, 1962, стр. 203—222; Ф. Крик, Генетический код (I), в сборнике: «Структура и функция клетки», М., «Мир», 1964, стр. 9—23; М. Ниренберг, Генетический код (II), там же, стр. 24—41; Ф. Крик, Генетический код (III) (F. H. C. Crick, The genetic code: III), Scientific American 215, № 4, 1966, стр. 55—61; М. В. Волькенштейн, Проблема генетического кода, «Природа», № 9, 1968, стр. 20—29.
157. М. Ичас, Биологический код, М., «Мир», 1971.

§ 4

158. К. Шеннон, Некоторые результаты теории кодирования для канала с шумами, в книге: «Работы по теории информации и кибернетике», (см. [1]), стр. 433—460.

159. П. Элайес, Кодирование для двух каналов с шумами, в сборнике: «Теория передачи сообщений» (см. [108]), стр. 114—138.
160. Р. Л. Галлагер, Простой вывод теоремы кодирования и некоторые применения, Кибернетический сборник (новая серия), вып. 3., М., «Мир», 1966, стр. 50—90.
161. Р. Л. Добрушин, Асимптотические оценки вероятности ошибки при передаче сообщения по дискретному каналу связи без памяти с симметричной матрицей вероятностей перехода, Теор. вероятн. и ее примен. 7, № 3, 1962, стр. 283—311.
162. К. Шеннон, Пропускная способность канала с шумом при нулевой ошибке, в книге: «Работы по теории информации и кибернетике» (см. [1]), стр. 464—487.
163. С. К. Заремба, Замечание к основной теореме для дискретного канала с шумами, в сборнике: «Теория передачи сообщений» (см. [108]), стр. 28—31.
164. Э. Н. Гилберт, Сравнение алфавитов сигналов (E. N. Gilbert, A comparison of signalling alphabets), Bell System Techn. Journ. 31, № 3, 1952, стр. 502—522.
165. Д. Слепян, Теория кодирования (D. Slepian, Coding theory), Nuovo Cimento Suppl., Ser. X. 13, № 2, 1959, стр. 373—388.
166. Г. А. Барнард, Простые доказательства простых случаев теоремы кодирования, в сборнике: «Теория передачи сообщений» (см. [108]), стр. 32—42.

§ 5

167. Н. Левинсон, Теория кодирования: противоречащий пример к принадлежащей Г. Х. Харди концепции прикладной математики (N. Levinson, Coding theory: a counterexample to G. H. Hardy's conception of applied mathematics), Amer. Math. Monthly. 77, № 3, 1970, стр. 249—258.
168. У. Питерсон, Коды, исправляющие ошибки, М., «Мир», 1964.
169. Э. Берлекэмп, Алгебраическая теория кодирования, М., «Мир», 1971.
170. Р. Л. Добрушин. Теория оптимального кодирования информации, в сборнике: «Кибернетика на службу коммунизму» (ред.— А. И. Берг), т. 3, 1966, стр. 13—45.
171. П. Элайс (или Элайес), Кодирование и декодирование, в сборнике: «Лекции по теории связи» (ред.— Е. Дж. Багдади), М., «Мир», 1964, стр. 289—317.
172. Р. В. Хэмминг, Коды с обнаружением и исправлением ошибок, в сборнике: «Коды с обнаружением и исправлением ошибок», М., ИЛ, 1956, стр. 7—23.
173. Р. Р. Варшамов, Оценка числа сигналов в кодах с коррекцией ошибок, Докл. Акад. Наук СССР, 117, № 5, 1957, стр. 739—741.
174. Р. К. Боуз, Д. К. Рой-Чоудхури, Об одном классе двоичных групповых кодов с исправлением ошибок, Кибернетический сборник, вып. 2, 1961, стр. 83—94; Дальнейшие ре-

зультаты относительно двоичных групповых кодов с исправлением ошибок, Кибернетический сборник, вып. 6, 1963, стр. 7—12.
175. А. Хоквингем, Коды, исправляющие ошибки (А. Hocquenghem, Codes correcteurs d'erreurs), Chiffres 2, 1959, стр. 147—156.
176. Р. Дж. Галлагер, Коды с малой плотностью проверок на четность, М., «Мир», 1966.
177. Д. Слепян, Класс двоичных сигнальных алфавитов, в сборнике: «Теория передачи сообщений» (см. [108]), стр. 82—113.
178. Р. Л. Добрушин, Асимптотическая оптимальность групповых и систематических кодов для некоторых каналов, Теор. вероятн. и ее примен. 8, № 1, 1963, стр. 52—66.
179. Г. Дригас, Теория кодирования для симметричных каналов (H. Drygas, Verschlüsselungstheorie für symmetrische Kanäle), Zeitschr. für Wahrscheinlichkeitstheorie und verw. Gebiete 4, 1965, стр. 121—143.
180. Э. М. Габидулин, Границы для вероятности ошибки декодирования при использовании линейных кодов без памяти, Пробл. передачи информ. 3, № 2, 1967, стр. 55—62.
181. Р. Алсведе, Групповые коды не позволяют достичь шенноновской пропускной способности для общих дискретных каналов (R. Ahlswede, Group codes do not achieve Shannon's channel capacity for general discrete channels), Ann. Mathem. Stat. 42, № 1, 1971, стр. 224—240.
182. Г. Е. Сакс, Исправление кратных ошибок с помощью проверок на четность (G. E. Sacks, Multiple error correction by means of parity checks), IRE Trans. on Inform. Theory IT-4, № 4, 1958, 145—147.
183. В. Д. Колесник, Е. Т. Мирончиков, Декодирование циклических кодов, М., «Связь», 1968.
184. Д. Форни, Каскадные коды, М., «Мир», 1970.
185. Д. Горенстейн, У. Питерсон, Н. Цирлер, Квазисовершенность кодов Боуза — Чоудхури с исправлением двух ошибок, Кибернетический сборник, вып. 6, 1963, стр. 20—24.
186. А. Тиетявяйнен, А. Перко, Не существует неизвестных совершенных двоичных кодов (A. Tietäväinen, A. Perko, There are no unknown perfect binary codes), Ann. Univ. Turku, Ser. A, I, № 148, 1971, стр. 3—10.
187. В. А. Зиновьев, В. К. Леонтьев, О совершенных кодах, Пробл. передачи информ. 8, № 1, 1972, стр. 26—35.

ИМЕННОЙ УКАЗАТЕЛЬ[1])

[1]) В настоящем указателе вслед за каждой фамилией перечислены все страницы, на которых либо упоминается данное лицо, либо имеются ссылки на его работы.

АЛФАВИТНЫЙ УКАЗАТЕЛЬ

Акива Моисеевич Яглом,
Исаак Моисеевич Яглом
ВЕРОЯТНОСТЬ И ИНФОРМАЦИЯ

М., 1973 г., 512 стр. с илл.

Редакторы: *С. З. Стамблер, В. В. Абгарян*

Техн. редактор *К. Ф. Брудно*
Корректоры: *З. В. Автонеева, Л. С. Сомова*

Сдано в набор 18/XII 1972 г.
Подписано к печати 28/V 1973 г.
Бумага 84×108¹/₃₂. Физ. печ. л. 16. Усл. печ. л. 26,88.
Уч.-изд. л. 28,73. Тираж 50 000 экз. Т-08147.
Цена книги 1 р. 11 к. Заказ № 1869

Издательство «Наука»
Главная редакция
физико-математической литературы
117071, Москва В-71, Ленинский проспект, 15

Типография № 2 издательства «Наука»
Москва Г-99, Шубинский пер., 10