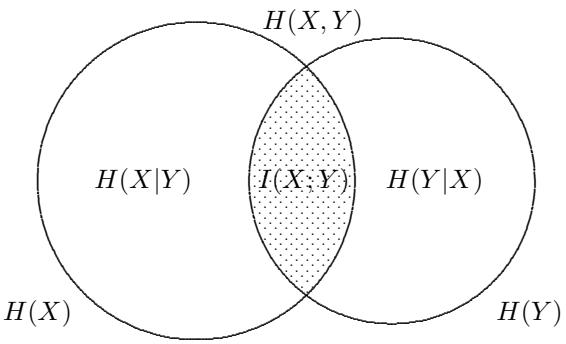
Ю. В. Свирид

ОСНОВЫ ТЕОРИИ ИНФОРМАЦИИ

Курс лекций

Издание 2-е, исправленное и дополненное



МИНСК
БГУ
2010

УДК 519.7(075.8)
ББК 22.18я73
С24

Печатается по решению
редакционно-издательского совета
Белорусского государственного университета

Р е ц е н з е н т ы:
доктор физико-математических наук, профессор *М. М. Ковалев*;
доктор технических наук, профессор *В. А. Чердынцев*

Свирид, Ю. В.

С24 Основы теории информации: курс лекций / Ю. В. Свирид. – 2-е изд., испр. и доп. – Минск : БГУ, 2010 – 151 с.

ISBN 978-985-518-378-6.

Курс лекций содержит введение в теорию информации и ее приложения. Понятие информации рассмотрено с синтаксической, семантической и прагматической точек зрения, а так же в связи с колмогоровской сложностью объектов. На основе понятия о типичных последовательностях доказаны три фундаментальные теоремы кодирования — о сжатии данных, о передаче данных и о сжатии и передаче данных. Рассмотрены некоторые важные каналы связи и вычислена их пропускная способность. Приведено описание и доказательство оптимальности ряда алгоритмов сжатия данных. Даны теоретико-информационные основы криптологии.

Для студентов и аспирантов математических, физических и инженерных специальностей университета.

УДК 519.7(075.8)
ББК 22.18я73

ISBN 978-985-518-378-6

Посвящается

моим родителям, *без которых этой книги определенно бы не было*[a],

Марине, *которая после стольких лет все еще остается загадкой для меня*[b],

Ксюше, *не раз отвлекавшей меня от компьютера, после чего у меня часто появлялись дельные мысли*[c].

[a] Последовательность {Мои родители, Я, Эта книга} образует цепь Маркова (с. 47).

[b] В терминах теории информации это означает, что поведенческая энтропия моей жены превышает мою способность к снятию неопределенности (1.30).

[c] Уровень помех не является определяющим фактором (раздел 2.5), важно чтобы скорость передачи информации была меньше пропускной способности канала!

От автора

В последнее время издавалось очень мало книг и статей на русском языке, которые бы касались проблем теории информации. Это связано в основном с тем, что специалисты советской школы, которая в 60–80-е гг. XX столетия наряду с американской была ведущей в мире и, безусловно, составляла одну из важных частей противостояния двух мировых систем, с началом перестройки получили возможность активно ездить за рубеж, писать книги для ведущих мировых издательств и публиковать новые результаты в международных журналах.

Пострадала от открытия границ и учебная литература в этой области на русском языке. Самым лучшим советским учебником по теории информации остается изданная в 1982 г. книга В. Д. Колесника и Г. Ш. Полтырева [5], а для второго чтения очень хороша монография Р. Л. Стратоновича [3], где теория информации излагается нетрадиционно, с позиций статистической термодинамики.

В англоязычной книге Т. М. Ковера и Дж. А. Томаса [7] собраны все важнейшие идеи, появившиеся в теории информации к началу 90-х гг. прошлого века, изложенные в едином стиле и единой терминологии. Это делает ее пригодной даже для первого чтения, поскольку фундаментальные понятия раскрыты в ней просто и доступно.

Хорошим введением в предмет являются: книга Р. Йоханнессона [8], охватывающая не только принципиальные разделы теории информации, но и частично теории кодирования и криптологии, книга Д. Хэнкерсона, Г. Харриса и П. Джонса [15], где в подробностях разобраны многие алгоритмы сжатия данных, появившиеся в последнее время русскоязычные пособия А. А. Духина [20], В. В. Панина [21] и Б. Д. Кудряшова [24], а также книга Э. Дезурвира [22], которая наряду с классической рассматривает и квантовую теорию информации.

С приложениями теории информации (от автоматизированных систем управления и теории динамических систем до биологии, теории эволюции и искусства) можно познакомиться по книгам И. М. Когана [4], В. И. Дмитриева [6], Д. С. Чернавского [17] и Д. Эвери [18]. Философско-естественнонаучное введение в понятие информации дает Х. Люре в [14], а в сборнике статей [23] дан обзор современных

достижений в создании единой концепции информации. В монографии А. И. Демина [19] сделана попытка понимания информации как сущности, сравнимой с пространством, временем и энергией.

Обзор самых существенных результатов, достигнутых в теории информации за 50 лет ее существования (т. е. к 1998 г.), дан С. Вердю в статье [9] с подробной библиографией из 440 позиций.

В отличие от работ по теории информации, книги по криптологии на русском языке (одними из лучших примеров являются [11], [13] и [16]) издавались в последнее время достаточно обильно. Именно наличием многих изданий на русском языке вызвано то, что в этой книге криптологии посвящена только одна глава, в которой изложены ее фундаментальные основы, заложенные Шенноном. Подробный очерк истории криптографии можно найти в [11].

Отдельного внимания, конечно, заслуживают пионерские статьи основоположника теории информации, кодирования и криптологии Клода Шеннона (1916–2001) [1] и [2]. Они являются редким примером, когда один человек почти на «пустом месте» в конце 40-х гг. сумел заложить фундаментальные основы науки, играющей все возрастающую роль с развитием общества и коммуникации.

Книга, которую Вы держите в руках, написана по результатам чтения автором лекций в Белорусском государственном университете и предназначена студентам старших курсов и аспирантам университетов, математикам и инженерам, желающим познакомиться с увлекательным предметом теории информации. Она призвана восполнить пробел в русскоязычной учебной литературе в данной области.

Я пытался излагать материал как можно проще и доступнее, однако определенная математическая культура для чтения книги все-таки требуется. Особо необходимо знание элементарных понятий и фактов из теории вероятностей и случайных процессов, с чем можно ознакомиться по [10] и [12]. Математик найдет в книге строгие доказательства приводимых фактов, а инженеру будут по душе приложения красивых математических концепций.

Книга состоит из глав, каждая из которых разбита на разделы. Впервые встречающиеся или вводимые понятия выделены **жирным шрифтом**. Эти понятия включены в предметный указатель в конце книги. Места, где стоит логическое ударение, подчеркнуты. Примеры приведены шрифтом с прямоугольной гарнитурой, тексты определений, теорем и лемм выделены *курсивом*, а знак □ означает окончание доказательства.

Для закрепления и расширения излагаемого материала в конце каждой главы приведены задачи. Их уровень сложности по пятибалльной шкале указан в квадратных скобках после номера, первые две цифры которого соответствуют номеру раздела, к которому задача относится. Задачи с уровнем сложности 1–2 балла являются прямым отображением материала книги и могут быть достаточно просто решены после прочтения и усвоения соответствующего раздела. Уровень сложности в 3–4 балла предполагает более глубокое понимание предмета с опорой на результаты различных разделов и умением провести ряд выкладок. Задачи со сложностью в 5 баллов представляют собой исследовательские проблемы или требуют значительных усилий для решения.

Всего имеется 3 задачи со сложностью 1 балл, 41 задача со сложностью 2 балла, 26 задач по 3 балла, 10 задач по 4 и 9 по 5 баллов. Общее количество баллов, которое можно набрать, решая задачи, равно 248, причем может считаться, что студент данный курс усвоил, если он правильно решил задачи с суммарным количеством баллов 125.

Введение

В повседневной жизни мы часто сталкиваемся с понятием **информации**. Обычно мы имеем в виду при этом какие-либо сведения, получаемые человеком из различных источников (люди, газеты, телевидение, радио), или сведения об окружающем мире и протекающих в нем процессах, воспринимаемые специальными устройствами или человеком, вызывающие при этом определенные его состояния (удивление, уныние, радость и т. п.). Социологи даже говорят об **информационном обществе** — новой стадии развития общества, в которой основным объектом деятельности является торговля информацией, а ведущая роль отводится сфере услуг, науке и образованию — то есть тому, что способствует созданию новой информации.

Этимологически слово информация восходит к ключевому понятию древнегреческой философии о **форме** (εἶδος, ἰδέα). У Гомера эйдос — это то, что видно, наружность, внешний вид (например, бронзовой статуи), у Платона форма (эйдос, идея) является вечным и неизменным прообразом вещей, всего преходящего и изменчивого бытия. Идеи (формы, эйдосы) у Аристотеля — это внутренние движущие силы вещей, неотделимые от них. Все вещи состоят, по Аристотелю, из материи и формы, причем материя – это лишь возможность, без которой не может существовать никакая вещь, а становится материя чем-то конкретным только в соединении с формой, придающей ей определенность и превращающей возможность в действительность.

Сегодняшнее понятие формы происходит от перевода греческого слова εἶδος на латинский язык — forma. Существительным informatio в латинском языке выражается действие формы, образа или внешности в том числе в смысле разъяснения, наставления, поучения, а глагол informare означает разъяснять, уведомлять, передавать какие-либо сведения.

В качестве формы, передающей какие-либо сведения, всегда используется некоторая система знаков. У людей она развивалась начиная от жестов и звуков в первобытном обществе до естественных или искусственный языков, до языка искусства в наши дни.

Наука, исследующая свойства знаков и знаковых систем, — семиотика — связывает с понятием знака четыре аспекта — синтаксический,

сигматический, семантический и прагматический. Эти аспекты могут быть однозначно перенесены и на понятие информации, так как ее передача и восприятие всегда основаны на той или иной системе знаков.

- **Синтаксическая часть** информации касается возможности появления различных знаков, правил образования их сочетаний и других внутренних свойств знаков безотносительно к их значениям или полезности.
- **Сигматическая часть** информации касается выбора тех или иных знаков для обозначения объектов.
- **Семантическая часть** информации касается смысла знаков и их сочетаний.
- **Прагматическая часть** информации касается эффекта, производимого появлением того или иного знака или их комбинации в смысле его полезности для тех или иных целей.

Другими словами, синтаксическая и сигматическая части информации касаются ее формы, а семантическая и прагматическая части — ее содержания.

Рассмотрим следующие высказывания:

1. Студентка А приглашена студентом В на свидание.
2. Студент В пригласил студентку А на свидание.
3. Студент В пригласил студентку А на свидание и подарил ей цветы.

Высказывания 1 и 2 синтаксически различны, но идентичны семантически и прагматически, поскольку они имеют один и тот же смысл и одинаково полезны (или бесполезны). Высказывания 1 и 3 различны не только синтаксически, но также семантически (поскольку высказывание 3 несет больше информации) и прагматически (поскольку вполне может существовать студентка С, которую высказывание 1 просто расстроит, а высказывание 3 приведет в истерику). Кроме того, все три высказывания идентичны сигматически, поскольку они используют одну и ту же систему знаков — русский язык.

В математической дисциплине, называемой **теорией информации**, изучается только синтаксическая часть информации. Подобно тому как, находя сумму чисел 7 и 8, математик абстрагируется от конкретных объектов, стоящих за этими числами, мы здесь полностью абстрагируемся от смысла или полезности информации и концентрируемся на вопросах измерения ее количества, фундаментальных границах для ее сжатия и надежной передачи. С сигматической точки

зрения здесь также обосновывается, например, применение бита, как знака, наиболее удобного для передачи и хранения информации.

Семантические и прагматические аспекты информации являются в своей основе предметом изучения в философии, лингвистике, психологии, медицине и биологии. Однако до настоящего времени не существует единого подхода, удовлетворительно описывающего данные части информации (на некоторых таких подходах мы остановимся в разделе 1.3), как нет и философской концепции понятия информации в единстве всех ее аспектов.

Следует ли, например, считать информацию некоторой объективной субстанцией, отличной от материи и энергии? Основоположник кибернетики Норберт Винер сказал в этой связи: «Информация есть информация, не материя, не энергия. Никакой материализм, который не допускает этого, не может существовать в настоящее время».

Или информация является свойством материи? Советский академик В. М. Глушков характеризовал информацию как меру неравномерности в распределении энергии (или материи) в пространстве и времени, то есть рассматривал ее как свойство материи. По Глушкову, информация существует, как только существуют материальные тела (например, звезды), а следовательно, и вызываемая ими неравномерность.

Информация носит субъективный характер, поскольку ею может считаться лишь то, что может быть понято и что создает новую информацию (Карл фон Вайцзеккер)? Является она отраженным разнообразием материального мира (А. Д. Урсул), или каждая материальная сущность имеет информационное происхождение (Джон Уилер)?

В любом случае понятие информации лежит где-то на стыке естественных и гуманитарных наук и, возможно, поможет создать для них единую основу. Оно выкристаллизовывается в одно из основных понятий естествознания наряду с понятиями материи и энергии, которые уже давно являются базисом всех естественных наук.

В физике давно научились измерять количество материи и энергии. О том же, как адекватно измерить количество информации, мы знаем только с 1948 г., когда появилась статья [1] Клода Шеннона «A Mathematical Theory of Communication», название которой при переиздании в виде книги год спустя было изменено на «The Mathematical Theory of Communication». То есть Шеннон к этому времени понял, что его результаты в теории информации имеют фундаментальное значение для передачи сообщений (коммуникации). Однако в научном ми-

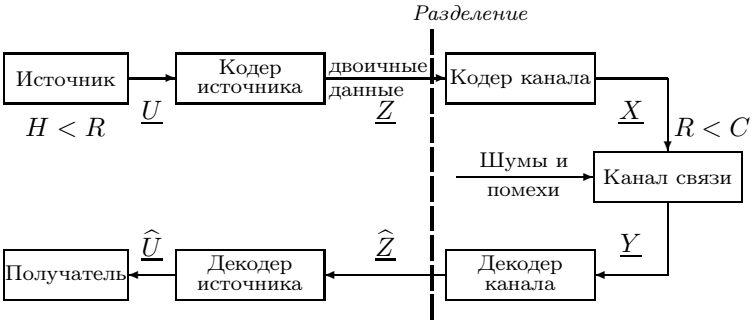


Рис. 1. Теоремы Шеннона в системе передачи информации

ре эти результаты были крайне новыми, а идеи необычными, поэтому потребовались годы для их осознания. И только через 35–40 лет фундаментальные результаты теории информации вошли в повседневное использование.

Например, каждый компакт-диск записан так, что наличие на нем царапин не всегда приводит к ухудшению качества звучания, а современные модемы для компьютеров содержат протоколы для исправления ошибок и сжатия данных. Или мобильные телефоны — в стандарте GSM они применяют алгоритм Витерби и сверточный код для повышения качества передаваемой речи, а в новом стандарте UMTS их работа основана на турбокодах.

Выделим в системе передачи информации наиболее существенные звенья, необходимые для понимания процессов, происходящих при передаче информации (рис. 1). **Источником** информации может быть природа, человек, ЭВМ и т. д. Данные, в которых заключена информация источника, называют **сообщением**. Оно может представлять собой результаты наблюдения какого-либо явления, устные или письменные фразы, последовательность двоичных символов и т. д. На протяжении всей книги мы будем обозначать данные источника вектором (последовательностью) $\underline{U}$.

Источник и получатель разделены в пространстве (при передаче информации) или во времени (при ее хранении) **каналом связи**. Как правило, работа канала несовершенна, и его выход может существенно отличаться от входа по причине наличия в канале шумов и/или помех, поэтому после всей необходимой обработки получатель имеет оценку $\widehat{\underline{U}}$ переданной последовательности $\underline{U}$.

Свойства источника и канала считаются фиксированными. Источнику ставится при этом в соответствие параметр H, называемый **энтропией**, а каналу связи — параметр C, называемый **пропускной способностью** канала связи. На понятиях энтропии и пропускной способности мы подробно остановимся в главах 1 и 2. Теперь лишь поясним энтропию источника как среднее количество информации, производимое им в единицу времени, а пропускную способность канала — как максимальное количество информации, способное быть переданным через канал в единицу времени при условии, что существует принципиальная возможность восстановить переданное с любой сколь угодно малой наперед заданной вероятностью ошибки. Пропускную способность канала связи можно сравнить с пропускной способностью автомагистрали, каждую секунду пропускающей максимум C автомобилей; попытка проехать за ту же секунду $C + 1$ автомобилям неизбежно приведет к аварии.

Если мы, с одной стороны, имеем фиксированную энтропию источника, а с другой стороны, фиксированную пропускную способность канала связи, то в выборе преобразований, которым подвергнутся данные источника перед поступлением в канал связи, мы не ограничены никак. Такие преобразования, называемые **кодированием** и осуществляемые в **кодерах**, необходимы всегда, когда сообщение содержит **избыточность** и/или не является одним из возможных входов канала. Например, письменный текст не может быть непосредственно передан по радио. Аналогично если сообщение было закодировано перед поступлением в канал, то необходима соответствующая обработка выхода канала для перевода информации в форму, приемлемую для получателя. Данный процесс называется **декодированием** и осуществляется в соответствующих **декодерах**. Кодеры и декодеры называют **кодеками**.

Шеннон сформулировал три теоремы, касающиеся кодирования данных источника и их передачи через канал связи.

Теорема о сжатии данных (выполняемом кодером источника) говорит, что каждый источник может как угодно точно быть задан с помощью R бит в единицу времени, если $H < R$, но не наоборот. Этой теореме посвящен раздел 2.2.

Теорема о передаче данных говорит, что R случайно выбранных бит в единицу времени могут быть переданы через канал и приняты с любой наперед заданной точностью в случае, если $R < C$, но не наоборот (раздел 2.5). Шеннон доказал, что важно не столько то, насколько

много шумов в канале, сколько то, как мы закодируем информацию (при выполнении условия $R < C$).

Из условий $H < R$ и $R < C$ естественно вытекает условие $H < C$, которое должно выполняться для любой надежной системы передачи информации. Об этом говорит теорема о сжатии и передаче данных, из которой, кроме того, следует, что проблема донесения данных источника к получателю может быть представлена без потери качества двумя независимо друг от друга решаемыми проблемами:

- Представление данных источника в виде двоичных последовательностей $\underline{Z}$ (**кодирование источника**).
- Передача двоичных последовательностей $\underline{X}$ через канал связи (**кодирование канала**), на выходе которого имеется в общем случае отличная от $\underline{X}$ и, возможно, недвоичная последовательность $\underline{Y}$. Ее декодирование дает оценки $\widehat{\underline{Z}}$ и $\widehat{\underline{U}}$.

То есть кодер канала может конструироваться отдельно от кодера источника, что естественно облегчает использование канала для различных источников. Данную теорему мы докажем в разделе 2.6.

Единственным недостатком теорем Шеннона является то, что он не указал, как необходимо сжимать и/или кодировать информацию при передаче для приема ее с любой заданной точностью, а только указал на такую принципиальную возможность. Иначе бы наука, родившись в 1948 г., одновременно бы и умерла.

Тому, как можно эффективно представлять сообщения источника в виде двоичных данных, посвящена глава 3. Под эффективностью здесь понимается такая возможность максимального сжатия сообщения (или удаление из него избыточности), что сообщение можно после него восстановить без каких-либо потерь. Например, вычеркивание некоторого количества букв в печатном тексте не обязательно приведет к потере его смысла, то есть текст заведомо содержит избыточность. Поэтому, имея канал связи с ограниченной пропускной способностью, нет смысла передавать текст в исходном виде: избыточность стоит удалить. Упор в данной главе сделан на описание идей сжатия данных и доказательство оптимальности рассматриваемых алгоритмов, а не на их практическую реализацию, с которой можно ознакомиться в обширной литературе.

Для возможности восстановления информации после ее искажения шумами и/или помехами в канале связи в нее до передачи в канал вводят избыточность по специальным правилам. Проблемы, свя-

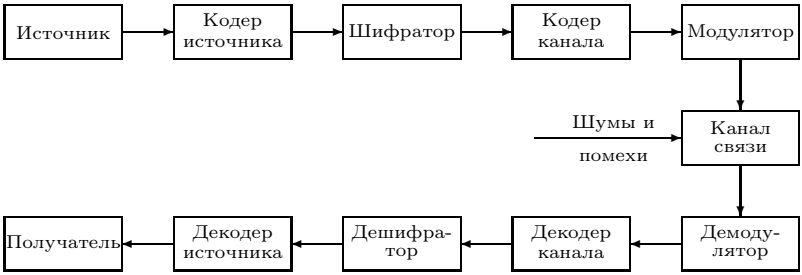


Рис. 2. Структурная схема системы передачи информации

занные с конкретными способами введения избыточности, являются объектом исследования **теории кодирования** и остаются за рамками этой книги.

Основам криптологии посвящена глава 4. Как мы увидим, избыточность играет и здесь существенную роль: чем менее избыточные данные поступают на шифратор, тем эффективнее шифрование, то есть если в системе передачи информации требуется шифрование, то место шифратора — между кодером источника (удаляющего избыточность) и кодером канала (вводящего ее), но ни в коем случае не после источника или кодера канала.

Поэтому более общая схема системы передачи информации изображена на рис. 2, где показаны также **модулятор**, преобразующий данные с выхода кодера канала в сигналы, способные быть переданными по каналу связи (например, в радиоволны специальной формы), и **демодулятор**, преобразующий сигналы из канала связи в данные, способные обрабатываться декодером канала. Модулятор и демодулятор называют **модемом**.

Глава 1

Что такое информация?

1.1. Некоторые понятия из теории вероятностей

Пусть некоторая **дискретная случайная величина** X принимает значения из **алфавита** $\mathcal{A}_X = \{x_1, x_2, \ldots, x_L\}$. Количество элементов в множестве $\mathcal{A}_X$ называется **мощностью множества** $\mathcal{A}_X$ и обозначается $\operatorname{card} \mathcal{A}_X$. В нашем случае $\operatorname{card} \mathcal{A}_X = L$.

Функцией вероятности $f(x)$ для X называется отображение множества $\mathcal{A}_X$ на действительные числа такое, что

$$f(x_i) = P(X = x_i), \quad i = 1, 2, \ldots, L,$$

где $P(X = x_i)$ означает вероятность того, что случайная величина X примет значение x_i.

Справедливо, что $f(x_i) \geq 0$ для всех i и что

$$\sum_{i=1}^{L} f(x_i) = 1.$$

Если X — случайная величина, для которой **функция распределения** $F(x) = P(X \leq x)$ является непрерывной функцией от x, то X называется **непрерывной случайной величиной**.

Пусть $f(x) = F'(x)$, причем производная существует. Тогда если

$$\int_Q f(x)\, dx = 1,$$

где Q — множество, на котором $f(x) > 0$, то функция $f(x)$ называется **плотностью вероятности** случайной величины X. Множество Q называется **ненулевым множеством** X.

Плотность вероятности для непрерывных случайных величин является аналогом функции вероятности для дискретных случайных величин.

Случайная величина может быть многомерной, то есть ее значения могут образовывать двойки, тройки и т. д. случайных величин. Например, пусть $Z = \{X, Y\}$, где X и Y являются случайными величинами с функциями вероятности $f(x)$ и $f(y)$ и алфавитами $\mathcal{A}_X = \{x_1, x_2, \ldots, x_L\}$ и $\mathcal{A}_Y = \{y_1, y_2, \ldots, y_M\}$. Тогда Z принимает значения из Декартова произведения множеств $\mathcal{A}_X$ и $\mathcal{A}_Y$: $\mathcal{A}_Z = \mathcal{A}_X \times \mathcal{A}_Y = \{(x_1, y_1), (x_1, y_2), \ldots, (x_L, y_M)\}$ и имеет функцию вероятности $f(z) = f(x, y)$, которая называется **совместной функцией вероятности** X и Y.

Случайные величины X и Y называются **статистически независимыми**, если для всех x и y

$$f(x, y) = f(x)f(y). \tag{1.1}$$

Условная функция вероятности для $f(y) > 0$ определяется как

$$f(x|y) = \frac{f(x, y)}{f(y)}. \tag{1.2}$$

Пусть случайные величины X и Y связаны некоторой функциональной зависимостью $y = G(x)$. Тогда математическое ожидание $\overline{Y} = \overline{G(X)}$ в дискретном случае определяется как

$$\overline{G(X)} = \sum_{i=1}^{L} f(x_i)G(x_i), \tag{1.3}$$

а в непрерывном — как

$$\overline{G(X)} = \int_{Q} f(x)G(x)dx. \tag{1.4}$$

Если $G(x) = x$, то формулы (1.3) и (1.4) задают математическое ожидание $\overline{X}$ случайной величины X.

Пример 1.1. Пусть X — случайная величина, принимающая значения из множества $\mathcal{A}_X = \{0, 1, 2\}$. Пример функции вероятности $f(x)$ для нее и некоторые другие функции от X приведены в таблице

x	$f(x)$	x^2	e^x	$-\log_2 f(x)$
0	0,25	0	1,00	2
1	0,5	1	2,72	1
2	0,25	4	7,39	2

При этом $\overline{X} = 1$; $\overline{X^2} = 1{,}5$; $\overline{e^X} \approx 3{,}46$; $\overline{f(X)} = 0{,}375$; $\overline{-\log_2 f(X)} = 1{,}5$.

В случае многомерной случайной величины $\underline{X}=\{X_1, X_2, \ldots, X_n\}$ математическое ожидание (1.3) и (1.4) вычисляется с учетом совместной функции или плотности вероятности $f(x_1, x_2, \ldots, x_n)$.

Если дискретные случайные величины X и Y связаны функцией $y = G(x)$, являющейся дифференцируемой и имеющей обратную функцию $x = \varphi(y)$, то их функции вероятности совпадают

$$f(y_i) = P(Y = y_i) = f(x_i), \text{ где } y_i = G(x_i) \text{ для каждого } i. \qquad (1.5)$$

Для непрерывной же случайной величины X с плотностью вероятности $f(x)$ плотность вероятности Y вычисляется по формуле

$$g(y) = f(\varphi(y))|\varphi'(y)|. \qquad (1.6)$$

Напомним, что функция $G(x)$ называется **выпуклой**, если она лежит не ниже любой ее хорды.

Определение 1.1. *Функция $G(x)$ называется выпуклой на интервале $[a, b]$, если для каждых $x_1, x_2 \in [a, b]$ и $0 \le \lambda \le 1$:*

$$G(\lambda x_1 + (1-\lambda)x_2) \ge \lambda G(x_1) + (1-\lambda)G(x_2). \qquad (1.7)$$

Функция $G(x)$ называется **вогнутой**, *если функция $-G(x)$ — выпуклая.*

Примерами вогнутых функций могут служить x^2, $|x|$, e^x, выпуклыми функциями являются $\sqrt{x}$ для $x > 0$ и $\log x$. Проверьте это, взяв вторую производную указанных функций.

Теперь мы докажем неравенство Йенсена, на котором основаны многие результаты в теории информации.

Теорема 1.1. **(Неравенство Йенсена):** *Если $G(x)$ является выпуклой функцией, то*

$$\overline{G(X)} \le G(\overline{X}). \qquad (1.8)$$

Доказательство. Рассмотрим касательную $G(\overline{X}) + G'(\overline{X})(x - \overline{X})$ к функции $G(x)$ в точке $x = \overline{X}$. Вследствие выпуклости $G(x)$ справедливо неравенство

$$G(x) \le G(\overline{X}) + G'(\overline{X})(x - \overline{X}),$$

усредняя которое, получаем

$$\begin{aligned} \overline{G(X)} &\le \overline{G(\overline{X}) + G'(\overline{X})(x - \overline{X})} = \\ &= G(\overline{X}) + G'(\overline{X})(\overline{X} - \overline{X}), \end{aligned}$$

откуда и следует (1.8). Очевидно, что знак равенства в (1.8) достигается при $x = \overline{X}$. □

1.2. Информация и энтропия

В теории информации речь всегда идет о ситуациях, в которых мы наблюдаем события, чтобы из них сделать вывод о других событиях. Например, в системе передачи данных мы наблюдаем принятый сигнал с целью вынести суждение о том, что было передано. Или, наблюдая погоду в Молодечно, пытаемся узнать погоду в Минске.

То есть мы пытаемся количественно измерить, как много говорит нам появление некоторого события B о возможности появления другого события A. Говоря на вероятностном языке, появление события B изменяет вероятность события A с априорной вероятности $P(A)$ до апостериорной вероятности $P(A|B)$. Если эти вероятности существенно различаются, то событие B говорит нам «много» о событии A, если эти вероятности равны, то не говорит ничего. Поэтому в качестве меры различия данных вероятностей может быть принят логарифм их отношения.

Определение 1.2. Количество информации, *которое получают о событии A, наблюдая событие B, обозначается $I(A;B)$ и определяется как*

$$I(A;B) = \log_b \frac{P(A|B)}{P(A)}, \tag{1.9}$$

где предполагается, что $P(A) \neq 0$ и $P(B) \neq 0$.

Если в качестве базиса логарифма применять $b = 2$, то **единицей измерения** информации будет бит; если $b = e$, то — нат. В дальнейшем мы будем, как правило, опускать запись базиса логарифма, поскольку он влияет только на выбор единицы измерения.

Определение 1.2 соответствует нашему интуитивному пониманию того, что из себя должно представлять количество информации. Действительно, $I(A;B) = 0$ означает, что события A и B независимы друг от друга, то есть $P(A|B) = P(A)$, и появление события B ничего не говорит о событии A.

Применяя формулу Байеса, мы видим из (1.9), что $I(A;B)$ симметрично относительно A и B:

$$I(A;B) = \log \frac{P(A|B)}{P(A)} = \log \frac{P(AB)}{P(A)P(B)} = \log \frac{P(B|A)}{P(B)} = I(B;A), \tag{1.10}$$

то есть безразлично, получаем мы информацию о B, наблюдая A, или наоборот. Поэтому $I(A;B)$ называется **взаимной информацией** событий A и B.

Далее очевидно, что

$$I(A;B) \leq -\log P(A), \tag{1.11}$$

где знак равенства достигается, когда $P(A|B) = 1$, а по причине симметрии $I(A;B)$ и тогда, когда $P(B|A) = 1$, то есть когда события A и B эквивалентны. Значение $-\log P(A)$ является количеством информации, которой достаточно для знания того, что событие A произошло. Поэтому для $A = B$ можно записать

$$I(A) = I(A;A) = \log \frac{P(A|A)}{P(A)} = -\log P(A). \tag{1.12}$$

Величина $I(A)$ называется **собственной информацией** события A.

Собственная информация всегда неотрицательна, она равна нулю только в том случае, если $P(A) = 1$, то есть событие A не является случайным. В самом деле, наблюдение такого события не может принести нам ничего нового, априори неизвестного. Чем меньше вероятность наступления события A, тем больше информации можно получить из его наблюдения, что и отражено в (1.12).

Например, если Гидрометеоцентр утверждает, что завтра будет дождь с вероятностью $7/8$, и дождь действительно пошел, то данный факт (реальный дождь) принесет нам $-\log_2(7/8) = 0{,}19$ бит информации, и мы вправе сказать, что мы и так знали, что дождь пойдет. Если же дождь не пошел, то данный факт принесет нам $-\log_2(1/8) = 3$ бита информации и мы почувствуем себя обогащенными знаниями (в частности, о качестве прогнозов Гидрометеоцентра). Если Гидрометеоцентр выдает такой прогноз изо дня в день, то вероятность того, что мы получим 0,19 бит информации, равна $7/8$, а вероятность получения нами 3 битов равна $1/8$. Поэтому в среднем высказывания Гидрометеоцентра приносят нам $-(7/8)\log_2(7/8) - (1/8)\log_2(1/8) = 0{,}54$ бита информации. Таким образом, мы впервые сталкиваемся с понятием **энтропии**.

Аналогично, если мы подбрасываем монету, наблюдаем исход данного эксперимента и нам известно, что вероятности выпадения аверса и реверса равны $1/2$, то наблюдение приносит нам в точности $-\log_2(1/2) = 1$ бит информации. Проводя подобный эксперимент с кубиком, у которого вероятность выпадения каждой грани равна $1/6$, мы получим $-\log_2(1/6) = 2{,}58$ бит информации.

Обобщая примеры с монетой и кубиком, можно заключить, что если все L исходов некоторого случайного эксперимента равновероят-

ны, то наблюдение каждого исхода приносит $-\log_2(1/L) = \log_2 L$ бит информации. Такой подход к информации дает «правильные» ответы на многие технические проблемы. Например, пусть в некоторой деревне имеется 8 телефонов и мы можем каждому из них поставить в соответствие три ($\log_2 8$) двоичных разряда. Тогда набор каждого номера эквивалентен выдаче в линию трех бит информации.

Аналогично в теории алгоритмов сложность алгоритма определяют как функцию (как правило, полиномиальную или экспоненциальную) от количества информации, содержащегося во входных данных задачи: если мы имеем на входе число L, то оно несет в себе $l = \log_2 L$ бит информации, а сложность алгоритма может быть, например, порядка l^2 (полином) или 2^l (экспонента).

Много это или мало — один бит информации? С одной стороны, это «всего лишь» столько информации, сколько мы получаем, подбрасывая монету. С другой стороны, представим себе, что нам необходимо дать уникальный телефонный номер не только каждому дому в упомянутой деревне, но и каждому жителю Земли. Сейчас на Земле живет около $6 \cdot 10^9 \approx 2^{32,5}$ человек, поэтому нам достаточно «всего лишь» 33 бита информации для идентификации любого человека на Земле, причем добавление «всего лишь» одного бита к уникальному номеру позволит «телефонизировать» вдвое большее количество людей!

Впервые ввести меру информации попытался Р. Хартли в 1928 г. В своих рассуждениях он исходил из интуитивной идеи о том, что сообщение, состоящее из n символов, должно нести в n раз больше информации, чем сообщение, состоящее из одного символа.

Пусть алфавит $\mathcal{A}_X$ некоторой случайной величины X состоит из L символов: $\mathcal{A}_X = \{x_1, x_2, \ldots, x_L\}$. Тогда алфавит сообщения, составленного из значений n последовательных случайных величин X, является Декартовым произведением n алфавитов $\mathcal{A}_X$ и обозначается $\mathcal{A}_X^n$. То есть если в множестве $\mathcal{A}_X$ L элементов, то в множестве $\mathcal{A}_X^n$ их L^n. По идее Хартли, в n раз должно различаться количество информации между сообщениями, выбранными из множеств $\mathcal{A}_X$ и $\mathcal{A}_X^n$. Обозначим эти количества информации через $H_0(L)$ и $H_0(L^n)$ соответственно.

Теорема 1.2. *Единственной непрерывной функцией, удовлетворяющей равенству*

$$H_0(L^n) = nH_0(L), \tag{1.13}$$

является логарифм $H_0(L) = K \log L$, где K — некоторая константа.

Доказательство. Обозначим $M = L^n$, тогда $n = \frac{\log M}{\log L}$ и из (1.13) имеем $H_0(M) = \frac{\log M}{\log L} H_0(L)$ или $\frac{H_0(M)}{\log M} = \frac{H_0(L)}{\log L} = K$, то есть $H_0(L) = K \log L$, где K — некоторая константа, влияющая только на базис логарифма. □

В дальнейшем, если нет опасности перепутать, мы будем для простоты вместо $H_0(L)$ записывать H_0. Мера информации H_0 называется в честь ее открывателя **информацией Хартли**. Теорема 1.2 дает еще одно объяснение появлению логарифма в определении 1.2.

Отметим, однако, что рассуждения Хартли справедливы исключительно в случае сообщений, символы которых равновероятны и независимы друг от друга. То есть информация Хартли, получаемая при бросании кубика, равна $H_0(6) = \log_2 6$ бит. Если же мы будем бросать кубик со смещенным центром тяжести, то, например, одна из граней будет выпадать чаще других, а так как мы это знаем, то и наблюдение такого эксперимента принесет нам, по нашим интуитивным ощущениям, в среднем меньше информации, чем при бросании обычного кубика.

Поэтому в 1948 г. Шеннон обобщил рассуждения Хартли на случай сообщений, имеющих неодинаковые вероятности отдельных символов, а также для сообщений, символы которых зависимы друг от друга. Оба этих явления имеют место, например, в любом естественном языке. Рассмотрим теперь идеи Шеннона более подробно.

Пусть случайные величины X и Y принимают значения из множеств $\mathcal{A}_X = \{x_1, x_2, \ldots, x_L\}$ и $\mathcal{A}_Y = \{y_1, y_2, \ldots, y_M\}$ соответственно. Применяя (1.12) и (1.9) к событиям $X = x_i$ и $Y = y_j$, получаем

$$I(X = x_i) = -\log f(x_i) \tag{1.14}$$

и

$$I(X = x_i;\ Y = y_j) = +\log \frac{f(x_i|y_j)}{f(x_i)}. \tag{1.15}$$

Мы видим, что информация (1.14) является функцией случайной величины X, а информация (1.15) — случайной величины $\{X, Y\}$, поэтому мы можем определить их среднее значение.

Определение 1.3. *Величина*

$$H(X) = \overline{I(X = x)} = \overline{-\log f(X)} = -\sum_{i=1}^{L} f(x_i) \log f(x_i) \tag{1.16}$$

называется **энтропией** *случайной величины* X*, а величина*

$$I(X;Y) = \overline{I(X=x;\ Y=y)} = \overline{\log \frac{f(X|Y)}{f(X)}} = \\ = \sum_{i=1}^{L}\sum_{j=1}^{M} f(x_i,y_j)\log\frac{f(x_i|y_j)}{f(x_i)} \qquad (1.17)$$

называется **взаимной информацией** *случайных величин* X *и* Y*.*

Из того, что по правилу Лопиталя $0\log 0 = \lim\limits_{\varepsilon\to 0}\varepsilon\log\varepsilon = 0$, следует, что величины $H(X)$ и $I(X;Y)$ определены и тогда, когда $f(x_i)=0$ для некоторого x_i или $f(x_i,y_j)=0$ для некоторой пары (x_i,y_j).

В отличие от информации, как функции двух переменных, энтропия определена как функция одной переменной. Определение (1.16) позволяет рассматривать ее как количество информации, необходимое в среднем для описания X.

Если X является случайной величиной, принимающей всего два значения x_1 и x_2, причем $f(x_1)=p$, а $f(x_2)=1-p$, то мы получим

$$H(X) = \eta(p) = -p\log_2 p - (1-p)\log_2(1-p). \qquad (1.18)$$

Данная функция называется **двоичной функцией энтропии** и ее график представлен на рис. 1.1.

С этой функцией мы уже сталкивались при рассмотрении прогнозов Гидрометеоцентра (на с. 18). Возможно и следующее шуточное толкование функции (1.18). Пусть для некоторого молодого человека $p=1$ означает, что девушка его любит, а $p=0$ — что не любит. Это два совершенно определенных состояния девушки, энтропия которых равна нулю. Встреча молодых людей начинается с состояния $p=0{,}5$, соответствующего безразличию и максимальной неопределен-

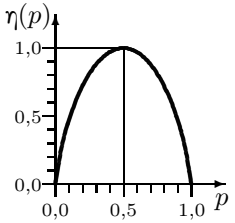


Рис. 1.1. Двоичная энтропия

ности (энтропии), а процесс ухаживания уменьшает эту неопределенность в ту или иную сторону.

Теорема 1.3. *Пусть X принимает значения из алфавита $\mathcal{A}_X = \{x_1, x_2, \ldots, x_L\}$. Тогда*

$$0 \le H(X) \le H_0, \tag{1.19}$$

причем равенство слева выполняется в точности тогда, когда существует такое i, что $f(x_i) = 1$, а равенство справа — когда $f(x_i) = 1/L$ для всех $i = 1, 2, \ldots, L$.

Доказательство.

$$-f(x_i)\log f(x_i) = f(x_i)\log\frac{1}{f(x_i)} = \begin{cases} = 0 & \text{для} \quad f(x_i) = 0, \\ > 0 & \text{для} \quad 0 < f(x_i) < 1, \\ = 0 & \text{для} \quad f(x_i) = 1. \end{cases}$$

То есть $H(X) = -\sum\limits_{i=1}^{L} f(x_i)\log f(x_i) \ge 0$. А равенство имеет место в точности тогда, когда $f(x_i) = 0$ для всех i, кроме как максимум одного, для которого $f(x_i) = 1$.

Для доказательства неравенства справа запишем

$$\begin{aligned} H(X) &= \overline{-\log f(X)} = \overline{\log\frac{1}{f(X)}} \le \\ &\le \log\overline{\frac{1}{f(X)}} = \log\sum_{i=1}^{L} f(x_i)\frac{1}{f(x_i)} = \log L, \end{aligned}$$

где неравенство следует из неравенства Йенсена (1.8), а во второй строке использовано (1.3). То есть

$$H(X) \le \log L = H_0, \tag{1.20}$$

а знак равенства в (1.20) достигается тогда же, когда и в неравенстве Йенсена — при $\frac{1}{f(x_i)} = \overline{\frac{1}{f(X)}} = L$, то есть при $f(x_i) = 1/L$ для любого i. □

Таким образом, если случайная величина X принимает единственное значение, то ее наблюдение не принесет информации и энтропия $H(X)$ равна нулю. Если же X принимает большое количество равновероятных значений, то энтропия максимальна (и равна информации Хартли), а исход случайного эксперимента максимально неопределен,

причем чем больше имеется равновероятных возможностей, тем больше и максимальная энтропия. Данное свойство позволяет толковать энтропию как **меру неопределенности** случайной величины X. Это подтверждает и пример с бросанием обычного кубика и кубика со смещенным центром тяжести.

Определение 1.4. Условная энтропия X относительно события $Y = y_j$ *определяется как*

$$H(X|Y = y_j) = \overline{-\log f(X|y_j)} = -\sum_{i=1}^{L} f(x_i|y_j) \log f(x_i|y_j), \qquad (1.21)$$

условная энтропия X относительно случайной величины Y — *как*

$$H(X|Y) = \overline{-\log f(X|Y)} = -\sum_{i=1}^{L}\sum_{j=1}^{M} f(x_i, y_j) \log f(x_i|y_j), \qquad (1.22)$$

а **совместная энтропия** X *и* Y — *как*

$$H(X, Y) = \overline{-\log f(X, Y)} = -\sum_{i=1}^{L}\sum_{j=1}^{M} f(x_i, y_j) \log f(x_i, y_j). \qquad (1.23)$$

Заметим, что в равенстве (1.21) усреднение берется только по одной переменной X, тогда как в равенствах (1.22) и (1.23) мы усредняем по обеим переменным X и Y, используя при этом их совместную функцию вероятности $f(x_i, y_j)$.

Условная энтропия относительно события (1.21) характеризует неопределенность случайной величины X после того, как случайная величина Y приняла некоторое конкретное значение, а условная энтропия (1.22) — после того, как случайная величина Y приняла произвольное значение.

Следствие 1.1. *Если $f(y_j) \neq 0$, то есть $f(x_i|y_j)$ существует, то*

$$0 \leq H(X|Y = y_j) \leq H_0, \qquad (1.24)$$

причем равенство слева выполняется в точности тогда, когда существует такое i, что $f(x_i|y_j) = 1$, а равенство справа — когда $f(x_i|y_j) = 1/L$ для всех $i = 1, 2, \ldots, L$.

Доказательство. Аналогично доказательству теоремы 1.3. □

Преобразуя (1.22), получаем

$$\begin{aligned} H(X|Y) &= -\sum_{i=1}^{L}\sum_{j=1}^{M} f(y_j)f(x_i|y_j)\log f(x_i|y_j) = \\ &= -\sum_{j=1}^{M} f(y_j)\sum_{i=1}^{L} f(x_i|y_j)\log f(x_i|y_j) = \\ &= \sum_{j=1}^{M} f(y_j)H(X|Y=y_j). \end{aligned} \tag{1.25}$$

Применение равенства (1.25) часто является наиболее простым способом вычисления $H(X|Y)$. Из (1.24) и (1.25) мы получаем

Следствие 1.2.

$$0 \le H(X|Y) \le H_0, \tag{1.26}$$

причем равенство слева выполняется в точности тогда, когда для каждого j с $f(y_j) \neq 0$ существует такое i, что $f(x_i|y_j) = 1$, а равенство справа — когда для каждого j с $f(y_j) \neq 0$ справедливо $f(x_i|y_j) = 1/L$ для всех $i = 1, 2, \ldots, L$.

Другими словами, равенство нулю условной энтропии в левой части (1.26) достигается, если (и только если) случайная величина Y однозначно определяет случайную величину X, то есть между ними нет статистической связи.

Теорема 1.4. *Условная энтропия не может превосходить безусловную:*

$$H(X|Y) \le H(X), \tag{1.27}$$

где знак равенства выполняется в точности тогда, когда случайные величины X и Y статистически независимы.

Доказательство. Рассмотрим величину

$$\begin{aligned} H(X|Y) - H(X) &= \overline{-\log f(X|Y)} - \overline{-\log f(X)} = \\ &= \overline{\log \frac{f(Y)}{f(X,Y)}} + \overline{\log f(X)} = \\ &= \overline{\log \frac{f(X)f(Y)}{f(X,Y)}} \le \log \overline{\frac{f(X)f(Y)}{f(X,Y)}} = \\ &= \log \sum_{i=1}^{L}\sum_{j=1}^{M} f(x_i,y_j)\frac{f(x_i)f(y_j)}{f(x_i,y_j)} = \log 1 = 0, \end{aligned}$$

где неравенство следует из неравенства Йенсена, причем равенство выполняется в точности тогда, когда $\frac{f(x_i)f(y_j)}{f(x_i,y_j)} = \overline{\left(\frac{f(X)f(Y)}{f(X,Y)}\right)} = 1$, то есть когда $f(x_i, y_j) = f(x_i)f(y_j)$ для всех x_i и y_j, что представляет собой определение статистической независимости случайных величин X и Y (1.1). □

Значит, добавление условий, связанных с X, энтропию X уменьшает, а добавление условий, не связанных с X, энтропию X не изменяет.

Теорема 1.5. *Взаимная информация $I(X;Y)$ никогда не отрицательна и не превосходит ни одну из энтропий $H(X)$ и $H(Y)$:*

$$0 \le I(X;Y) \le \min\{H(X), H(Y)\}, \tag{1.28}$$

причем знак равенства слева достигается в точности тогда, когда случайные величины X и Y статистически независимы, а справа — когда одна из случайных величин однозначно определяет другую.

Доказательство.

$$\begin{aligned} I(X;Y) &= \overline{\log \frac{f(X|Y)}{f(X)}} = \overline{\log f(X|Y)} - \overline{\log f(X)} = \\ &= H(X) - H(X|Y) \ge 0, \end{aligned} \tag{1.29}$$

где неравенство и условия достижения равенства в нем вытекают из теоремы 1.4. Поэтому левая часть в (1.28) доказана.

Так как по (1.26) $H(X|Y) \ge 0$, то из (1.29) следует, что $I(X;Y) \le$ $\le H(X)$. Рассуждая аналогично и учитывая симметрию взаимной информации, можно записать

$$0 \le H(Y|X) = H(Y) - I(X;Y). \tag{1.30}$$

То есть $I(X;Y) \le H(Y)$. Если X и Y определяют друг друга взаимно однозначно, то $H(X|Y) = H(Y|X) = 0$ и $I(X;Y) = H(X) = H(Y)$, что доказывает правую часть (1.28). □

Заметим, что выражение (1.30) позволяет интерпретировать получение информации $I(X;Y)$ как уменьшение неопределенности $H(Y)$ до величины $H(Y|X)$.

Пример 1.2. Пусть мы имеем аптекарские весы и $n = 3^k$ одинаковых на вид монет, причем известно, что среди них имеется одна фальшивая, которая легче всех остальных. Разбивая все монеты на три равные части и взвешивая две из них, мы можем определить, в какой части находится фальшивая монета: если легче одна из чаш весов, то фальшивая монета находится там, а если веса одинаковы, то — в оставшейся части. Действуя таким образом с локализованной частью и далее, мы найдем фальшивую монету за k взвешиваний.

Однако обоснование того, что применять надо именно такую стратегию взвешивания и что k является минимальным необходимым числом взвешиваний, следует из теории информации. Действительно, имея $n = 3^k$ монет, мы стоим перед неопределенностью $H(X) = \log_2 n$ бит, поскольку каждая из монет может быть либо нормальной, либо фальшивой. Наша задача состоит в том, чтобы за минимальное число взвешиваний $Y_1, Y_2, \ldots$ снизить эту неопределенность до нуля, когда о каждой монете достоверно известно, фальшивая она или нет.

На каждую из чаш аптекарских весов имеет смысл класть только одинаковое количество монет, поэтому разделить монеты на части можно, только оставив i монет и положив на каждую чашу по $(n-i)/2$ монет при условии, что число $(n-i)/2$ является целым и $i = 0, 1, \ldots, n$. Тогда с вероятностью i/n фальшивая монета находится в оставшейся части, и мы после взвешивания остаемся перед неопределенностью в $\log_2 i$ бит, а с вероятностью $(n-i)/n$ фальшивая монета находится в числе взвешиваемых монет, причем после взвешивания мы останемся перед неопределенностью в $\log_2 \frac{n-i}{2}$ бит. Поэтому остающаяся после первого взвешивания неопределенность равна

$$H(X|Y_1) = \left(\frac{i}{n}\right) \log_2 i + \left(\frac{n-i}{n}\right) \log_2 \left(\frac{n-i}{2}\right) \quad \text{бит},$$

причем при $i = n/3$ она минимальна и равна $\log_2(n/3)$ бит. То есть максимальная информация, получаемая после первого взвешивания, равна $I(X;Y_1) = H(X) - H(X|Y_1) = \log_2 n - \log_2(n/3) = \log_2 3$ бит, и достигается это при делении всех монет перед взвешиванием на три равные части ($i = n/3$). В результате следования и далее стратегии получения максимальной информации (или минимальной остающейся неопределенности) после j-го взвешивания у нас останется $n/3^j$ монет и неопределенность в $\log_2(n/3^j)$ бит, а после k-го взвешивания — одна фальшивая монета и соответственно нулевая неопределенность.

Пример 1.3. Пусть двоичные случайные величины X и Y имеют совместную функцию вероятности $f(x,y)$, задаваемую следующей таблицей:

	x_1	x_2
y_1	0	3/4
y_2	1/8	1/8

Тогда

$$H(X) = \eta(1/8) = 0{,}54 \text{ и } H(Y) = \eta(1/4) = 0{,}81.$$

Для получения условных энтропий по (1.25) вычислим условные энтропии относительно соответствующих событий

$$H(X|Y=y_1) = \eta(0) = 0, \qquad H(X|Y=y_2) = \eta(1/2) = 1{,}00,$$
$$H(Y|X=x_1) = \eta(0) = 0, \qquad H(Y|X=x_2) = \eta(1/7) = 0{,}59.$$

Поэтому

$$\begin{aligned} H(X|Y) &= \frac{3}{4}H(X|Y=y_1) + \frac{1}{4}H(X|Y=y_2) = 0{,}25, \\ H(Y|X) &= \frac{1}{8}H(Y|X=x_1) + \frac{7}{8}H(Y|X=x_2) = 0{,}52. \end{aligned}$$

Вычисляя совместную энтропию и взаимную информацию двумя способами, убеждаемся в их симметричности относительно аргументов:

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) = 0{,}54 + 0{,}52 = 1{,}06, \\ &= H(Y) + H(X|Y) = 0{,}81 + 0{,}25 = 1{,}06, \\ I(X;Y) &= H(X) - H(X|Y) = 0{,}54 - 0{,}25 = 0{,}29, \\ &= H(Y) - H(Y|X) = 0{,}81 - 0{,}52 = 0{,}29. \end{aligned}$$

Другими словами, наблюдение случайной величины Y принесет нам 0,29 бит информации о случайной величине X и наоборот.

Из определения энтропии (1.16) и определения условной функции вероятности (1.2) следует, что взаимная информация может быть вычислена и как

$$\begin{aligned} I(X;Y) &= \overline{\log \frac{f(X|Y)}{f(X)}} = \overline{\log \frac{f(X,Y)}{f(X)f(Y)}} = \\ &= \overline{\log f(X,Y) - \log f(X) - \log f(Y)} = \\ &= \overline{-\log f(X)} + \overline{-\log f(Y)} + \overline{\log f(X,Y)} = \\ &= H(X) + H(Y) - H(X,Y), \end{aligned} \tag{1.31}$$

а условная энтропия $H(X|Y)$ — как

$$H(X|Y) = \overline{-\log f(X|Y)} = \overline{-\log \frac{f(X,Y)}{f(Y)}} = H(X,Y) - H(Y). \quad (1.32)$$

Иллюстрация к соотношениям (1.29), (1.30), (1.31) и (1.32) между величинами $H(X)$, $H(Y)$, $H(X,Y)$, $H(X|Y)$, $H(Y|X)$ и $I(X;Y)$ дана на рис. 1.2 в виде диаграммы Эйлера.

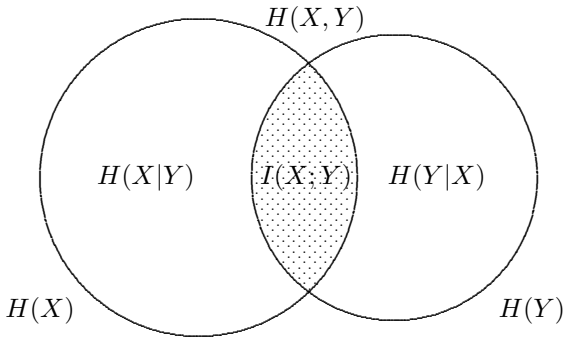


Рис. 1.2. Диаграмма Эйлера

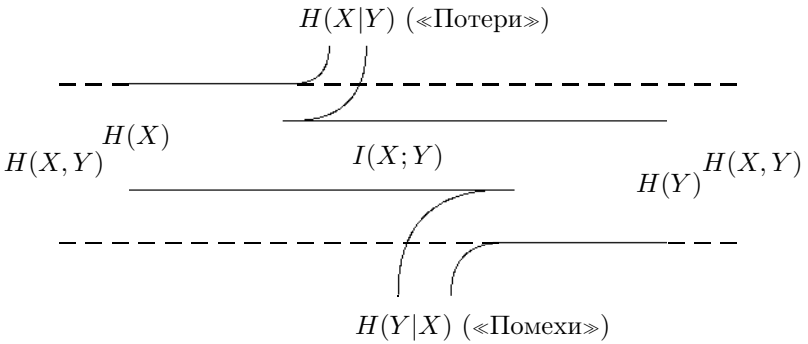


Рис. 1.3. Передача информации с точки зрения энтропии

На рис. 1.3 дана интерпретация процесса передачи информации в терминах энтропии и взаимной информации. Передавая сигнал X через канал связи, мы наблюдаем на его выходе сигнал Y, по которому хотим определить преданный сигнал X. При этом из-за шумов и помех

безвозвратно теряется в среднем $H(X|Y)$ бит информации о переданном сигнале и добавляется в среднем $H(Y|X)$ бит «лишней» информации. Поэтому задачей кодирования является такое (качественное и количественное) введение избыточности в передаваемый сигнал, которое позволяет скомпенсировать как потери исходной, так и появление «лишней» информации.

1.3. Семантическая и прагматическая информация

В то время как синтаксическую информацию можно описать и измерить количественно, семантическая и прагматическая информации поддаются только качественному описанию. Отвлекаясь от нашего чисто математического изложения, приведем здесь некоторые философские размышления о семантической и прагматической частях информации.

Семантическая часть связана со смыслом, содержащимся в сообщении. Смысл же сообщения, в свою очередь, связан со способностью получателя (человека, животного, искусственной самообучающейся системы) это сообщение понять. Причем понимание зависит от совокупности сведений и связей между ними, которыми получатель располагает на момент получения сообщения. Данная совокупность сведений, по предложению Ю. Шрейдера, разработавшего наиболее адекватную теорию семантической информации, называется **тезаурусом** получателя.

В зависимости от соотношения между смысловым содержанием информации и тезаурусом получателя T изменяется и количество семантической информации $I_{semant}(T)$, воспринимаемой получателем и включаемой им в дальнейшем в свой тезаурус. Характер такой зависимости показан на рис. 1.4. Если тезаурус получателя в какой-либо области знаний меньше некоторого порогового значения T_1, то сообщения из этой области он будет просто не в состоянии понять и семантическая информация равна нулю. С другой стороны, если получатель «все знает» ($T > T_2$ на рис. 1.4), то и поступление сообщения не принесет ему дополнительной информации.

Например, получая информацию по телевидению из выпуска новостей, мы сравниваем ее с уже имевшейся у нас и выносим суждение об информативности данного выпуска. При этом количество информации в повторно просмотренном выпуске новостей равно нулю, если, конечно, отвлечься от таких возможных событий, как просмотр вы-

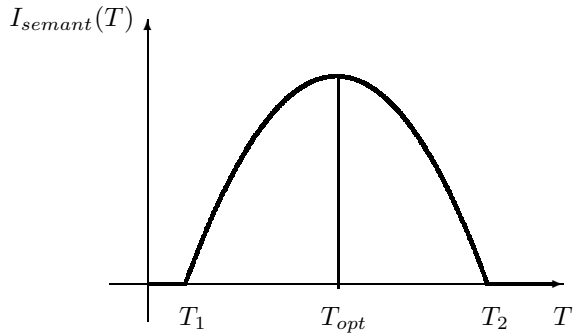


Рис. 1.4. Семантическая информация

пуска новостей в первый раз по черно-белому, а во второй раз — по цветному телевизору.

Наибольшее количество семантической информации достигается, когда тезаурус получателя таков ($T = T_{opt}$), что сведения, содержащиеся в сообщении, получателю понятны и ранее не были ему известны (отсутствовали в его тезаурусе). Следовательно, одно и то же сообщение принесет разное количество информации для получателей с разным тезаурусом.

В предположении, что имеется количественное описание для величины конкретного тезауруса (например, содержащееся в нем количество элементарных фактов и их связей), в качестве количества семантической информации может быть принято изменение величины тезауруса получателя после восприятия и анализа поступившего сообщения.

Нельзя, однако, забывать, что даже получателям с одинаковым тезаурусом, но имеющим различную способность, возможность и желание к анализу, одно и то же сообщение может принести различное количество семантической информации. То есть семантическая информация относительна не только применительно к тезаурусу, но и применительно к конкретному носителю данного тезауруса в некоторый момент времени.

Рассматривая информацию в ее семанто-прагматическом аспекте, Карл фон Вайцзеккер сформулировал тезис, что «информацией является только то, может быть понято, и только то, что создает информацию». Поясним это на следующих примерах.

Пусть студент А просит студента Б выключить свет. Тогда о передаче информации можно говорить, если А понял, что от него хочет Б (семантика), и за этим следует некоторая реакция Б (прагматика). Эта реакция может состоять и в том, что Б на самом деле выключит свет. Но даже если этого не произойдет, о передаче информации можно говорить все равно, если тем или иным способом возможно установить действие (эффект) информации на Б (например, замерив нейронную активность в его центре слуха).

Или, когда ребенок учится говорить, он издает сначала произвольные звуки, а через некоторое время замечает, что, например, за двойным слогом «ма» следует определенная реакция матери. То есть, за наблюдением эффекта, производимого словом «мама» (прагматика), следует понимание его семантики. Когда же ребенок чувствует себя уверенно на уровне звуков и слогов, он по той же схеме изучает следующие семантические уровни языка — слова и предложения.

Поэтому семантический аспект информации объективно подтверждается прагматическим, а эффект, производимый информацией, снова представляется в виде информации, что очень сходно с герменевтическим кругом: когда понимание целого складывается из понимания отдельных его частей, а для понимания частей необходимо предварительное понимание целого. То есть акт передачи информации является актом информационного обмена, который воздействует как на получателя, так и на источник.

С синтаксической точки зрения количество информации, содержащееся в событиях, имеющих вероятности $2^{-10^{100}}$ и $2^{-10^{101}}$, отличается в 10 раз, а с прагматической точки зрения мы не можем получить информацию из такого рода событий — ее невозможно подтвердить.

Поэтому при близких к нулю вероятностях событий равна нулю и прагматическая информация о них. Как и их синтаксическая информация, она равна нулю также и для заведомо известных событий.

То есть качественно изобразить прагматическую информацию в зависимости от вероятности p наступления какого-либо события можно, как это показано на рис. 1.5, где изображена также синтаксическая информация $-\log p$ (1.12).

Элемент подтверждения очень важен в прагматическом аспекте информации. В примере с изучением языка ребенком только на основе уже подтвержденных знаний на более низком семантическом уровне происходит переход на более высокий уровень при встрече новых, неожиданных элементов.

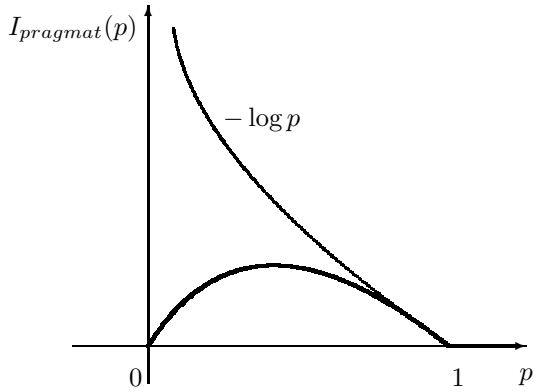


Рис. 1.5. Прагматическая информация

Все аспекты информации можно наблюдать в генетике: с одной стороны ген представляет собой последовательность нуклеотидов с определенным синтаксисом, с другой — эта последовательность перенимает определенные функции в клетке (семантика) и оказывает определенные действия (прагматика), которые существенны для выживания клетки.

Размышления об информации находят отражение не только в научной, но и в художественной литературе. Вот что пишет о ее семантопрагматических аспектах современный испанский писатель Артуро Перес-Реверте[1]:

≪— ...Читатель формируется из того, что он прочел раньше, но также из кино и телепередач, которые он посмотрел. К той информации, которую предлагает ему автор, он непременно добавляет свою собственную. Тут и кроется опасность: из-за избытка аллюзий может получиться неверный или даже вовсе не соответствующий действительности образ...

— Значит, информация была ложной.

— Не обязательно. Информация, которую дает вам книга, обычно бывает объективной. Хотя злонамеренный автор может представить ее в таком виде, что читатель поймет ее превратно, но сама по себе информация никогда не бывает ложной. Это сам читатель прочитывает книгу неверно≫.

[1] *Перес-Реверте А.* Клуб Дюма, или Тень Ришелье/пер. с исп. М.: Иностранка, 2003. С. 518.

1.4. Дифференциальная и относительная энтропия

В этом разделе мы введем понятие дифференциальной энтропии, которая является энтропией для непрерывных случайных величин, и относительной энтропии, которая может служить мерой различия между функциями или плотностями вероятности.

Определение 1.5. Дифференциальной энтропией $h(X)$ *непрерывной случайной величины* X *с плотностью вероятности* $f(x)$ *называется*

$$h(X) = \overline{-\log f(X)} = -\int\limits_{Q} f(x)\log f(x)dx, \tag{1.33}$$

если интеграл существует, и где Q *— ненулевое множество* X*.*

Так же как и для дискретных случайных величин (1.16), дифференциальная энтропия зависит только от плотности вероятности $f(x)$, поэтому мы будем в некоторых случаях трактовать ее как функционал $h(f(x))$, для простоты записывая $h(f)$.

Пример 1.4. Рассмотрим случайную величину X, равномерно распределенную на интервале от a до b. В данном случае $Q = [a, b]$, а $f(x) = 1/(b-a)$. Тогда

$$h(X) = -\int\limits_{a}^{b} \frac{1}{b-a}\,\log\frac{1}{b-a}\,dx = \log(b-a). \tag{1.34}$$

Заметим, что $\log(b-a) < 0$ при $(b-a) < 1$, и поэтому дифференциальная энтропия может быть отрицательной.

Пример 1.5. Рассмотрим случайную величину X с нормальным (**гауссовским**) распределением ($Q = (-\infty, \infty)$), математическим ожиданием μ и дисперсией σ^2:

$$f(x) = -\frac{1}{\sqrt{2\pi}\sigma}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{1.35}$$

Ее дифференциальная энтропия равна

$$\begin{aligned} h(X) &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)dx = \\ &= \log\sqrt{2\pi\sigma^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx + \\ &\quad + \frac{\log e}{2\sigma^2}\int_{-\infty}^{\infty}(x-\mu)^2\frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx = \end{aligned}$$

$$= \log\sqrt{2\pi\sigma^2} + \frac{\log e}{2\sigma^2}\,\sigma^2 = \log\sqrt{2\pi e\sigma^2}, \qquad (1.36)$$

где использовано соотношение

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

В зависимости от σ дифференциальная энтропия здесь также может быть положительной, нулевой или отрицательной.

В отличие от энтропии дискретной случайной величины, дифференциальная энтропия не может быть истолкована как неопределенность X в абсолютном смысле, поскольку она зависит от преобразования координат, в которых находится (возможно, многомерная — см. задачу 1.6.2) случайная величина.

Действительно, пусть случайные величины X и Y связаны дифференцируемой функцией $y = G(x)$, имеющей обратную функцию $x = \varphi(y)$. Если X и Y — дискретные случайные величины, то их функции вероятности (1.5), а следовательно, и энтропии (1.16) совпадают: $H(Y) = H(X)$ (для произвольной функции $y = G(x)$ выполняется $H(Y) \le H(X)$ — см. задачу 1.2.7). Если же X — непрерывная случайная величина с плотностью вероятности $f(x)$, то плотность вероятности Y равна $g(y) = f(\varphi(y))|\varphi'(y)|$ (1.6) и дифференциальная энтропия (1.33) будет зависеть от величины $|\varphi'(y)|$.

Пример 1.6. Пусть X — непрерывная случайная величина с дифференциальной энтропией $h(X)$, а $y = ax+b$, где a и b — некоторые константы. Тогда $\varphi(y) = \frac{y-b}{a}$, $|\varphi'(y)| = \frac{1}{|a|}$ и $g(y) = f\left(\frac{y-b}{a}\right)\frac{1}{|a|}$. Поэтому

$$\begin{aligned} h(Y) &= -\int_Q g(y)\log g(y)dy = \\ &= -\int_Q f\left(\frac{y-b}{a}\right)\frac{1}{|a|}\log\left[f\left(\frac{y-b}{a}\right)\frac{1}{|a|}\right]dy = \\ &= -\int_Q f\left(\frac{y-b}{a}\right)\frac{1}{|a|}\log\left[f\left(\frac{y-b}{a}\right)\right]dy - \\ &\quad -\int_Q f\left(\frac{y-b}{a}\right)\frac{1}{|a|}\log\frac{1}{|a|}dy = h(X) + \log|a|, \end{aligned}$$

где последнее равенство получается после замены переменных.

Несмотря на данное различие, целый ряд свойств дифференциальной энтропии совпадает по форме со свойствами энтропии дискретных случайных величин.

Определение 1.6. *Если X и Y имеют совместную плотность вероятности $f(x,y) > 0$ на множестве Q, то* **условная дифференциальная энтропия** $h(X|Y)$ *определяется как*

$$h(X|Y) = -\int\limits_Q f(x,y)\log f(x|y)\,dx\,dy. \tag{1.37}$$

Так как

$$f(x|y) = \frac{f(x,y)}{f(y)}, \tag{1.38}$$

где $f(y)$ рассматривается на своем ненулевом множестве, то непосредственно из (1.37) следует аналогичное (1.30) свойство

$$h(X|Y) = h(X,Y) - h(Y).$$

Определение 1.7. *Если X и Y имеют совместную плотность вероятности $f(x,y) > 0$ на множестве Q, то* **взаимная информация** $I(X;Y)$ *определяется как*

$$I(X;Y) = \int\limits_Q f(x,y)\log\frac{f(x|y)}{f(x)}\,dx\,dy. \tag{1.39}$$

Из (1.37) и (1.39) следует аналогичное (1.29) равенство

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X). \tag{1.40}$$

Теорема 1.6. *Дифференциальная энтропия $h(f)$ является выпуклым функционалом плотности f.*

Доказательство. Пусть случайные величины X_1 и X_2 с плотностями вероятности f_1 и f_2 определены на множестве Q. Пусть $\vartheta = 1$ с вероятностью λ, а $\vartheta = 2$ с вероятностью $1-\lambda$. Пусть далее $Z = X_\vartheta$. Тогда случайная величина Z имеет плотность вероятности $\lambda f_1 + (1-\lambda)f_2$. Так как добавление условий не увеличивает энтропию (1.27), мы имеем $h(Z) \geq h(Z|\vartheta)$, или, в эквивалентной записи,

$$h(\lambda f_1 + (1-\lambda)f_2) \geq \lambda h(f_1) + (1-\lambda)h(f_2), \tag{1.41}$$

откуда, учитывая определение выпуклости (1.7), получаем утверждение теоремы. Аналогичное утверждение верно и в дискретном случае (см., например, рис. 1.1). □

Одним из физических следствий **выпуклости энтропии** (1.41) является то, что смешивание двух газов с одинаковыми энтропиями приводит к газу с более высокой энтропией.

Определение 1.8. *Пусть $f(x)$ и $g(x)$ — две функции или плотности вероятности. Обозначим $G(x) = \frac{f(x)}{g(x)}$. Тогда* **относительная энтропия** *для дискретных случайных величин определяется как*

$$H(f||g) = \overline{\log G(X)} = \sum_{i=1}^{L} f(x_i) \log \frac{f(x_i)}{g(x_i)}, \tag{1.42}$$

а для непрерывных случайных величин — как

$$h(f||g) = \overline{\log G(X)} = \int\limits_{Q} f(x) \log \frac{f(x)}{g(x)}\, dx. \tag{1.43}$$

Заметим, что $h(f||g)$ (1.43) конечна, если множество, на котором $f(x) > 0$, содержится в множестве, на котором $g(x) > 0$. Поэтому под множеством Q в (1.43) понимается множество, на котором $f(x) > 0$. Заметим также, что относительная энтропия не является симметричной функцией своих аргументов, то есть в общем случае $h(f||g) \neq h(g||f)$ (см. задачу 1.4.5).

Теорема 1.7. *Относительная энтропия неотрицательна*

$$h(f||g) \geq 0, \tag{1.44}$$

причем равенство достигается тогда, и только тогда, когда $f(x) = g(x)$.

Доказательство. Ограничимся рассмотрением только непрерывного случая, доказательство для дискретных случайных величин проводится аналогично:

$$\begin{aligned} -h(f||g) &= \overline{\log \frac{1}{G(X)}} \leq \log \overline{\frac{1}{G(X)}} = \log \int\limits_{Q} f(x) \frac{g(x)}{f(x)}\, dx = \\ &= \log \int\limits_{Q} g(x)\, dx = \log 1 = 0, \end{aligned}$$

где первое неравенство следует из неравенства Йенсена, а равенство достигается тогда же, когда и в неравенстве Йенсена, то есть при

$$\frac{f(x)}{g(x)} = \overline{\left(\frac{f(X)}{g(X)}\right)} = 1.$$

□

Итак, мы видим, что относительная энтропия всегда неотрицательна и равна нулю тогда, и только тогда, когда $f(x) = g(x)$, то есть она представляет собой как бы расстояние между $f(x)$ и $g(x)$. Однако она не является «настоящим» расстоянием между ними, так как она не симметрична относительно аргументов и не удовлетворяет неравенству треугольника. Но ее применение в качестве меры различия двух плотностей вероятности случайных величин часто оправдано.

Заметим, что из (1.38) и (1.39) следует, что взаимная информация может быть представлена через относительную энтропию как

$$I(X;Y) = h(f(x,y)||f(x)f(y)). \tag{1.45}$$

Поэтому из неотрицательности относительной энтропии (1.44) следует альтернативное доказательство того факта, что $I(X;Y) \geq 0$ (см. (1.28)), с равенством в точности тогда, когда X и Y статистически независимы. В свою очередь, из неотрицательности взаимной информации и (1.40) следует, что добавление условий не увеличивает дифференциальную энтропию: $h(X) \geq h(X|Y)$.

1.5. Максимум энтропии

Рассмотрим проблему **максимизации энтропии**

$$h(f) = -\int\limits_Q f(x) \log f(x) dx \tag{1.46}$$

на всех плотностях вероятности f, удовлетворяющих следующим условиям:

$$f(x) \geq 0, \tag{1.47}$$

$$\int\limits_Q f(x) dx = 1, \tag{1.48}$$

$$\int\limits_Q f(x) r_i(x) dx = \alpha_i, \text{ при } i = 1, 2, \ldots, \tag{1.49}$$

где $r_i(x)$ — некоторые функции от x. Если, например, $r_i(x) = x^i$, то условие (1.49) задает i-е моменты случайной величины X.

Эта проблема соответствует стандартной изопериметрической задаче вариационного исчисления, поэтому для ее решения применим метод множителей Лагранжа $(\lambda_0, \lambda_1, \ldots)$ и сформируем функционал из подынтегральных выражений в (1.46), (1.48) и (1.49)

$$\Phi(f) = -f(x) \ln f(x) + \lambda_0 f(x) + f(x) \sum_{i>0} \lambda_i r_i(x).$$

Необходимым условием экстремума определенного интеграла (1.46) является удовлетворение функции $f(x)$ уравнению Эйлера

$$\frac{\partial \Phi(f)}{\partial f} = -\ln f(x) - 1 + \lambda_0 + \sum_{i>0} \lambda_i r_i(x) = 0, \qquad (1.50)$$

где при взятии частной производной учтено, что $(y \ln y)' = \ln y + 1$.

Так как дифференциальная энтропия $h(f)$ является выпуклой функцией (теорема 1.6), из (1.50) получаем форму плотности вероятности, необходимую для максимизации энтропии

$$f^*(x) = e^{\lambda_0 - 1 + \sum\limits_{i>0} \lambda_i r_i(x)}, \; x \in Q, \qquad (1.51)$$

где параметры λ_i выбираются так, что $f^*(x)$ удовлетворяет условиям (1.47), (1.48) и (1.49). Докажем теперь, что плотность $f^*(x)$ (1.51) на самом деле максимизирует интеграл (1.46).

Теорема 1.8. *Функция $f^*(x)$ из (1.51) (и только она) максимизирует дифференциальную энтропию $h(f)$ (1.46) на всех плотностях вероятности f, удовлетворяющих условиям (1.47), (1.48) и (1.49).*

Доказательство. Пусть некоторая другая плотность $g \neq f^*$ также удовлетворяет условиям (1.47), (1.48) и (1.49). Тогда

$$\begin{aligned}
h(g) &= -\int\limits_Q g \ln g \, dx = -\int\limits_Q g \ln \left(\frac{g}{f^*} f^* \right) dx = \\
&= -h(g||f^*) - \int\limits_Q g \ln f^* dx < -\int\limits_Q g \ln f^* dx = \\
&= -\int\limits_Q g \left(\lambda_0 - 1 + \sum_i \lambda_i r_i(x) \right) dx = \\
&= -\int\limits_Q f^* \left(\lambda_0 - 1 + \sum_i \lambda_i r_i(x) \right) dx = \\
&= -\int\limits_Q f^* \ln f^* dx = h(f^*),
\end{aligned}$$

где во второй строке использовано то, что $h(g||f^*) > 0$ при $g \neq f^*$ (1.44), а при переходе от третьей к четвертой строке — то, что обе функции g и f^* удовлетворяют условиям (1.47), (1.48) и (1.49). Поэтому $h(g) < h(f^*)$ и теорема доказана. □

Аналогичный результат справедлив и для дискретного случая (см. задачу 1.5.3).

Пример 1.7. Пусть $Q = [a, b]$, а других условий не задано. Тогда в соответствии с (1.51) форма максимизирующего дифференциальную энтропию распределения $f^*(x) = e^{\lambda_0 - 1}$. Из необходимости выполнения условий (1.47) и (1.48) получаем $f^*(x) = 1/(b-a)$. То есть равномерное распределение имеет наибольшую энтропию (1.34) из всех распределений, заданных на интервале $Q = [a, b]$, при отсутствии других ограничений.

Пример 1.8. Найдем максимизирующее дифференциальную энтропию распределение при следующих условиях: $Q = [0, \infty)$ и $\overline{X} = \mu$. Так как задано только математическое ожидание ($r_1(x) = x$, $r_i(x) = 0$ для $i > 1$), будем искать такое распределение в форме (1.51):

$$f^*(x) = e^{\lambda_0 - 1 + \lambda_1 x}. \tag{1.52}$$

В соответствии с условиями (1.48) и (1.49) запишем систему уравнений

$$\begin{cases} \int\limits_0^\infty e^{\lambda_0 - 1 + \lambda_1 x} dx = 1, \\ \int\limits_0^\infty e^{\lambda_0 - 1 + \lambda_1 x} x dx = \mu, \end{cases}$$

откуда найдем, что $\lambda_0 = 1 - \ln \mu$ и $\lambda_1 = -1/\mu$. Подставляя найденные значения в (1.52), получим

$$f^*(x) = \frac{1}{\mu}\, e^{-x/\mu} \text{ при } x \geq 0, \tag{1.53}$$

что соответствует **экспоненциальному** распределению с дифференциальной энтропией

$$h(f^*) = -\int\limits_0^\infty f^*(x) \log f^*(x) dx = \log(\mu e). \tag{1.54}$$

Условия последнего примера соответствуют в физике постановке задачи нахождения плотности распределения молекул газа в атмосфере по высоте $x \geq 0$, когда фиксированной считается средняя (по высоте) потенциальная энергия молекул газа μ и известно, что газ стремится перейти в состояние с максимальной энтропией. Плотность нашей атмосферы на самом деле описывается формулой (1.53).

Теперь рассмотрим аналогичную задачу для дискретного случая.

Пример 1.9. Пусть $P(X = k) = f(k)$ при $k = 0, 1, \ldots$ и $\overline{X} = \mu$. Тогда форма искомого распределения $f^*(k) = e^{\lambda_0 - 1 + \lambda_1 k}$ и необходимо выполнение условий:

$$\begin{cases} \sum\limits_{k=0}^{\infty} f^*(k) = 1, \\ \sum\limits_{k=0}^{\infty} f^*(k)k = \mu, \end{cases}$$

откуда находим, что $\lambda_0 = 1 - \ln(\mu + 1)$ и $\lambda_1 = \ln \mu - \ln(\mu + 1)$. Поэтому

$$f^*(k) = \frac{1}{\mu + 1} \left(\frac{\mu}{\mu + 1} \right)^k \text{ при } k = 0, 1, 2, \ldots,$$

что соответствует **геометрическому** распределению с энтропией

$$H(X) = -\sum_{k=0}^{\infty} f^*(k) \log f^*(k) = (\mu + 1) \log(\mu + 1) - \mu \log \mu. \qquad (1.55)$$

Заметим, что, несмотря на кажущуюся одинаковость постановок задачи для непрерывного и дискретного случаев, в двух последних примерах мы получаем различные результаты.

Заметим также, что формулы (1.54) и (1.55) задают **верхнюю границу энтропии** для неотрицательных случайных величин с заданным математическим ожиданием. Используя такой подход, можно получать верхние границы энтропии и при других условиях.

Пример 1.10. Пусть $Q = (-\infty, \infty)$, $\overline{X} = \mu$ и $\overline{X^2} = \sigma^2 + \mu^2$. Тогда форма искомого распределения

$$f^*(x) = e^{\lambda_0 - 1 + \lambda_1 x + \lambda_2 x^2}. \qquad (1.56)$$

Не находя параметры, мы можем заметить, что (1.56) совпадает по форме с гауссовским распределением (1.35), поэтому оно максимизирует дифференциальную энтропию при фиксированных первом и втором моментах и бесконечном интервале значений X.

В этом смысле гауссовское распределение при данных условиях является наиболее неопределенным из всех распределений, а аддитивный гауссовский шум является ≪наихудшим≫ из всех шумов (см. задачу 1.5.4).

Если же $Q = (-\infty, \infty)$ и задано только $\overline{X} = \mu$, то максимальная энтропия бесконечна и не существует максимизирующего энтропию распределения. В качестве пояснения можно рассмотреть гауссовское распределение с увеличивающейся до бесконечности дисперсией.

1.6. Энтропия дискретных случайных процессов

Определение 1.9. Дискретным случайным процессом $\underline{X}$ *называется проиндексированная последовательность* $\underline{X} = \{X_1, X_2, \ldots, X_n\}$ *случайных величин* X_t, $t = 1, 2, \ldots, n$, *каждая с конечным алфавитом* $\mathcal{A}_X$. **Алфавит случайного процесса** $\underline{X}$ *представляет собой Декартово произведение алфавитов каждой из случайных величин и обозначается через* $\mathcal{A}_X^n$. *Случайный процесс задается* n**-мерной совместной функцией вероятности** $f(\underline{x})$, *где последовательность*

$$\underline{x} = \{x_1, x_2, \ldots, x_n\} = \{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}$$

называется **реализацией** *случайного процесса.*

Типичной интерпретацией индекса t является время.

Определение 1.10. Энтропия $H(\underline{X})$ *случайного процесса* $\underline{X}$ *определяется как*

$$H(\underline{X}) = \overline{I(\underline{X} = \underline{x})} = \overline{-\log f(\underline{X})} = -\sum_{\underline{x}} f(\underline{x}) \log f(\underline{x}), \tag{1.57}$$

а **взаимная информация случайных процессов** $\underline{X}$ *и* $\underline{Y}$ *— как*

$$\begin{aligned} I(\underline{X};\underline{Y}) &= \overline{I(\underline{X} = \underline{x}; \underline{Y} = \underline{y})} = \\ &= \overline{\log \frac{f(\underline{X}|\underline{Y})}{f(\underline{X})}} = \sum_{\underline{x}} \sum_{\underline{y}} f(\underline{x}, \underline{y}) \log \frac{f(\underline{x}|\underline{y})}{f(\underline{x})}. \end{aligned} \tag{1.58}$$

Данные определения аналогичны определениям $H(X)$ (1.16) и $I(X;Y)$ (1.17) для случайных величин. По аналогии с (1.23) и (1.32) можно определить и **совместную энтропию** $H(\underline{X}, \underline{Y})$, а также **условную энтропию** $H(\underline{X}|\underline{Y})$ случайных процессов.

Рассмотренные в разделе 1.2 свойства информации и энтропии отдельных случайных величин могут быть без труда перенесены на случайные процессы, однако для них имеется и ряд новых или обобщающих свойств. Например, энтропия случайного процесса обладает свойством **иерархической аддитивности**.

Теорема 1.9. *Пусть* $\underline{X} = \{X_1, X_2, \ldots, X_n\}$ *— некоторый случайный процесс. Тогда*

$$H(\underline{X}) = H(X_1) + \sum_{t=2}^{n} H(X_t | X_{t-1}, \ldots, X_1). \tag{1.59}$$

Доказательство. В соответствии с диаграммой Эйлера (с. 28) имеем

$$\begin{aligned} H(X_1, X_2) &= H(X_1) + H(X_2|X_1), \\ H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3|X_1) = \\ &= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1), \\ &\vdots \\ H(X_1, X_2, \ldots, X_n) &= H(X_1) + H(X_2|X_1) + \ldots + \\ &\quad + H(X_n|X_{n-1}, \ldots, X_1). \end{aligned}$$

Альтернативное доказательство равенства (1.59) состоит в многократном применении формулы Байеса к функции вероятности $f(\underline{x})$:

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= f(x_1) f(x_2|x_1) \ldots f(x_n|x_{n-1}, \ldots, x_1) = \\ &= f(x_1) \prod_{t=2}^{n} f(x_t|x_{t-1}, \ldots, x_1), \end{aligned} \tag{1.60}$$

что после логарифмирования и усреднения дает (1.59). □

Так как условная энтропия не может превосходить безусловную (1.27), из (1.59) получаем

$$H(\underline{X}) \le \sum_{t=1}^{n} H(X_t), \tag{1.61}$$

причем знак равенства в (1.61) достигается, если все $X_1, X_2, \ldots, X_n$ независимы друг от друга. Неравенство (1.61) называется **границей независимости для энтропии**.

Определение 1.11. Энтропия n-го порядка *случайного процесса* $\underline{X}$ *определяется как*

$$H_n(\underline{X}) = \frac{H(X_1, X_2, \ldots, X_n)}{n}, \tag{1.62}$$

а **абсолютная энтропия** *случайного процесса* $\underline{X}$ *— как*

$$H_\infty(\underline{X}) = \lim_{n \to \infty} H_n(\underline{X}), \tag{1.63}$$

если предел существует.

Величины $H_n(\underline{X})$ (1.62) и $H_\infty(\underline{X})$ (1.63) представляют собой среднюю энтропию, порождаемую случайным процессом в единицу времени, и измеряются в двоичном случае в битах, приходящихся на один символ случайного процесса.

Пример 1.11. Рассмотрим печатную машинку, которая имеет m букв. Пусть эта машинка выдает m^n равновероятных последовательностей $\underline{x}$ длины n. Тогда $H(\underline{X}) = \log_2 m^n$, а $H_n(\underline{X}) = H_\infty(\underline{X}) = \log_2 m$ бит/символ.

Определение 1.12. *Пусть* $\{i_1, i_2, \ldots, i_k\}$, *где* $k \leq n$, — *некоторое подмножество индексов* $\{1, 2, \ldots, n\}$.

Случайный процесс $\underline{X}$ *называется* **стационарным**, *если* k*-мерная функция вероятности* $f(\underline{x})$ *инвариантна относительно сдвига* t *по времени*

$$\begin{aligned} &f(X_{i_1} = x_{i_1}, X_{i_2} = x_{i_2}, \ldots, X_{i_k} = x_{i_k}) = \\ &= f(X_{i_1+t} = x_{i_1}, X_{i_2+t} = x_{i_2}, \ldots, X_{i_k+t} = x_{i_k}) \end{aligned} \tag{1.64}$$

для любого сдвига t, *любого* k *и любых* $i_1, i_2, \ldots, i_k$.

Случайный процесс $\underline{X}$ *называется процессом* **без памяти**, *если для* k*-мерной функции вероятности* $f(\underline{x})$ *имеет место равенство*

$$f(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = \prod_{j=1}^{k} f_{i_j}(x_{i_j}) \tag{1.65}$$

для любого k *и любых* $i_1, i_2, \ldots, i_k$.

Для произвольных процессов без памяти одномерные функции вероятности зависят от момента времени $t = i_j$, поэтому в правой части (1.65) они записаны с соответствующими индексами, однако для стационарных процессов, как следует из (1.64), все они одинаковы для всех моментов времени, поэтому

$$f(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = \prod_{j=1}^{k} f(x_{i_j}), \tag{1.66}$$

откуда следует, что случайные величины, входящие в стационарный случайный процесс без памяти, являются **независимыми и одинаково распределенными (н.о.р.)** случайными величинами. В дальнейшем, рассматривая процессы без памяти, мы будем иметь в виду только стационарные случайные процессы.

Пример 1.12. Пусть $X_1, X_2, \ldots, X_n$ — последовательность н.о.р. случайных величин, энтропия каждой из которых равна $H(X)$. Тогда

$$H_n(\underline{X}) = H_\infty(\underline{X}) = \lim_{n\to\infty} \frac{H(\underline{X})}{n} = \frac{nH(X)}{n} = H(X). \tag{1.67}$$

Если же величины $X_1, X_2, \ldots, X_n$ независимы, однако неодинаково распределены, то в соответствии с границей независимости энтропии (1.61)

$$H(\underline{X}) = \sum_{t=1}^{n} H(X_t),$$

причем в данном случае возможны такие распределения величин X_t, что предел

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} H(X_t) \tag{1.68}$$

либо существует, либо не существует (см. задачу 1.6.3).

Теорема 1.10. *Для стационарного случайного процесса* $\underline{X}$

$$H_\infty(\underline{X}) = \lim_{n\to\infty} H_n(\underline{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, \ldots, X_2, X_1), \tag{1.69}$$

причем последовательности

$$H(X_n|X_{n-1}, \ldots, X_2, X_1), \;\; n = 2, 3, \ldots \tag{1.70}$$

и $H_n(\underline{X})$ *не возрастают с ростом* n.

Доказательство. Так как

$$\begin{aligned} H(X_n|X_{n-1}, \ldots, X_2, X_1) &\leq H(X_n|X_{n-1}, \ldots, X_2) = \\ &= H(X_{n-1}|X_{n-2}, \ldots, X_1), \end{aligned}$$

где неравенство следует из (1.27) (добавление условий не увеличивает энтропию), а равенство следует из стационарности (1.64), последовательность (1.70) в самом деле не возрастает с ростом n. Поэтому предел в правой части (1.69) существует.

Далее, по свойству иерархической аддитивности энтропии (1.59), имеем

$$H_\infty(\underline{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) = \tag{1.71}$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} H(X_t|X_{t-1}, \ldots, X_1), \tag{1.72}$$

причем под знаком предела в (1.72) мы имеем среднее по времени значение невозрастающей последовательности (1.70). Так как со временем каждое из слагаемых в (1.72) стремится к существующему пределу (1.69), то и их среднее под знаком предела в (1.71) не возрастает с ростом n:

$$H_\infty(\underline{X}) \leq H_n(\underline{X}) \tag{1.73}$$

и стремится к этому же пределу, что доказывает теорему. □

1.7. Эргодические и марковские случайные процессы

Пусть $\underline{X} = \{X_{1-k}, X_{2-k}, \ldots, X_0, X_1, X_2, \ldots, X_n\}$ является стационарным случайным процессом с алфавитом $\mathcal{A}_X^{n+k}$. Обозначим

$$z_t = G(x_{t-k}, x_{t-k+1}, \ldots, x_t), \;\; t = 1, 2, \ldots, n,$$

где $G(x_{t-k}, x_{t-k+1}, \ldots, x_t)$ — произвольная действительная функция, определенная на множестве $\mathcal{A}_X^{k+1}$. Свяжем со значениями z_t данной функции случайную величину Z_t. В силу стационарности случайного процесса $\underline{X}$ случайные величины Z_t, $t = 1, 2, \ldots, n$, имеют одинаковую (не зависящую от t, которое для простоты записи примем равным $k+1$) функцию вероятности, а значит, и одинаковое математическое ожидание $\overline{Z}$:

$$\begin{aligned} \overline{Z} &= \overline{G(x_{t-k}, x_{t-k+1}, \ldots, x_t)} = \\ &= \sum_{x_1} \sum_{x_2} \ldots \sum_{x_{k+1}} f(x_1, x_2, \ldots, x_{k+1}) G(x_1, x_2, \ldots, x_{k+1}), \end{aligned}$$

которое может быть истолковано как **среднее по множеству реализаций** отрезков длины $k+1$ случайного процесса $\underline{X}$.

Определим теперь **среднее по времени** величин z_t как

$$\frac{1}{n} \sum_{t=1}^{n} z_t = \frac{1}{n} \sum_{t=1}^{n} G(x_{t-k}, x_{t-k+1}, \ldots, x_t). \tag{1.74}$$

Определение 1.13. *Дискретный стационарный случайный процесс называется* **эргодическим**, *если для любого фиксированного k ($0 \leq \leq k < n$) при $n \to \infty$ и любой определенной на $\mathcal{A}_X^{k+1}$ действительной функции G равна единице вероятность того, что среднее по времени совпадает со средним по множеству реализаций:*

$$P\left(\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} z_t = \overline{Z} \right) = 1, \tag{1.75}$$

при условии, что эти средние существуют.

Имея в виду (1.75), иногда говорят, что среднее по времени стремится **по вероятности** к среднему по множеству реализаций. Суть записи (1.75) заключается в том, что среднее по времени является случайной величиной, а среднее по множеству реализаций — детерминированной, поэтому просто поставить между ними знак равенства нельзя. Для упрощения (1.75) введем понятие **вероятностного предела** plim и получим эквивалентную запись:

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} z_t = \overline{Z}. \tag{1.76}$$

Другими словами (1.75) может быть сформулировано так: для любых $\delta > 0$ и $\varepsilon > 0$ существует такое значение n_0, что для любого $n > n_0$ справедливо

$$P\left(\left|\frac{1}{n}\sum_{t=1}^{n} z_t - \overline{Z}\right| \le \varepsilon\right) \ge 1 - \delta. \tag{1.77}$$

Эргодические случайные процессы являются важным классом стационарных случайных процессов, поскольку многие их характеристики могут быть найдены экспериментально с помощью наблюдения, возможно достаточно длительного, одной случайной величины, входящей в процесс. Действительно, наблюдая одну случайную величину, вычисляя по ней величины z_t, $t = 1, 2, \ldots, n$, и усредняя их по времени, можно достаточно хорошо оценить величину $\overline{Z}$.

Пример 1.13. Пусть дана реализация $\underline{x}$ дискретного эргодического процесса $\underline{X}$ и требуется по ней оценить вероятность, с которой возникает событие $x_1 \in \mathcal{A}_X$. Интуитивно ясно, что для этого достаточно подсчитать частоту появления этого события в данной реализации, однако точное обоснование того, что нужно сделать именно так, базируется на свойстве эргодичности. Пусть $k = 0$, а

$$G(x) = \begin{cases} 1, & \text{если } x = x_1, \\ 0, & \text{если } x \ne x_1. \end{cases}$$

Тогда $\overline{Z} = \overline{G(X)} = P(x = x_1)\cdot 1 + P(x \ne x_1)\cdot 0 = P(x = x_1)$, а $z_t = G(x_t)$. Очевидно, что $\frac{1}{n}\sum\limits_{t=1}^{n} z_t$ есть частота появления события x_1, что и есть обоснование того, что частота является хорошей оценкой вероятности.

Однако не все стационарные случайные процессы обладают свойством эргодичности, что показывает следующий пример.

Пример 1.14. Пусть имеются две урны, в первой из которых лежат два белых и один черный шар, а во второй — два черных и один белый. Пусть экспериментатор сначала случайно выбирает одну из урн, причем первая выбирается с вероятностью $0 < p < 1$. Затем он многократно извлекает из выбранной урны случайно некоторый шар, фиксирует его цвет и возвращает шар назад. Если в начале эксперимента выбрана первая урна, то частота появления белого шара будет равна 2/3, в противном случае — 1/3. Если имеется n независимо работающих экспериментаторов, то примерно np из них зарегистрируют частоту 2/3, а примерно $n(1-p)$ — частоту 1/3. Частота появления белого шара в среднем по множеству всех экспериментаторов будет примерно равна $p(2/3)+(1-p)(1/3) = (1+p)/3$ и не будет совпадать с частотой, полученной любым из экспериментаторов. Поэтому описанный случайный процесс не будет эргодическим.

Для произвольного случайного процесса $\underline{X}$ возможна любая зависимость между величинами X_t. Примером процесса с зависимостью является **марковский процесс** k-го порядка, для которого каждая случайная величина зависит напрямую только от k предшествующих случайных величин и тем самым зависит только косвенно от всех остальных предшествующих случайных величин

$$\begin{aligned} &f(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_{2-k} = x_{2-k}, X_{1-k} = x_{1-k}) = \\ &= f(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_{t-k} = x_{t-k}) \end{aligned} \tag{1.78}$$

для любого $t = 1, 2, \ldots, n$.

Марковский процесс первого порядка называется **цепью Маркова**. Функция вероятности (1.60) для нее может быть записана как

$$f(\underline{x}) = f(x_{1-k}) f(x_{2-k}|x_{1-k}) \ldots f(x_n|x_{n-1}), \tag{1.79}$$

а значение X_t называется **состоянием** цепи в момент времени t. Переходы между состояниями (при изменении t) характеризуются своими вероятностями $P(X_t = i|X_{t-1} = j)$, $i, j \in \mathcal{A}_X$, $t = 2-k, \ldots, n$, причем если из любого состояния возможно перейти в любое другое состояние с ненулевой вероятностью за конечное число шагов, то цепь Маркова называется **неразложимой**. Суть этого термина заключается в том, что множество состояний цепи Маркова не разложимо на такие подмножества, что, попав в какое-либо состояние одного из подмножеств, невозможно из этого подмножества выйти.

Необходимым условием стационарности цепи Маркова является ее **однородность**, то есть независимость от времени условных функций

вероятности $f(x_t|x_{t-1})$ в (1.79). Но одномерные функции вероятности $f(x_t)$ (представляющие собой начальные условия), как правило, зависят от t, и весь процесс не стационарен. Однако во многих случаях после некоторого начального переходного процесса он стремится к стационарному при $t \to \infty$. И более того, если цепь Маркова стационарна и неразложима, то она эргодична. Два последних факта мы приводим без доказательств, поскольку их можно найти в любом учебнике по случайным процессам.

Если в записи (1.60) для $f(\underline{x})$ произвольного случайного процесса $\underline{X}$ учитывать зависимость каждого символа x_t от k предыдущих символов (а не от всех предыдущих, как в (1.60), и не от одного, как в (1.79)), получим функцию

$$Q_k(\underline{x}) = f(x_{1-k}, x_{2-k}, \ldots, x_0) \prod_{t=1}^{n} f(x_t|x_{t-1}, \ldots, x_{t-k}), \tag{1.80}$$

которая называется **марковской аппроксимацией k-го порядка** функции $f(\underline{x})$. Так же как и для $f(\underline{x})$, для нее выполняется свойство

$$\sum_{\underline{x}} Q_k(\underline{x}) = 1, \tag{1.81}$$

что легко проверяется суммированием сначала по x_n, потом по x_{n-1} и, наконец, по x_1 и $\{x_{1-k}, x_{2-k}, \ldots, x_0\}$.

Лемма 1.1. *Для эргодического случайного процесса $\underline{X}$ имеет место вероятностный предел*

$$\operatorname*{plim}_{n\to\infty} -\frac{1}{n} \log Q_k(\underline{x}) = H(X_{k+1}|X_k, \ldots, X_1), \tag{1.82}$$

причем для достаточно большого k справедливо

$$\operatorname*{plim}_{n\to\infty} -\frac{1}{n} \log Q_k(\underline{x}) = H_\infty(\underline{X}). \tag{1.83}$$

Доказательство. Из (1.80) имеем

$$-\frac{1}{n} \log Q_k(\underline{x}) = -\frac{1}{n} \log f(x_{1-k}, \ldots, x_0) + \frac{1}{n} \sum_{t=1}^{n} -\log f(x_t|x_{t-1}, \ldots, x_{t-k}),$$

где первое слагаемое стремится к нулю при $n \to \infty$, поскольку $f(x_{1-k}, \ldots, x_0) \neq 0$ (то есть первые k символов случайного процесса существуют), а второе слагаемое является средним по времени (1.74) для функции $G(x_{t-k}, x_{t-k+1}, \ldots, x_t) = -\log f(x_t|x_{t-1}, \ldots, x_{t-k})$. Среднее же по

реализации для той же функции для любого $t = 1, 2, \ldots, n$ равно

$$\overline{-\log f(x_t|x_{t-1}, \ldots, x_{t-k})} = H(X_{k+1}|X_k, \ldots, X_1),$$

поскольку для $t = k+1$ это следует из определения условной энтропии, а для остальных t — из стационарности случайного процесса. Поэтому (1.82) выполняется в силу свойства эргодичности.

Кроме того, в силу стационарности случайного процесса из теоремы 1.10 (равенство (1.69)) следует, что для любого сколь угодно малого, но фиксированного $\delta > 0$ найдется такое (возможно, достаточно большое) k, что

$$H(X_{k+1}|X_k, \ldots, X_1) - H_\infty(\underline{X}) \leq \delta, \tag{1.84}$$

поэтому (1.83) следует из (1.82) и (1.84). □

Лемма 1.2. *Марковская аппроксимация k-го порядка при достаточно большом k хорошо приближает функцию $f(\underline{x})$ эргодического случайного процесса $\underline{X}$:*

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} \log Q_k(\underline{x}) = \frac{1}{n} \log f(\underline{x}). \tag{1.85}$$

Доказательство. Перепишем (1.85) в форме (1.77):

$$P\left(\left|\frac{1}{n} \log Q_k(\underline{x}) - \frac{1}{n} \log f(\underline{x})\right| \geq \varepsilon\right) \leq \delta. \tag{1.86}$$

Используя **неравенство Чебышева**, оценим левую часть (1.86) как

$$P\left(\left|\log \frac{Q_k(\underline{x})}{f(\underline{x})}\right| \geq n\varepsilon\right) \leq \frac{1}{n\varepsilon} \overline{\left|\log \frac{Q_k(\underline{x})}{f(\underline{x})}\right|}.$$

При $Q_k(\underline{x})/f(\underline{x}) \geq 1$ имеем

$$\begin{aligned}\overline{\left|\log \frac{Q_k(\underline{x})}{f(\underline{x})}\right|} &= \overline{\log \frac{Q_k(\underline{x})}{f(\underline{x})}} \leq \log \overline{\frac{Q_k(\underline{x})}{f(\underline{x})}} = \\ &= \log \sum_{\underline{x}} f(\underline{x}) \frac{Q_k(\underline{x})}{f(\underline{x})} = \log \sum_{\underline{x}} Q_k(\underline{x}) = 0,\end{aligned}$$

где неравенство следует из неравенства Йенсена, а последнее равенство — из (1.81). Поэтому (1.86) выполняется. При $Q_k(\underline{x})/f(\underline{x}) < 1$

имеем

$$\begin{aligned}
\frac{1}{n}\overline{\left|\log\frac{Q_k(\underline{x})}{f(\underline{x})}\right|} &= \frac{1}{n}\overline{\left(-\log\frac{Q_k(\underline{x})}{f(\underline{x})}\right)} = \\
&= \frac{1}{n}\overline{-\log Q_k(\underline{x})} - \frac{1}{n}\overline{-\log f(\underline{x})} = \\
&= \sum_{\underline{x}} f(\underline{x})\left(-\frac{1}{n}\log Q_k(\underline{x})\right) - H_n(\underline{X}) \overset{n\to\infty}{\to} \\
&\overset{n\to\infty}{\to} H(X_{k+1}|X_k,\ldots,X_1) - H_\infty(\underline{X}) \le \delta,
\end{aligned}$$

где в предельном переходе использовано (1.82), а неравенство есть (1.84). Поэтому (1.86) выполняется всегда, а вместе с ним и (1.85) для достаточно больших k. □

Таким образом, из лемм 1.1 (равенство (1.83)) и 1.2 следует теорема **Шеннона – МакМиллана**.

Теорема 1.11. *Для эргодического случайного процесса $\underline{X}$ справедливо*

$$\operatorname*{plim}_{n\to\infty} -\frac{1}{n}\log f(\underline{x}) = H_\infty(\underline{X}). \tag{1.87}$$

Справедливость леммы 1.2 для какого-либо случайного процесса может служить основанием для предположения о его эргодичности. Проиллюстрируем этот факт на примере проведенного Шенноном моделирования английского языка в его «упрощенном» варианте, содержащем только 26 букв и пробел. При этом мы игнорируем все остальные особенности языка — знаки препинания, большие и маленькие буквы, абзацы и т. п.

Для моделирования нам необходимо знать вероятности, с которыми в языке встречаются символы, их пары, тройки и т. д. Эти вероятности можно вычислить, анализируя реальные тексты. Известно, например, что вероятность пробела в английском языке равна 0,18, вероятность самой вероятной буквы «E» равна 0,13, а самой «невероятной» буквы «Z» — 0,05. Самой вероятной парой букв является «TH» (0,033), за ней следует «HE» (0,027) и так далее. Результаты моделирования оказались следующими:

1. Аппроксимация нулевого порядка (все символы независимы и равновероятны):

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ
   FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD

2. Аппроксимация первого порядка (символы независимы, но их частоты соответствуют английскому тексту):

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA
   TH EEI ALHENHTTPA OOBTTVA NAH BRL

3. Аппроксимация второго порядка (частоты пар соответствуют английскому тексту):

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY
   ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO
   TIZIN ANDY TOBE SEACE CTISBE

4. Аппроксимация третьего порядка (частоты троек соответствуют английскому тексту):

   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID
   PONDENOME OF DEMONSTURES OF THE REPTAGIN IS
   REGOACTIONA OF CRE

5. Аппроксимация четвертого порядка (частоты четверок соответствуют английскому тексту):

   THE GENERATED JOB PROVIDUAL BETTER TRAND
   THE DISPLAYED CODE ABOVERY UPONDULTS WELL
   THE CODERST IN THESTICAL IT DO HOCK BOTHE
   MERG INSTATES CONS ERATION NEVER ANY OF
   PUBLE AND TO THEORY EVENTIAL CALLEGAND TO
   ELAST BENERATED IN WITH PIES AS IS WITH THE

6. По аналогии можно моделировать язык не только по символам, но и по словам. В этом примере мы имеем пословную аппроксимацию первого порядка (слова выбраны независимо, но в соответствии с их частотой в английском языке):

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR
   COME CAN DIFFERENT NATURAL HERE HE THE A IN
   CAME THE TO OF TO EXPERT GRAY COME TO
   FURNISHES THE LINE MESSAGE HAD BE THESE

7. Пословная аппроксимация второго порядка (частоты пар слов соответствуют английскому тексту):

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH
   WRITER THAT THE CHARACTER OF THIS POINT IS

THEREFORE ANOTHER METHOD FOR THE LETTERS
THAT THE TIME OF WHO EVER TOLD THE PROBLEM
FOR AN UNEXPECTED[2]

Мы видим, что аппроксимации все больше и больше приближаются к реальному языку, хотя даже аппроксимация четвертого порядка еще не учитывает всей его структуры. Исследования показывают, что статистические зависимости в реальных языках сохраняются примерно до тридцатой буквы. Поэтому если язык является эргодическим процессом, то упоминаемое в лемме 1.2 ≪достаточно большое k≫ равно для него тридцати.

Подобные аппроксимации могут быть использованы для вычисления энтропии языка как случайного процесса. С увеличением сложности модели мы учитываем все больше особенностей языка, а условная неопределенность каждой следующей буквы уменьшается. Для английского языка имеем следующую убывающую последовательность энтропий:

$$H_0 = \log_2 27 \approx 4.75, \;\; H_1(X_1) \approx 4{,}03,$$
$$H_2(X_1, X_2) \approx 3.32, \;\; H_3(X_1, X_2, X_3) \approx 3{,}10,$$
$$H_4(X_1, X_2, X_3, X_4) \approx 2.80, \ldots, H_\infty(\underline{X}) \approx 1{,}34, \qquad (1.88)$$

как и следует ожидать для стационарного процесса в соответствии с (1.73). Для вычисления этих энтропий были найдены вероятности отдельных символов, их пар, троек и четверок и т. д. по некоторому конкретному тексту (реализации случайного процесса). При этом использовано предположение об эргодичности и свойства одной реализации распространены на весь язык.

Однако известно, что тексты разного содержания имеют разные вероятностные характеристики. Например, русская разговорная речь имеет абсолютную энтропию 1,40 бит/символ, а деловой текст ≪всего≫ 0,83 бит/символ. Поэтому предположение об эргодичности языка, если и справедливо, то в пределах определенной группы текстов.

Тем не менее статистические модели языка играют ключевую роль в системах распознавания речи, когда без знания сложных грамматических правил на основании уже сказанных слов можно оценить наиболее вероятное следующее слово, которое по каким-либо причинам нераспознаваемо.

[2]В блестящем переводе Елены Саракваш по-русски эта тарабарщина звучит так: ≪Интеллект и в лобовой атаке на английского писателя, что характерной особенностью этого вопроса являются поэтому другие литературные средства, ритм которых неизменно указывал на проблему для непредсказуемого≫.

1.8. Колмогоровская сложность

Отличный от шенноновского подход к определению количества информации, основанный на понятии сложности определенного объекта, предложил в 1965 году великий математик А. Колмогоров.

Рассмотрим три последовательности:

$\underline{x}_1 = 01010101010101010101010101010101010101010101010101\ldots$

$\underline{x}_2 = 00100100001111110110101010001000100001011010001100001000\ldots$

$\underline{x}_3 = 10110101010010011110100011001001001011011110010100011010\ldots$

Первая из них выглядит простой и представляет собой повторение 28 раз последовательности 01. Вторая и третья выглядят случайными, однако вторая представляет собой первые 56 бит после запятой в двоичном представлении числа π, тогда как третья получена в результате подбрасывания монеты 56 раз.

Идея Колмогорова состоит в том, чтобы описать сложность объекта (в данном случае последовательности) как длину самой короткой компьютерной программы, которая воспроизводит данный объект:

$$\mathcal{K}_{\mathcal{U}}(\underline{x}) = \min_{P:\ \mathcal{U}(P)=\underline{x}} l(P),$$

где $\mathcal{K}_{\mathcal{U}}(\underline{x})$ — колмогоровская сложность последовательности $\underline{x}$ относительно универсального компьютера[3] $\mathcal{U}$ (которым может быть, например, универсальная машина Тьюринга), $\mathcal{U}(P)$ — выход компьютера $\mathcal{U}$ после выполнения программы P, $l(P)$ — длина программы P.

Очевидно, что для первой последовательности компьютеру необходимо написать инструкцию вроде: «Повторить 28 раз 01», для второй — «Вычислить первые n бит после запятой в представлении числа π», задав при этом какой-либо алгоритм для вычисления, например с помощью ряда $\pi = 4\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{2k-1}$, и определив точность вычисления, а для третьей — придется написать: «Выдать последовательность 1011010101001001...».

Если компьютеру заранее известна длина последовательности $l(\underline{x})$, которую необходимо выдать, то длина программы в первом и втором случаях является константой, причем очевидно, что во втором случае эта константа несколько больше, чем в первом, а в третьем случае длина программы равна длине последовательности плюс некоторая константа, отражающая инструкции компьютеру. Поэтому понятен следующий результат.

[3]Напомним, что универсальным компьютером называется компьютер, способный моделировать работу любого другого компьютера.

Теорема 1.12. *Условная колмогоровская сложность последовательности* $\underline{x}$ *не превосходит ее длину:*

$$\mathcal{K}_{\mathcal{U}}(\underline{x}|l(\underline{x})) = \min_{P:\ \mathcal{U}(P,\ l(\underline{x}))=\underline{x}} l(P) \le l(\underline{x}) + c,$$

где c *— некоторая константа, не зависящая от* $\underline{x}$*.*

Понятно также и то, что если $\mathcal{U}$ — универсальный компьютер, то для любого другого компьютера $\mathcal{V}$: $\mathcal{K}_{\mathcal{U}}(\underline{x}) \le \mathcal{K}_{\mathcal{V}}(\underline{x}) + c_{\mathcal{V}}$, где константа $c_{\mathcal{V}}$ определяет длину инструкций для моделирования компьютера $\mathcal{V}$ на компьютере $\mathcal{U}$ и не зависит от $\underline{x}$. Поэтому в записи колмогоровской сложности можно без потери общности опускать индекс, означающий конкретный компьютер.

Если длина последовательности $l(\underline{x}) = n$ компьютеру заранее не известна, то ему необходимо каким-то образом сообщить, когда в процессе генерации $\underline{x}$ он должен остановиться. Для описания длины n повторим каждый бит в ее двоичном представлении дважды, а в конце поставим 01. Таким образом компьютер будет знать, что представление n закончено. Поэтому дополнение к длине программы информации о длине последовательности $l(\underline{x})$ не составит более $2\log n + 2$ бит, и справедлива следующая теорема.

Теорема 1.13. *Колмогоровская сложность последовательности* $\underline{x}$ *удовлетворяет неравенству*

$$\mathcal{K}(\underline{x}) \le \mathcal{K}(\underline{x}|n) + 2\log n + c. \tag{1.89}$$

Определим теперь колмогоровскую сложность двоичной последовательности длины n с k единицами. Компьютерная программа для ее воспроизведения может выглядеть следующим образом: «Сгенегировать в лексикографическом порядке все последовательности длины n с k единицами и выбрать i-ю из них».

Для представления k нам достаточно $2\log k + 2$ бит, а для представления i — достаточно $\log \binom{n}{k}$ бит (здесь мы уже можем отказаться от повторения битов дважды, так как i может заканчивать программу). Поэтому колмогоровская сложность такой последовательности будет равна

$$\mathcal{K}(\underline{x}|n) = 2\log k + \log \binom{n}{k} + c \le 2\log k + n\eta\left(\frac{k}{n}\right) + c, \tag{1.90}$$

где η — двоичная функция энтропии (1.18), а неравенство следует из формулы Стирлинга: $\binom{n}{k} \le 2^{n\eta\left(\frac{k}{n}\right)}$.

Обобщая последние рассуждения на последовательности с произвольным количеством единиц и учитывая, что $\log k \le \log n$, из (1.90) получаем следующую теорему.

Теорема 1.14. *Колмогоровская сложность двоичной последовательности длины n удовлетворяет неравенству*

$$\mathcal{K}(\underline{x}|n) \le n\eta\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + 2\log n + c, \tag{1.91}$$

где сумма представляет собой количество единиц в последовательности $\underline{x}$.

Теперь покажем, что ожидаемое значение колмогоровской сложности случайной последовательности близко к энтропии источника, ее породившего.

Теорема 1.15. *Если двоичная последовательность $\underline{x}$ состоит из значений н.о.р. случайных величин X_i, $i = 1, 2, \ldots, n$, с функцией вероятности $f(1) = p$, $f(0) = 1 - p$ и энтропией $\eta(p) = H(X)$, то для любого n справедливо*

$$H(X) \le \frac{1}{n}\overline{\mathcal{K}(\underline{x}|n)} = \frac{1}{n}\sum_{\underline{x}} f(\underline{x})\mathcal{K}(\underline{x}|n) \le H(X) + \frac{2\log n}{n} + \frac{c}{n}. \tag{1.92}$$

Доказательство. Действительно, усредним (1.91) по всем последовательностям $\underline{x}$:

$$\begin{aligned}
\overline{\mathcal{K}(\underline{x}|n)} &\le n\overline{\eta\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)} + 2\log n + c \le \\
&\le n\eta\left(\frac{1}{n}\sum_{i=1}^{n}\overline{X_i}\right) + 2\log n + c = \\
&= n\eta(p) + 2\log n + c,
\end{aligned} \tag{1.93}$$

где второе неравенство следует из выпуклости двоичной функции энтропии (рис. 1.1) и неравенства Йенсена (1.8). Поэтому из (1.93) следует верхняя граница в (1.92).

Для доказательства нижней границы в (1.92) заметим, что каждой строке $\underline{x}$ соответствует некоторая программа P минимальной длины, после выполнения которой компьютер произведет строку $\underline{x}$. Из результатов главы 3 будет ясно, что средняя длина такой программы, как и средняя длина кодового слова оптимального кода, не может быть меньше энтропии источника (3.8). □

Результат теоремы 1.15 показывает, что колмогоровская сложность и шенноновская энтропия дают очень схожие меры информации, однако концептуальное различие между ними состоит в том, что колмогоровская сложность не оперирует понятием вероятности, а рассматривает тот или иной объект целиком. Кроме того, она оперирует не только синтаксическим аспектом, но через инструкции компьютеру учитывает также семантику объектов.

1.9. Задачи

1.1.1. [1 балл] Дайте графическую интерпретацию определения выпуклости функции по неравенству (1.7).

1.2.1. [2 балла] Пусть мы имеем две монеты, одна из которых настоящая, а у другой с двух сторон «орел». Выберем случайно одну из монет и подбросим ее дважды. Сколько информации о том, настоящая монета или нет, мы при этом получим?

1.2.2. [2 балла] Пусть мы сначала бросаем кубик с гранями, пронумерованными от 1 до 6, а затем подбрасываем монету такое количество раз, которое выпало на кубике. Сколько информации мы при этом получаем?

1.2.3. [2 балла] Пусть мы подбрасываем монету до тех пор, пока не выпадет «орел», а величина X описывает необходимое число подбрасываний. Найдите $H(X)$.

1.2.4. [2 балла] На шахматной доске в одной из клеток поставлена фигура, причем все положения фигуры одинаково вероятны. Какое минимальное количество вопросов, на которые можно ответить «да» или «нет», нужно поставить, чтобы определить положение фигуры? Как нужно задавать вопросы, чтобы минимизировать их количество?

1.2.5. [2 балла] Докажите справедливость пословицы «Лучше один раз увидеть, чем сто раз услышать», исходя из предположения, что увидеть мы собираемся случайную картинку на экране компьютерного монитора с разрешением 1024×1280 точек и цветом, описываемым 32 битами на точку, а услышать — 100 случайно выбранных слов из словаря объемом в 300 тысяч слов.

1.2.6. [2 балла] Статистика говорит, что 70 % мужчин имеют темные волосы, 25 % женщин — блондинки, а 80 % блондинок выходят замуж за темноволосых мужчин. Сколько информации о цвете волос мужа несет цвет волос его жены?

1.2.7. [2 балла] Докажите, что энтропия случайной величины Y, связанной со случайной величиной X произвольной функцией $y = G(x)$, не превосходит энтропию X: $H(Y) \leq H(X)$. Для каких функций $G(x)$ выполняется знак равенства?

1.2.8. [2 балла] При каких условиях

(а) из того, что $H(X|Y) = 0$, следует, что $H(Y|X) = 0$?

(б) выполняется равенство $H(X|G(Y)) = H(X|Y)$?

1.2.9. [2 балла] Определим расстояние между случайными величинами X и Y как $d(X, Y) = H(X|Y) + H(Y|X)$. Покажите, что для любых трех случайных величин X, Y и Z выполняется неравенство треугольника $d(X, Y) + d(Y, Z) \geq d(X, Z)$.

1.2.10. [3 балла] Пусть для некоторой дискретной случайной величины X определена функция вероятности $f(x_i)$, $i = 1, 2, \ldots, L$, а функция вероятности случайной величины Y отличается от $f(x_i)$ только в j-й и l-й позициях так, что $f(y_j) = f(y_l) = (f(x_j) + f(x_l))/2$. Докажите, что $H(X) \leq H(Y)$, а также что любое перемещение вероятности, приводящее к более равномерному распределению, увеличивает энтропию.

1.2.11. [3 балла] Пусть случайная величина X имеет биномиальное распределение: $P(X = k) = \binom{n}{k} p^k q^{n-k}$, где $0 \leq k \leq n$, $0 < p < 1$ и $q = 1 - p$. Покажите, что $H(X) \leq -n(p \log p + q \log q)$.

1.2.12. [5 баллов] Пусть мы имеем аптекарские весы и $n \geq 3$ монет, среди которых может быть, а может и не быть одна фальшивая, которая отличается от других только по весу. Найдите верхнюю границу для количества монет n при условии, что k взвешиваний помогут найти фальшивую монету, если она имеется, и если имеется, помогут определить, тяжелее она или легче остальных. Найдите при этих условиях стратегию для $n = 12$ монет и $k = 3$ взвешиваний.

1.2.13. [5 баллов] Рассмотрим последовательность непрерывных по всем аргументам функций $H_L(p_1, p_2, \ldots, p_L)$, где $L = 2, 3, \ldots$, а значения p_i задают функцию вероятности некоторой случайной величины X: $p_i = f(x_i)$ для всех $i = 1, 2, \ldots, L$, причем $H_2(\frac{1}{2}, \frac{1}{2}) = 1$. Докажите, что если для любого L выполняется **условие группировки**

$$H_L(p_1, p_2, \ldots, p_L) = H_{L-1}(p_1 + p_2, p_3 \ldots, p_L) + \\ + (p_1 + p_2) H_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right),$$

то $H_L(p_1, p_2, \ldots, p_L) = -\sum\limits_{i=1}^{L} p_i \log p_i = H(X)$, причем энтропия — единственная функция, удовлетворяющая данному условию.

Таким образом может быть дано **аксиоматическое определение энтропии**, так как условие группировки соответствует интуитивному пониманию того, что неопределенность конечных событий с вероятностями $p_1, p_2, \ldots, p_L$ не может измениться от введения промежуточного события с вероятностью $p_1 + p_2$.

1.3.1. [2 балла] Дайте интерпретацию количества семантической информации при различных значениях T_1, T_2, T_{opt} и $I_{semant}(T_{opt})$ на рис. 1.4.

1.4.1. [2 балла] Пусть $f_1(x)$ и $f_2(x)$ — плотности вероятности двух непрерывных случайных величин, отличающихся только математическим ожиданием. Докажите, что дифференциальные энтропии $h_1(X)$ и $h_2(X)$ этих случайных величин одинаковы.

1.4.2. [2 балла] Вычислите дифференциальную энтропию суммы двух независимых гауссовских случайных величин с различающимися математическими ожиданиями и дисперсиями.

1.4.3. [3 балла] Постройте непрерывную случайную величину X с $h(X) = -\infty$.

1.4.4. [4 балла] Предположим, мы имеем непрерывную случайную величину X с плотностью вероятности $f(x)$, причем величина $f(x) \log f(x)$ интегрируема по Риману. Разделим ненулевое множество X на n отрезков длины Δ и предположим, что $f(x)$ непрерывна внутри каждого отрезка. Определим дискретную случайную величину Y как $y_i = x_i$, если $(i-1)\Delta \le x < i\Delta$, $i = 1, 2, \ldots, n$, где x_i — среднее значение функции $f(x)$ в i-м интервале: $f(x_i)\Delta = \int_{(i-1)\Delta}^{i\Delta} f(x)dx$. Выразите $H(Y)$ через $h(X)$ и Δ. Что произойдет с $H(Y)$, если $\Delta \to 0$?

1.4.5. [2 балла] Пусть $f(x)$ и $g(x)$ — две функции вероятности случайных величин, определенных на множестве $\{x_1, x_2\}$, причем $f(x_1) = r$, $f(x_2) = 1 - r$, а $g(x_1) = s$, $g(x_2) = 1 - s$. Постройте графики относительных энтропий $h(f||g)$ и $h(g||f)$ в зависимости от r и s. При каких условиях $h(f||g) = h(g||f)$?

1.5.1. [3 балла] Найдите максимизирующее энтропию распределение случайной величины X, определенной на множестве $Q = [0, \infty)$, если задано:

(а) $\overline{X} = \alpha_1$ и $\overline{\ln X} = \alpha_2$,

(б) $\overline{X} = \alpha_1$ и $\overline{X^2} = \alpha_2$.

1.5.2. [5 баллов] Предположим, n кубиков брошено на стол и нам известно, что общая сумма выпавших значений равна $n\alpha$. Какова ожидаемая часть кубиков, на которых выпала i-я грань ($i = 1, 2, \ldots, 6$)?

1.5.3. [4 балла] Используя метод множителей Лагранжа, докажите, что величина

$$f^*(x) = g(x)e^{\sum_i^\infty \lambda_i r_i(x) + \lambda_0}$$

минимизирует относительную энтропию $H(f||g)$ (1.42) по всем функциям вероятности $f(x)$, $x \in \{1, 2, \ldots\}$ таким, что $\sum\limits_x f(x) r_i(x) = \alpha_i$, при $i = 1, 2, \ldots$ Данный результат является обобщением теоремы 1.8.

1.5.4. [3 балла] Пусть $Y = X + Z$, где случайная величина X имеет гауссовское распределение, а Z — независимая от X случайная величина с нулевым математическим ожиданием и дисперсией σ^2. Докажите, что $I(X;Y)$ минимально, если Z также имеет гауссовское распределение.

1.6.1. [2 балла] Пусть $\ldots, X_{-1}, X_0, X_1, \ldots$ — стационарный случайный процесс. Определите истинность или ложность следующих утверждений:

(а) $H(X_n|X_0) = H(X_{-n}|X_0)$,

(б) $H(X_n|X_0) \geq H(X_{n-1}|X_0)$,

(в) $H(X_n|X_1, \ldots, X_{n-1}, X_{n+1})$ не возрастает с ростом n.

1.6.2. [3 балла] Определим **дифференциальную энтропию** случайного процесса $\underline{X}$ с совместной плотностью вероятности $f(\underline{x})$ по аналогии с (1.33) и (1.57) как $h(\underline{X}) = -\int\limits_Q f(\underline{x}) \log f(\underline{x}) d\underline{x}$, где Q — ненулевое множество $\underline{X}$. Докажите, что

(а) $h(\underline{X}A) = h(\underline{X}) + \log|\det A|$, где A — квадратная матрица соответствующего размера, а $\det A$ — ее определитель,

(б) любое преобразование координат с отличным от единицы якобианом приведет к изменению дифференциальной энтропии.

1.6.3. [4 балла] Пусть $X_1, X_2, \ldots, X_n$ — независимые, но не одинаково распределенные двоичные случайные величины и пусть

$$P(X_t = 1) = \begin{cases} 0{,}5, & \text{если } 2k < \log\log t \leq 2k+1, \\ 0, & \text{если } 2k+1 < \log\log t \leq 2k+2, \end{cases}$$

при $k = 0, 1, \ldots$ Докажите, что в этом случае предел (1.68) не существует.

1.7.1. [1 балл] Оцените вероятность появления не более k единиц в двоичной последовательности длины n, порождаемой эргодическим случайным процессом.

1.7.2. [4 балла] Докажите, что стационарный случайный процесс без памяти является эргодическим.

1.7.3. [3 балла] Пусть мы бросаем кубик и в зависимости от выпавшей i-й грани ($i = 1, 2, \ldots, 6$) выбираем параметр $p = 1/i$. Затем с вероятностью p мы выбираем в качестве символа случайного процесса 1, а 0 — с вероятностью $1 - p$. Докажите, что таким образом мы получаем стационарный, но не эргодический случайный процесс. Найдите его абсолютную энтропию.

1.7.4. [2 балла] Какова средняя длина слова в марковской аппроксимации первого порядка для английского языка?

1.7.5. [2 балла] Оцените количество различных имеющих смысл фраз длины n в английском и русском разговорном языке.

1.7.6. [3 балла] Рассмотрим модель некоторого языка, в котором все гласные заменены одним символом, а все согласные — другим. Докажите, что если данный язык представляет собой марковский процесс, то и наша модель также является марковским процессом. Что можно сказать о порядках этих процессов?

1.7.7. [2 балла] Докажите, что $H(X_0|X_n)$ не убывает с возрастанием n для любой цепи Маркова.

1.8.1. [2 балла] Найдите колмогоровскую сложность чисел e, $\sqrt{2}$, $(100!)!$.

1.8.2. [3 балла] Найдите верхнюю границу колмогоровской сложности произвольного натурального числа.

1.8.3. [2 балла] Найдите верхнюю границу колмогоровской сложности изображения, состоящего из n точек, если известно, что данное изображение может быть сжато в три раза.

1.8.4. [3 балла] Докажите, что граница в (1.89) может быть улучшена до

$$\mathcal{K}(\underline{x}) \leq \mathcal{K}(\underline{x}|n) + \log^* n + c,$$

где $\log^* n = \log n + \log\log n + \log\log\log n + \ldots$

1.8.5. [3 балла] Докажите, что количество последовательностей $\underline{x}$, имеющих сложность $\mathcal{K}(\underline{x}) < k$, менее чем 2^k.

1.8.6. [3 балла] Докажите, что если символы последовательности в теореме 1.15 имеют алфавит с мощностью L, то верхняя граница в этой теореме может быть задана как $H(X) + \frac{L \log n}{n} + \frac{c}{n}$.

Глава 2

Фундаментальные теоремы кодирования

2.1. Типичные последовательности

В теореме 1.11 Шеннона – МакМиллана мы показали, что для эргодического случайного процесса близка к единице вероятность того, что величина $-\frac{1}{n}\log f(\underline{x})$ совпадает с абсолютной энтропией этого процесса. В доказательстве этой теоремы мы существенным образом опирались на свойства стационарности и эргодичности.

Данный результат справедлив и для стационарных случайных процессов без памяти, поскольку эти процессы также являются эргодическими (см. задачу 1.7.2). Однако для них подобное свойство доказывается значительно проще, так как является прямым следствием закона больших чисел.

Напомним, что закон больших чисел в теории вероятностей утверждает, что среднее арифметическое значение $\frac{1}{n}\sum\limits_{t=1}^{n} u_t$ н.о.р. случайных величин при возрастании n перестает зависеть от конкретных особенностей отдельных случайных величин и приближается к математическому ожиданию $\overline{U}$ каждой из них. Случайные отклонения от среднего, неизбежные для каждой из случайных величин, в своей массе взаимно гасятся, так что их среднее становится устойчивым к отдельным отклонениям.

Более точно закон больших чисел формулируется так: для любого $\delta > 0$ и заданного $\varepsilon > 0$ существует такое значение n_0, что для любого $n > n_0$ справедливо

$$P\left(\left|\frac{1}{n}\sum_{t=1}^{n} u_t - \overline{U}\right| \le \varepsilon\right) \ge 1 - \delta, \tag{2.1}$$

или же

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} u_t = \overline{U}. \tag{2.2}$$

Поскольку стационарный процесс без памяти состоит из н.о.р. случайных величин (1.66), ясно, что

$$f(\underline{u}) = \prod_{t=1}^{n} f(u_t). \tag{2.3}$$

Теорема 2.1. *Для стационарного случайного процесса без памяти* $\underline{U}$ *справедливо*

$$\operatorname*{plim}_{n\to\infty} -\frac{1}{n} \log f(\underline{u}) = H(U), \tag{2.4}$$

где $H(U)$ *является энтропией каждой из случайных величин, входящих в* $\underline{U}$.

Доказательство. Функции от независимых случайных величин также являются независимыми случайными величинами. Так как U_t $(t = 1, 2, \ldots, n)$ независимы и одинаково распределены, то и $G(u_t) = \log f(u_t)$ независимы и одинаково распределены. Поэтому в силу (2.3) и закона больших чисел (2.2)

$$\operatorname*{plim}_{n\to\infty} -\frac{1}{n} \log f(\underline{u}) = \operatorname*{plim}_{n\to\infty} -\frac{1}{n} \sum_{t=1}^{n} \log f(u_t) = \overline{-\log f(U)} = H(U). \qquad \square$$

Свойство (2.4) и аналогичное ему свойство (1.11) для эргодических случайных процессов называется свойством **энтропийной устойчивости** случайного процесса $\underline{U}$. Оно играет фундаментальную роль в доказательстве всех классических теорем кодирования.

В форме (2.1) выражение (2.4) выглядит так:

$$P\left(\left|-\frac{1}{n} \log f(\underline{u}) - H(U)\right| \le \varepsilon\right) \ge 1 - \delta. \tag{2.5}$$

Определение 2.1. *Назовем последовательность* $\underline{u} = \{u_1, u_2, \ldots, u_n\}$ **типичной**, *если она обладает свойством энтропийной устойчивости (2.4), а множество* $\mathcal{T}_\varepsilon(U)$ *всех типичных последовательностей определим как*

$$\mathcal{T}_\varepsilon(U) = \left\{ \underline{u} \in \mathcal{A}_U^n \;\middle|\; \left|-\frac{1}{n} \log f(\underline{u}) - H(U)\right| \le \varepsilon \right\}. \tag{2.6}$$

Теорема 2.2. *Для каждого* $\varepsilon > 0$ *и* $\delta > 0$ *существует такое значение* n_0*, что для всех* $n > n_0$ *и типичных последовательностей* $\underline{u}$

$$P\left(\underline{u} \in \mathcal{T}_\varepsilon(U)\right) \geq 1 - \delta, \tag{2.7}$$

$$2^{-n(H(U)+\varepsilon)} \leq f(\underline{u}) \leq 2^{-n(H(U)-\varepsilon)}, \tag{2.8}$$

$$(1-\delta)2^{n(H(U)-\varepsilon)} \leq \operatorname{card} \mathcal{T}_\varepsilon(U) \leq 2^{n(H(U)+\varepsilon)}. \tag{2.9}$$

Доказательство. Неравенство (2.7) следует из (2.5) и (2.6), то есть множество типичных последовательностей несет в себе, по существу, всю вероятность (для достаточно больших n).

Неравенства (2.8) следуют непосредственно из определения (2.6) в предположении двоичного логарифма. Они означают, что все типичные последовательности имеют примерно одинаковую вероятность $2^{-nH(U)}$.

Для доказательства правой части в (2.9) запишем, учитывая левую часть в (2.8),

$$\begin{aligned} 1 = \sum_{\underline{u} \in \mathcal{A}_U^n} f(\underline{u}) \geq \sum_{\underline{u} \in \mathcal{T}_\varepsilon(U)} f(\underline{u}) \quad &\geq \quad \sum_{\underline{u} \in \mathcal{T}_\varepsilon(U)} 2^{-n(H(U)+\varepsilon)} = \\ &= \quad \operatorname{card} \mathcal{T}_\varepsilon(U) 2^{-n(H(U)+\varepsilon)}, \end{aligned}$$

то есть $\operatorname{card} \mathcal{T}_\varepsilon(U) \leq 2^{n(H(U)+\varepsilon)}$.

Для доказательства левой части в (2.9), учитывая (2.7) и правую часть в (2.8), имеем

$$1 - \delta \leq \sum_{\underline{u} \in \mathcal{T}_\varepsilon(U)} f(\underline{u}) \leq \sum_{\underline{u} \in \mathcal{T}_\varepsilon(U)} 2^{-n(H(U)-\varepsilon)} = \operatorname{card} \mathcal{T}_\varepsilon(U) 2^{-n(H(U)-\varepsilon)},$$

то есть $\operatorname{card} \mathcal{T}_\varepsilon(U) \geq (1-\delta)2^{n(H(U)-\varepsilon)}$. □

Пример 2.1. Пусть случайная величина U принимает значения из множества $\{0, 1\}$, а ее функция вероятности $f(u)$ определяется как $f(1) = p \leq 0,5$ и $f(0) = 1 - p$. Энтропия U является двоичной энтропией (1.18)

$$H(U) = \eta(p) = -p \log_2 p - (1-p) \log_2 (1-p).$$

Пусть теперь $\underline{U} = \{U_1, U_2, \ldots, U_n\}$ — случайный процесс, состоящий из н.о.р. случайных величин, каждая с функцией вероятности $f(u)$. Вероятность любой реализации такого случайного процесса равна $f(\underline{u}) = p^t (1-p)^{n-t}$, где t — количество единиц в ней ($t = 1, 2, \ldots, n$). То есть

$$-\frac{1}{n} \log_2 f(\underline{u}) = -\frac{t}{n} \log_2 p - \left(1 - \frac{t}{n}\right) \log_2 (1-p).$$

Определим, какие реализации данного процесса будут относиться к множеству типичных:

$$\begin{aligned}
\mathcal{T}_\varepsilon(U) &= \left\{ \underline{u} \;\middle|\; \left| -\frac{1}{n}\log_2 f(\underline{u}) - H(U) \right| \le \varepsilon \right\} = \\
&= \left\{ \underline{u} \;\middle|\; \left| \left(p - \frac{t}{n}\right)\log_2 p + \left(\frac{t}{n} - p\right)\log_2(1-p) \right| \le \varepsilon \right\} = \\
&= \left\{ \underline{u} \;\middle|\; \left| \left(\frac{t}{n} - p\right)\log_2 \frac{1-p}{p} \right| \le \varepsilon \right\} = \\
&= \left\{ \underline{u} \;\middle|\; \left| \frac{t}{n} - p \right| \le \varepsilon_1 \right\},
\end{aligned}$$

где $\varepsilon_1 = \varepsilon / \log_2((1-p)/p)$. Поэтому множество типичных последовательностей состоит в данном случае из последовательностей, число единиц в которых близко к np.

Например, если $p = 0{,}1$, то множество типичных последовательностей состоит из последовательностей с количеством единиц, близким к $n/10$. Количество последовательностей в этом множестве близко к $2^{n\eta(0,1)} = 2^{n0{,}47}$, а доля множества типичных последовательностей среди всех последовательностей близка к $2^{n\eta(0,1)}/2^n = 2^{-n(1-\eta(0,1))} = 2^{-n0{,}53}$. При $n = 100$ эта доля составляет примерно 10^{-16} и тем не менее суммарная вероятность типичных последовательностей близка к единице!

Если же $p = 0{,}5$, то $\eta(0{,}5) = 1$, вероятность любой последовательности длины n не зависит от последовательности и равна 2^{-n}, а множество типичных последовательностей совпадает с множеством всех последовательностей.

2.2. Сжатие данных и избыточность

Теперь мы хотим закодировать случайный процесс $\underline{U}$ так, чтобы, во-первых, было использовано как можно меньше бит для его представления, а, во-вторых, по этим битам можно было полностью восстановить исходный процесс.

Назовем **скоростью кодирования** R среднее количество бит, используемых для кодирования одного символа случайного процесса. Использование именно двоичных символов для кодирования диктуется требованием удобства реализации практических алгоритмов. Так же как и выбор базиса логарифма в определениях информации, эн-

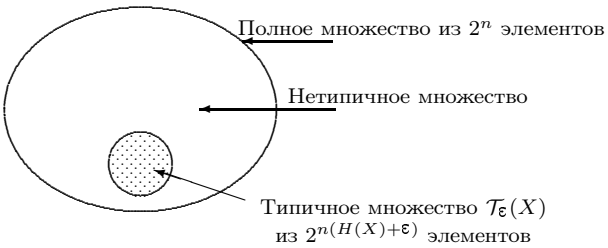


Рис. 2.1. Типичное и полное множество

тропии или множества типичных последовательностей, данный выбор не оказывает влияния на теоретические выкладки.

Разделим полное множество реализаций случайного процесса $\underline{U}$ на два множества: типичное и комплементарное (рис. 2.1).

Упорядочим все элементы в каждом множестве (например, в лексикографическом порядке). Мы можем представить каждую последовательность в $\mathcal{T}_\varepsilon(U)$, указав ее номер в множестве. Так как мы имеем не более $2^{n(H(U)+\varepsilon)}$ типичных последовательностей, то нам для этого потребуется не более чем $n(H(U)+\varepsilon)$ бит. А так как сумма вероятностей всех типичных последовательностей стремится к единице с возрастанием n, причем вероятность возникновения нетипичной последовательности стремится к нулю, то и для представления полного множества достаточно всего лишь $n(H(U)+\varepsilon)$ бит, по которым возможно восстановить как множество типичных последовательностей, так и полное множество. Если же мы применим меньше чем $n(H(U)-\varepsilon)$ бит (например, $n(H(U)-\varepsilon)-1$), то их будет достаточно только для представления ничтожно малой части типичных последовательностей (см. левое неравенство в (2.9)) и восстановление даже множества $\mathcal{T}_\varepsilon(U)$ невозможно.

Таким образом, мы доказали теорему **Шеннона о сжатии данных**, которую сформулируем в общем для всех удовлетворяющих свойству энтропийной устойчивости процессов виде.

Теорема 2.3. *Для того, чтобы вероятность полного восстановления обладающего свойством энтропийной устойчивости случайного процесса $\underline{U} = \{U_1, U_2, \ldots, U_n\}$ стремилась к единице, если $n \to \infty$, необходимо и достаточно, чтобы скорость кодирования R была больше, чем $H_\infty(\underline{U})$ бит на каждый символ U_t.*

Теорема 2.3 устанавливает теоретическую возможность сжатия данных и задает меру качества ($R > H_\infty(\underline{U})$ бит на символ) для практических алгоритмов кодирования.

В общем случае, когда алфавит каждой из случайных величин, входящих в случайный процесс $\underline{U}$, состоит из L символов, а информация Хартли равна $H_0 = \log_2 L$, для представления всего множества реализаций необходимо nH_0 бит, а для представления множества типичных последовательностей асимптотически достаточно $nH_\infty(\underline{U})$ бит. Поэтому говорят о возможности сжатия данных до (абсолютной) энтропии. При этом все сжатые последовательности имеют асимптотически одинаковую вероятность $2^{-nH_\infty(\underline{U})}$ (2.8), а их количество равно $2^{nH_\infty(\underline{U})}$ (2.9).

Поскольку исходную последовательность можно полностью восстановить по $nH_\infty(\underline{U})$ бит, то ее оставшиеся $n(H_0 - H_\infty(\underline{U}))$ бит являются избыточными. Исходя из этого избыточность определяется следующим образом.

Определение 2.2. *Величина*

$$\rho_n(\underline{U}) = H_0 - H_n(\underline{U}) \tag{2.10}$$

называется **избыточностью n-го порядка**, *величина*

$$\rho_\infty(\underline{U}) = \lim_{n\to\infty} \rho_n(\underline{U}), \tag{2.11}$$

если предел существует, называется **абсолютной избыточностью**, *а величина*

$$\mathsf{P}_\infty(\underline{U}) = 1 - \frac{H_\infty(\underline{U})}{H_0} \tag{2.12}$$

называется **относительной избыточностью** *случайного процесса* $\underline{U}$.

Для стационарных процессов в силу (1.69) предел в (2.11) всегда существует, при этом для процессов с двоичным алфавитом, когда $H_0 = 1$, $\rho_\infty(\underline{U}) = \mathsf{P}_\infty(\underline{U})$.

Пример 2.2. Из (1.88) имеем, что абсолютная энтропия английского языка равна $H_\infty(\underline{U}) \approx 1{,}34$ бит (на букву), а информация Хартли равна $H_0 = \log_2 27 \approx 4{,}75$ бит, значит, абсолютная избыточность есть $\rho_\infty(\underline{U}) = H_0 - H_\infty(\underline{U}) \approx 3{,}41$ бит, а относительная избыточность $\mathsf{P}_\infty(\underline{U}) \approx 0{,}72$.

Определение избыточности обосновывается тем, что максимальная информация, которую мог бы нести каждый символ случайного

процесса, равна H_0, что имело бы место, если бы его символы появлялись независимо и равновероятно. Избыточность такого случайного процесса равна нулю, как это, например, имеет место для случайного процесса из примера 2.1 при $p = 0{,}5$. При $p = 0{,}1$ мы имеем $\varrho_\infty(\underline{U}) = 1 - \eta(0{,}1) \approx 0{,}53$, то есть в этом случае более половины символов каждой реализации являются избыточными. Из примера 2.1 следует также, что не все последовательности одинаково хорошо сжимаемы: чем меньше энтропия последовательности, тем эффективнее возможно сжатие, то есть тем больше избыточности последовательность содержит.

На примере языка (1.88) мы также видим, что средняя энтропия символа значительно меньше H_0, а величина $\varrho_n(\underline{U})$ характеризует то, насколько реальный язык может быть сжат, если известны вероятностные характеристики его n-буквенных сочетаний. Но это не означает буквально то, что при $\mathsf{P}_\infty(\underline{U}) = 0{,}75$ три из четырех букв текста можно вычеркнуть без всяких потерь. Это значит, что, используя знания о статистических особенностях текста, его можно сжать до четверти длины и тем не менее иметь возможность восстановить его исходный вид. Некоторые методы такого сжатия рассмотрены в главе 3. Важную роль избыточность играет также и в криптографии (глава 4).

2.3. Совместно-типичные последовательности

До сих пор мы рассматривали типичные последовательности для одиночных случайных процессов. Для доказательства двух других теорем Шеннона нам потребуется обобщение понятия типичных последовательностей на пары случайных процессов. Так же как и в доказательстве теоремы о сжатии данных, ограничимся здесь для простоты рассмотрением стационарных процессов без памяти.

Пусть X и Y — случайные величины с функциями вероятности $f(x)$ и $f(y)$ и алфавитами $\mathcal{A}_X$ и $\mathcal{A}_Y$, а $Z = \{X, Y\}$ — случайная величина с функцией вероятности $f(z) = f(x, y)$ и алфавитом $\mathcal{A}_Z = \mathcal{A}_X \times \mathcal{A}_Y$.

Пусть далее $\underline{X} = \{X_1, X_2, \ldots, X_n\}$, $\underline{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ и $\underline{Z} = \{Z_1, Z_2, \ldots, Z_n\}$, где $Z_t = \{X_t, Y_t\}$ $(t = 1, 2, \ldots, n)$, являются стационарными случайными процессами без памяти. Для случайных процессов $\underline{X}$ и $\underline{Y}$ это означает, что справедливо (2.3), а для случайного процесса $\underline{Z}$ — что

$$f(\underline{x},\underline{y}) = \prod_{t=1}^{n} f(x_t, y_t). \tag{2.13}$$

Определение 2.3. *Назовем последовательность* $\{\underline{x},\underline{y}\}$*, удовлетворяющую (2.13),* **совместно-типичной**, *если каждая из последовательностей* $\underline{x}$, $\underline{y}$ *и* $\{\underline{x},\underline{y}\}$ *является типичной, то есть обладает свойством энтропийной устойчивости*

$$\operatorname*{plim}_{n\to\infty} -\frac{1}{n}\log f(\underline{x}) = H(X), \qquad \operatorname*{plim}_{n\to\infty} -\frac{1}{n}\log f(\underline{y}) = H(Y), \tag{2.14}$$

$$\operatorname*{plim}_{n\to\infty} -\frac{1}{n}\log f(\underline{x},\underline{y}) = H(X,Y), \tag{2.15}$$

а множество $\mathcal{T}_\varepsilon(X,Y)$ *совместно-типичных последовательностей* $\{\underline{x},\underline{y}\}$ *определим как*

$$\mathcal{T}_\varepsilon(X,Y) = \left\{ \{\underline{x},\underline{y}\} \in \mathcal{A}_X^n \times \mathcal{A}_Y^n \;\middle|\; \begin{array}{l} \left| -\frac{1}{n}\log f\left(\underline{x}\right) - H(X) \right| \le \varepsilon, \\ \left| -\frac{1}{n}\log f\left(\underline{y}\right) - H(Y) \right| \le \varepsilon, \\ \left| -\frac{1}{n}\log f\left(\underline{x},\underline{y}\right) - H(X,Y) \right| \le \varepsilon \end{array} \right\}.$$

По аналогии с «одномерным» случаем непосредственно из определения 2.3 получаем следующий результат.

Теорема 2.4. *Для каждого* $\varepsilon > 0$ *и* $\delta > 0$ *существует такое значение* n_0*, что для всех* $n > n_0$ *и совместно-типичных последовательностей* $\{\underline{x},\underline{y}\}$*:*

$$P\left(\{\underline{x},\underline{y}\} \in \mathcal{T}_\varepsilon(X,Y)\right) \ge 1-\delta,$$

$$2^{-n(H(X,Y)+\varepsilon)} \le f\left(\underline{x},\underline{y}\right) \le 2^{-n(H(X,Y)-\varepsilon)}, \tag{2.16}$$

$$(1-\delta)2^{n(H(X,Y)-\varepsilon)} \le \operatorname{card}\mathcal{T}_\varepsilon(X,Y) \le 2^{n(H(X,Y)+\varepsilon)}. \tag{2.17}$$

На рис. 2.2 все типичные последовательности показаны черными точками, а нетипичные — белыми.

Если X и Y статистически независимы ($I(X;Y)=0$), то черными точками заполнен весь левый верхний квадрат на рис. 2.2, то есть для

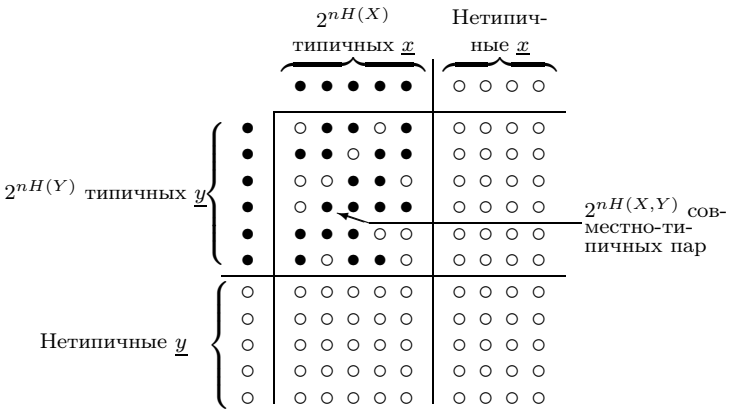


Рис. 2.2. Множество совместно-типичных последовательностей

этого случая, если $\underline{x}$ и $\underline{y}$ обе являются типичными, то они и совместно-типичны, поскольку из равенств (2.14) следует (2.15).

С другой стороны, если $H(X|Y) = 0$ (то есть Y однозначно определяет X), то черными точками заполнена диагональ верхнего левого квадрата на рис. 2.2.

Следствие 2.1. *Для всех* $\{\underline{x}, \underline{y}\} \in \mathcal{T}_\varepsilon(X, Y)$ *справедливо*

$$2^{-n(H(X|Y)+2\varepsilon)} \leq f\left(\underline{x}\,\middle|\,\underline{y}\right) \leq 2^{-n(H(X|Y)-2\varepsilon)}. \tag{2.18}$$

Доказательство. Так как $f\left(\underline{x}\,\middle|\,\underline{y}\right) = \frac{f(\underline{x},\underline{y})}{f(\underline{y})}$, доказательство следует из (2.16) и из того, что

$$2^{-n(H(Y)+\varepsilon)} \leq f\left(\underline{y}\right) \leq 2^{-n(H(Y)-\varepsilon)},$$

поскольку $\underline{y} \in \mathcal{T}_\varepsilon(Y)$. □

Попробуем теперь оценить количество типичных последовательностей $\underline{x}$, если задана последовательность $\underline{y}$, такая, что пара $\{\underline{x}, \underline{y}\}$ является совместно-типичной. Пусть $\mathcal{T}_\varepsilon\left(X\,\middle|\,\underline{y}\right) = \left\{\underline{x}\,\middle|\,(\underline{x}, \underline{y}) \in \mathcal{T}_\varepsilon(X, Y), \underline{y}\right\}$.

Теорема 2.5. *Для любого* $\varepsilon > 0$

$$\operatorname{card} \mathcal{T}_\varepsilon\left(X\,\middle|\,\underline{y}\right) \leq 2^{n(H(X|Y)+2\varepsilon)}. \tag{2.19}$$

Доказательство. Если $\underline{y} \notin \mathcal{T}_\varepsilon(Y)$ (то есть $\underline{y}$ сама не является типичной), то $\operatorname{card} \mathcal{T}_\varepsilon\left(X\,\middle|\,\underline{y}\right) = 0$ и теорема тривиальна. Поэтому будем пред-

полагать, что $y \in \mathcal{T}_\varepsilon(Y)$. Тогда, учитывая левую часть в (2.18), запишем

$$\begin{aligned} 1 &= \sum_{\underline{x}\in\mathcal{A}_X^n} f\left(\underline{x}\,|\underline{y}\right) \geq \sum_{\underline{x}\in\mathcal{T}_\varepsilon\left(X|\underline{y}\right)} f\left(\underline{x}\,|\underline{y}\right) \geq \\ &\geq \sum_{\underline{x}\in\mathcal{T}_\varepsilon\left(X|\underline{y}\right)} 2^{-n(H(X|Y)+2\varepsilon)} = \operatorname{card}\mathcal{T}_\varepsilon\left(X\,|\underline{y}\right) 2^{-n(H(X|Y)+2\varepsilon)}, \end{aligned}$$

откуда и следует (2.19). □

Аналогично можно доказать, что $\operatorname{card}\mathcal{T}_\varepsilon\left(Y\,|\underline{x}\right) \leq 2^{n(H(Y|X)+2\varepsilon)}$, поэтому имеется максимум $2^{n(H(Y|X)+2\varepsilon)}$ черных точек в каждом столбце и максимум $2^{n(H(X|Y)+2\varepsilon)}$ черных точек в каждой строке на рис. 2.2.

Заметим, что для множества $\mathcal{T}_\varepsilon\left(X\,|\underline{y}\right)$ условно типичных последовательностей не существует нижней границы для $\operatorname{card}\mathcal{T}_\varepsilon\left(X\,|\underline{y}\right)$, доказываемой по аналогии с нижней границей для $\operatorname{card}\mathcal{T}_\varepsilon(X)$ в теореме 2.2 (см. задачу 2.3.1).

Теорема 2.6. *Если некоторая последовательность $\underline{x}$ выбрана случайно в соответствии с функцией вероятности $f(\underline{x})$, а последовательность $\underline{y}$ — в соответствии с $f(\underline{y})$, то вероятность того, что пара $\left\{\underline{x},\underline{y}\right\}$ совместно-типична, имеет границы*

$$(1-\delta)2^{-n(I(X;Y)+3\varepsilon)} \leq \sum_{\left\{\underline{x},\underline{y}\right\}\in\mathcal{T}_\varepsilon(X,Y)} f\left(\underline{x}\right) f\left(\underline{y}\right) \leq 2^{-n(I(X;Y)-3\varepsilon)}.$$

Доказательство. Из (2.8) получаем

$$\begin{aligned} &\operatorname{card}\mathcal{T}_\varepsilon(X,Y)2^{-n(H(X)+\varepsilon)}2^{-n(H(Y)+\varepsilon)} \leq \\ \leq &\sum_{\left\{\underline{x},\underline{y}\right\}\in\mathcal{T}_\varepsilon(X,Y)} f\left(\underline{x}\right) f\left(\underline{y}\right) \leq \\ \leq &\operatorname{card}\mathcal{T}_\varepsilon(X,Y)2^{-n(H(X)-\varepsilon)}2^{-n(H(Y)-\varepsilon)}. \end{aligned}$$

Используя далее верхнюю и нижнюю границы для $\operatorname{card}\mathcal{T}_\varepsilon(X,Y)$ (2.17) и равенство $I(X;Y) = H(X) + H(Y) - H(X,Y)$ (см. рис. 1.2), имеем результат теоремы. □

Таким образом, вероятность того, что при случайном выборе последовательностей $\underline{x}$ и $\underline{y}$ пара $\left\{\underline{x},\underline{y}\right\}$ окажется совместно-типичной, асимптотически (при $n \to \infty$) равна $2^{-nI(X;Y)}$.

2.4. Лемма об обработке данных и лемма Фано

В этом разделе мы рассмотрим два важных вспомогательных результата, применяемых в доказательствах теорем Шеннона.

Рассмотрим ситуацию, когда данные проходят через два последовательно соединенных обработчика (рис. 2.3). Эти обработчики мо-

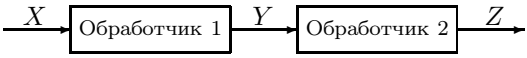


Рис. 2.3. Иллюстрация к лемме об обработке данных

гут быть любыми, даже статистическими машинами, но единственное ограничение состоит в том, что X, Y и Z образуют цепь Маркова $X \to Y \to Z$, то есть X оказывает влияние на Z только косвенно — через Y, и при заданном Y величины X и Z считаются независимыми. Лемма **об обработке данных** утверждает следующее.

Лемма 2.1. *Любые обработчики могут только «потерять» информацию, но не добавить ее*

$$I(X;Z) \leq \left\{ \begin{array}{c} I(X;Y) \\ I(Y;Z) \end{array} \right. . \tag{2.20}$$

Доказательство. Используя диаграмму Эйлера (см. рис. 1.2), имеем

$$\begin{aligned} I(X;Z) &= H(X) - H(X|Z) \leq \\ &\leq H(X) - H(X|Z,Y) = \\ &= H(X) - H(X|Y) = I(X;Y), \end{aligned}$$

где неравенство следует из того, что $H(X|Z) \geq H(X|Z,Y)$, а следующее за ним равенство — из того, что X и Z независимы. Аналогично

$$\begin{aligned} I(X;Z) &= H(Z) - H(Z|X) \leq \\ &\leq H(Z) - H(Z|X,Y) = \\ &= H(Z) - H(Z|Y) = I(Y;Z). \end{aligned}$$

□

Лемму об обработке данных в теории информации можно в некотором смысле сравнить с законом сохранения энергии в физике. Так же как с точки зрения закона сохранения энергии бессмысленно говорить, что, например, электростанция производит электроэнергию, так и с точки зрения леммы об обработке данных не имеет смысла выражение «компьютер производит информацию». В обоих случаях речь может идти только о преобразовании энергии или информации.

Предположим теперь, что нам надо оценить случайную величину X, если нам известна статистически зависимая от нее случайная величина Y.

Обозначим **вероятность ошибки** такой оценки через $P_e = P(X \neq Y)$. Из (1.26) ясно, что вероятность ошибки будет равна нулю, если между X и Y нет статистической связи, то есть $H(X|Y) = 0$. Интуитивно можно ожидать, что вероятность ошибки будет малой, если и энтропия $H(X|Y)$ достаточно мала. Лемма **Фано** устанавливает количественную связь между вероятностью ошибки и условной энтропией.

Лемма 2.2. *Пусть случайные величины X и Y принимают значения из множества $\{x_1, x_2, \ldots, x_L\}$, тогда*

$$H(X|Y) \leq \eta(P_e) + P_e \log_2(L-1). \tag{2.21}$$

Доказательство. Введем флаг ошибки

$$E = \begin{cases} 0, & \text{если } X = Y, \\ 1, & \text{если } X \neq Y. \end{cases}$$

Очевидно, что $H(E) = \eta(P_e)$ (так как E — двоичная переменная). Из диаграммы на рис. 1.2 мы наблюдаем, что $H(X,E) = H(X) + H(E|X)$, или по аналогии для условных энтропий

$$H(X,E|Y) = H(X|Y) + H(E|X,Y) = H(X|Y),$$

так как X и Y однозначно определяют E.

То есть мы имеем

$$\begin{aligned} H(X|Y) &= H(X,E|Y) = \\ &= H(E|Y) + H(X|Y,E) \leq \\ &\leq H(E) + H(X|Y,E). \end{aligned} \tag{2.22}$$

Очевидно, что $H(X|Y,E=0) = 0$, так как в данном случае X определено однозначно ($X = Y$); очевидно и то, что $H(X|Y,E=1) \leq \leq \log_2(L-1)$, так как для каждого значения Y существует только $L-1$ значений X, если $E = 1$. То есть

$$\begin{aligned} H(X|Y,E) &= P(E=1)H(X|Y,E=1) + \\ &\quad + P(E=0)H(X|Y,E=0) \leq \\ &\leq P(E=1)\log_2(L-1) = P_e \log_2(L-1), \end{aligned} \tag{2.23}$$

и поэтому результат леммы следует из (2.22) и (2.23). □

Лемму Фано можно эвристически пояснить следующим образом. Пусть нам известно $H(X|Y)$ и мы хотим определить X. Проверим сначала, выполняется ли равенство $X = Y$, и если да, то мы X определили. Чтобы в этой ситуации определить, справедливо ли $X = Y$, необходимо $\eta(P_e)$ бит. Если же $X \neq Y$, то остается еще $L-1$ возможностей определить X, а если $X \neq Y$ (что возникает с вероятностью P_e), необходимо еще максимум $\log_2(L-1)$ бит.

Равенство в лемме Фано достигается тогда, когда все $(L-1)$ «неверных» значений равновероятны (при условии, что ошибка возникла, то есть $X \neq Y$).

График функции $\eta(P_e) + P_e \log_2(L-1)$ для $0 \leq P_e \leq 1$ изображен на рис. 2.4. Мы видим, что если $P_e = 0$, то и $H(X|Y) = 0$, как мы интуитивно и предполагали.

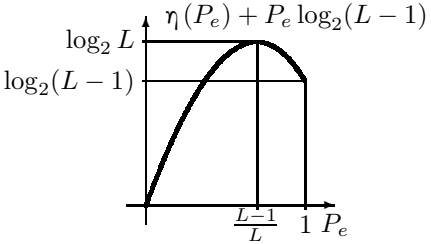


Рис. 2.4. График функции в лемме Фано

Если энтропия $H(X|Y) > 0$, то из леммы Фано следует положительная нижняя граница для P_e.

Пример 2.3. Пусть X, Y — двоичные случайные величины и пусть $H(X|Y) = 0{,}5$, тогда из леммы Фано: $0{,}5 \leq \eta(P_e)$ или $0{,}110 \leq P_e \leq 0{,}810$. Если же X, Y — троичные переменные и опять $H(X|Y) = 0{,}5$, то $0{,}5 \leq \eta(P_e) + P_e$, откуда следует, что $0{,}084 \leq P_e \leq 1$.

Заметим, что так как функция двоичной энтропии (1.18) ограничена сверху единицей (см. рис. 1.1), из (2.21) следует более слабая, однако часто достаточная верхняя граница для условной энтропии

$$H(X|Y) \leq 1 + P_e \log_2 L. \tag{2.24}$$

Лемма об обработке данных и лемма Фано могут быть без труда обобщены на случайные процессы $\underline{X}$, $\underline{Y}$ и $\underline{Z}$, причем вид лемм при этом сохранится, а мощность алфавита L в лемме Фано будет соответствовать общему количеству последовательностей $\underline{x}$.

2.5. Теорема о передаче данных

Дискретным каналом связи называется изображенная на рис. 2.5 совместно с кодером и декодером канала совокупность

$$\left(\mathcal{A}_X^n, f(\underline{y}|\underline{x}), \mathcal{A}_Y^n\right), \tag{2.25}$$

состоящая из множества $\mathcal{A}_X^n$ входных последовательностей $\underline{X}$ длины n, множества $\mathcal{A}_Y^n$ входных последовательностей $\underline{Y}$ длины n и условной функции вероятности $f(\underline{y}|\underline{x})$. Это означает, что если была послана последовательность $\underline{x}$, то принятая последовательность $\underline{y}$ выбирается случайно в соответствии с $f(\underline{y}|\underline{x})$.

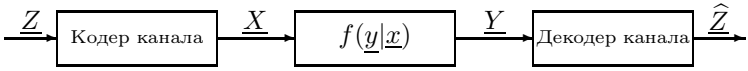


Рис. 2.5. К теореме о передаче данных

Дискретный канал связи называется **каналом без памяти**, если была послана последовательность $\underline{x} = \{x_1, x_2, \ldots, x_n\}$, а принятая последовательность $\underline{y} = \{y_1, y_2, \ldots, y_n\}$ выбирается случайно в соответствии с

$$f(\underline{y}|\underline{x}) = \prod_{t=1}^{n} f\left(y_t\,|x_t\right). \tag{2.26}$$

Это означает, что каждый символ y_t на выходе канала зависит только от соответствующего ему символа x_t на входе, а не от предыдущих символов $x_{t'}$ и $y_{t'}$, $t' < t$, и не от времени t.[1] При этом функция распределения $f(\underline{y}, \underline{x})$ удовлетворяет (2.13) (см. (1.2)), а совокупность (2.25) может быть записана в виде $(\mathcal{A}_X, f(y|x), \mathcal{A}_Y)$.

Стоит заметить, что $\operatorname{card}\mathcal{A}_X$ и $\operatorname{card}\mathcal{A}_Y$ необязательно одинаковы. Часто мощность алфавита передаваемых последовательностей равна двум, а принимаемых с выхода канала — больше двух (случай так называемых ≪**мягких**≫ решений демодулятора). Без существенного ограничения общности мы в дальнейшем ограничимся рассмотрением случая $\operatorname{card}\mathcal{A}_X = 2$.

[1] Заметим, что зависимость выходных символов канала связи от предыдущих входных символов означала бы прямое наличие памяти канала, зависимость от предыдущих выходных символов означала бы наличие в канале связи обратной связи, а зависимость от времени — динамическое изменение параметров канала.

Предположим, кодером источника было произведено сжатие до энтропии последовательности $\underline{u}$ из источника, то есть, как следует из раздела 2.1, все последовательности $\underline{z}$ равновероятны.

Организуем передачу последовательности $\underline{z}$ следующим образом. Пусть кодер делит ее на последовательности по k бит в каждой. То есть имеется всего $K = 2^k$ различных последовательностей длины k, каждая из которых по некоторому правилу, называемому **кодированием канала**, ставится в соответствие **кодовому слову** $\underline{x}$ длины n бит. Число n называется **длиной блока**.

Поскольку общее количество двоичных последовательностей длины n равно $L = 2^n$, каждая такая последовательность (при условии их равновероятности) могла бы нести $H_0 = \log_2 L$ бит информации. Однако общее количество наших исходных равновероятных последовательностей равно $K = 2^k$, поэтому относительная избыточность (2.12) случайного процесса $\underline{X}$, описывающего кодовые слова, равна

$$\mathsf{P}_\infty(\underline{X}) = 1 - \frac{H_\infty(\underline{X})}{H_0} = 1 - \frac{\log_2 K}{\log_2 L}.$$

Тесно связанная с относительной избыточностью и также используемая как мера избыточности величина

$$R = \frac{\log_2 K}{\log_2 L} \tag{2.27}$$

называется **скоростью кода**. Для двоичного случая скорость $R = \frac{k}{n}$ задает, сколько бит последовательности $\underline{z}$ передается в канал связи в единицу времени. Понятно, что $0 \le R \le 1$, причем случай $R = 0$ соответствует «бесконечной избыточности» (в канал последовательность $\underline{z}$ не передается), а случай $R = 1$ соответствует безызбыточной передаче данных.

Заметим также, что кодирование канала должно быть взаимнооднозначным преобразованием $\underline{Z}$ в $\underline{X}$, поэтому

$$H(\underline{Z}) = H(\underline{X}) = -\log_2 \frac{1}{K} = \log_2 2^{nR} = nR. \tag{2.28}$$

Определение 2.4. $\big(2^{nR}, n\big)$**-код канала связи** *состоит из множества* 2^{nR} *кодовых слов длины* n*, функции кодирования* $\mathsf{K}(\underline{Z}) = \underline{X}$ *и функции декодирования* $\mathsf{D}(\underline{Y}) = \widehat{Z}$.

По принимаемой последовательности $\underline{y}$, которая из-за влияния канала связи оказывается статистически связанной с переданной последовательностью $\underline{x}$, декодер делает оценку переданной последовательности. Определим **среднюю вероятность ошибки декодиро-**

вания как

$$\overline{P_e^{(n)}} = \frac{1}{2^{nR}} \sum_{\underline{x}} P(\underline{z} \neq \widehat{\underline{z}}). \tag{2.29}$$

Определение 2.5. *Скорость R $\left(2^{nR}, n\right)$-кода называется* **достижимой** *дискретным каналом без памяти, если для каждого $\varepsilon > 0$ найдется такое n_0, что для этого канала существует последовательность $\left(2^{nR}, n\right)$-кодов с длинами блоков $n > n_0$ и с $\overline{P_e^{(n)}} < \varepsilon$. Супремум*[2] *всех достижимых скоростей называется* **пропускной способностью** *C дискретного канала без памяти.*

Другими словами, пропускная способность определяет наименьшую избыточность, при которой возможна надежная передача данных. Следующая теорема называется **теоремой Шеннона о передаче данных**.

Теорема 2.7. *Пропускная способность дискретного канала без памяти $(\mathcal{A}_X, f(y|x), \mathcal{A}_Y)$ задается выражением*

$$C = \sup_{f(x)} I(X;Y). \tag{2.30}$$

Доказательство. Для того чтобы показать, что C из (2.30) является пропускной способностью канала связи (то есть супремумом всех достижимых скоростей), мы должны доказать, что:

1) каждая скорость $R < C$ достижима, то есть существует такой $\left(2^{nR}, n\right)$-код, что $\overline{P_e^{(n)}} \to 0$, если $n \to \infty$;

2) верно обратное утверждение: если $R > C$ (2.30), то существует некоторое $\delta > 0$, такое, что $\overline{P_e^{(n)}} \geq \delta$.

Начнем с доказательства обратного утверждения, то есть что невозможно передать информацию надежно (с любой сколь угодно

[2] **Супремум**, или точная верхняя грань $\sup\limits_{\mathcal{A}} G(x)$, где $x \in \mathcal{A}$, а $G(x)$ — некоторая функция на множестве $\mathcal{A}$, есть наименьшее число G_0, такое, что $G_0 \geq G(x)$ для каждого $x \in \mathcal{A}$. Если в множестве $\mathcal{A}$ существует такой элемент x_0, для которого $G_0 = G(x_0)$, то говорят, что верхняя грань достигается на $\mathcal{A}$. Тогда $G(x_0) = \sup\limits_{\mathcal{A}} G(x) = \max\limits_{\mathcal{A}} G(x)$. Если $\mathcal{A}$ — конечное множество, то точная верхняя грань всегда достигается и всегда $\sup\limits_{\mathcal{A}} G(x) = \max\limits_{\mathcal{A}} G(x)$. Если $\mathcal{A}$ — бесконечное множество, то точная верхняя грань может не достигаться ни на одном элементе из $\mathcal{A}$. Например, если $\mathcal{A}$ — множество натуральных чисел и $G(x) = 1 - 1/x$, то $\sup\limits_{\mathcal{A}} G(x) = 1$, но $G(x) \neq 1$ ни для одного элемента из $\mathcal{A}$.

малой вероятностью ошибки) через дискретный канал без памяти со скоростью, большей его пропускной способности.

Так как общее количество кодовых слов равно 2^{nR}, из леммы Фано (2.24) следует, что

$$H(\underline{Z}|\widehat{\underline{Z}}) \leq 1 + nR\overline{P_e^{(n)}}. \tag{2.31}$$

С другой стороны,

$$I(\underline{Z};\widehat{\underline{Z}}) = H(\underline{Z}) - H(\underline{Z}|\widehat{\underline{Z}}). \tag{2.32}$$

Так как $\underline{Z} \to \underline{X} \to \underline{Y} \to \widehat{\underline{Z}}$ является цепью Маркова, то по лемме об обработке данных (2.20)

$$I(\underline{Z};\widehat{\underline{Z}}) \leq I(\underline{X};\underline{Y}), \tag{2.33}$$

а, в свою очередь, для $I(\underline{X};\underline{Y})$, используя свойство иерархической аддитивности (1.59) и границу независимости энтропии (1.61), для канала связи без памяти (то есть когда Y_t зависит только от X_t) получим

$$\begin{aligned} I(\underline{X};\underline{Y}) &= H(\underline{Y}) - H(\underline{Y}|\underline{X}) = \\ &= \sum_{t=1}^{n} H(Y_t|Y_{t-1},\ldots,Y_1) - \sum_{t=1}^{n} H(Y_t|Y_{t-1},\ldots,Y_1,\underline{X}) = \\ &= \sum_{t=1}^{n} H(Y_t) - \sum_{t=1}^{n} H(Y_t|X_t) = \sum_{t=1}^{n} I(X_t;Y_t) \leq nC, \end{aligned} \tag{2.34}$$

где неравенство следует из (2.30).

Комбинируя (2.31) с (2.32), (2.33), (2.28) и (2.34), получаем

$$1 + nR\overline{P_e^{(n)}} \geq H(\underline{Z}|\widehat{\underline{Z}}) \geq H(\underline{Z}) - I(\underline{X};\underline{Y}) \geq nR - nC,$$

что эквивалентно

$$\overline{P_e^{(n)}} \geq 1 - \frac{C}{R} - \frac{1}{nR}. \tag{2.35}$$

Неравенство (2.35) показывает, что при $R > C$ вероятность ошибки ограничена снизу положительной величиной при достаточно больших n (а значит, и для всех n, так как если $\overline{P_e^{(n)}} = 0$ для некоторого малого n, то мы могли бы построить код с большим n и $\overline{P_e^{(n)}} = 0$ простой конкатенацией малых кодов). Иллюстрация к неравенству (2.35) дана на рис. 2.6.

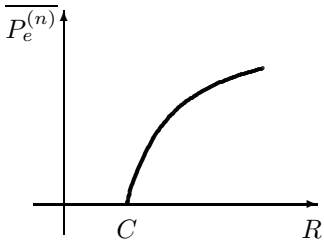


Рис. 2.6. Нижняя граница для $\overline{P_e^{(n)}}$

Если потребовать, что $\overline{P_e^{(n)}} = 0$, то из леммы Фано (2.21) следует, что $H(\underline{Z}|\widehat{\underline{Z}}) = 0$, поэтому $nR = H(\underline{Z}) = I(\underline{Z};\widehat{\underline{Z}}) \leq I(\underline{X};\underline{Y}) \leq nC$. Значит, для любого n и любого $\left(2^{nR}, n\right)$-кода, для которого ошибка декодирования недопустима вообще, должно выполняться $R \leq C$.

Теперь докажем асимптотическую достижимость (пункт 1).

Назовем функцию вероятности, при которой достигается супремум в (2.30), **согласованной с каналом связи**, и пусть $f^*(x)$ есть такая функция. Тогда, так как $\underline{x}$ состоит из н.о.р. случайных величин, то

$$f^*\left(\underline{x}\right) = \prod_{t=1}^{n} f^*\left(x_t\right). \tag{2.36}$$

Выберем в соответствии с $f^*\left(\underline{x}\right)$ случайно 2^{nR} кодовых слов $\underline{x}$ и назовем это множество кодовых слов **случайным** $\left(2^{nR}, n\right)$**-кодом**.

Построим правило декодирования так, что для каждой принятой последовательности $\underline{y}$ декодер выносит решение о том, что была послана последовательность $\widehat{\underline{x}}$, если

$$\begin{aligned} \left\{\widehat{\underline{x}}, \underline{y}\right\} &\in \mathcal{T}_\varepsilon(X,Y) \text{ и} \\ \left\{\underline{e}, \underline{y}\right\} &\notin \mathcal{T}_\varepsilon(X,Y) \text{ при любом } \underline{e} \neq \widehat{\underline{x}}, \end{aligned}$$

то есть последовательность $\widehat{\underline{x}}$ существует и однозначна. В противном случае, когда

$$\begin{aligned} \left\{\widehat{\underline{x}}, \underline{y}\right\} &\notin \mathcal{T}_\varepsilon(X,Y) \text{ или} \\ \left\{\underline{e}, \underline{y}\right\} &\in \mathcal{T}_\varepsilon(X,Y) \text{ хотя бы при одном } \underline{e} \neq \widehat{\underline{x}}, \end{aligned} \tag{2.37}$$

то есть когда последовательность $\widehat{\underline{x}}$ не существует или не однозначна, декодер совершает ошибку — выносит решение $\widehat{\underline{x}} \neq \underline{x}$ или, что эквива-

лентно, $\widehat{\underline{z}} \neq \underline{z}$. Докажем, что средняя вероятность ошибки (2.29) при таком декодировании может быть сделана сколь угодно малой.

Для некоторой последовательности $\underline{\alpha}$ рассмотрим два события, образующих при объединении полное множество:

$$E_{\underline{\alpha}} = \left\{\left\{\underline{\alpha}, \underline{y}\right\} \in \mathcal{T}_\varepsilon(X,Y)\right\} \text{ и } E^c_{\underline{\alpha}} = \left\{\left\{\underline{\alpha}, \underline{y}\right\} \notin \mathcal{T}_\varepsilon(X,Y)\right\}.$$

Предположим, мы решаем, что была послана последовательность $\widehat{\underline{x}}$. Тогда в соответствии с (2.37) событие, заключающееся в том, что произошла ошибка, есть

$$E^c_{\widehat{\underline{x}}} \bigcup \left(\bigcup_{\underline{e} \neq \widehat{\underline{x}}} E_{\underline{e}} \right). \tag{2.38}$$

Для определения вероятности этого события используем известную из теории вероятностей границу объединения

$$P\left(\bigcup_i E_i\right) \leq \sum_i P\left(E_i\right) \tag{2.39}$$

и получим

$$\overline{P_e^{(n)}} = \frac{1}{2^{nR}} \, 2^{nR} P\left(E^c_{\widehat{\underline{x}}} \bigcup \left(\bigcup_{\underline{e} \neq \widehat{\underline{x}}} E_{\underline{e}} \right)\right) \leq P\left(E^c_{\widehat{\underline{x}}}\right) + \sum_{\underline{e} \neq \widehat{\underline{x}}} P\left(E_{\underline{e}}\right), \tag{2.40}$$

где равенство следует из того, что событие (2.38) имеет одинаковую вероятность для каждой последовательности $\widehat{\underline{x}}$.

Первое слагаемое в (2.40) может быть оценено, если мы заметим, что пара $\left\{\widehat{\underline{x}}, \underline{y}\right\}$ выбиралась как совместно-типичная в соответствии с распределением $f(\underline{x}, \underline{y})$, поэтому по теореме 2.4

$$P\left(E^c_{\widehat{\underline{x}}}\right) \to 0, \text{ если } n \to \infty.$$

Для $\underline{e} \neq \widehat{\underline{x}}$ вероятность того, что пара $\left\{\underline{e}, \underline{y}\right\}$ является совместно-типичной, задается распределением $f(\underline{x})f(\underline{y})$, поэтому в соответствии с теоремой 2.6

$$P\left(E_{\underline{e}}\right) \leq 2^{-n(I(X,Y)-3\varepsilon)} = 2^{-n(C-3\varepsilon)},$$

где знак равенства следует из того, что при рассматриваемой функции вероятности $f^*(x)$ достигается супремум в (2.30).

Так как всего имеется 2^{nR}–1 последовательностей $\underline{e} \neq \widehat{\underline{x}}$, получаем

$$\sum_{\underline{e}\neq\widehat{\underline{x}}} P\left(E_{\underline{e}}\right) \leq \left(2^{nR}-1\right) 2^{-n(C-3\varepsilon)} \leq 2^{n(R-C+3\varepsilon)},$$

где правая часть стремится к нулю, если $n \to \infty$, а $R < C - 3\varepsilon$.

Таким образом, мы доказали, что среднее значение вероятности ошибки для случайного кода стремится к нулю. Это означает, что существует как минимум один код с $\overline{P_e^{(n)}} \to 0$, если $n \to \infty$. Поэтому скорость $R < C$ достижима.

В действительности же мы доказали и нечто большее, а именно что для любого $\varepsilon > 0$ найдется такое n_0, что существует последовательность кодов с длинами блока $n > n_0$, вероятностью ошибки, не превосходящей $\varepsilon + 2^{n(R-C+3\varepsilon)}$, и скоростью $R < C - 3\varepsilon$. □

Однако мы не дали руководства, как именно могут быть построены такие коды. Если бы мы использовали схему, предложенную в доказательстве, то нам бы надо было сгенерировать случайный код в соответствии с распределением (2.36), а для его декодирования хранить все 2^{nR} кодовых слов, количество которых экспоненциально возрастает с ростом n, тогда как $\overline{P_e^{(n)}} \to 0$ только при $n \to \infty$.

Теорема о передаче данных была сенсацией в статье Шеннона в 1948 г. Ранее предполагалось, что повышение надежности может достигаться уменьшением скорости передачи или повышением отношения сигнал/шум. Шеннон доказал, что, пока $R < C$, повышение надежности (малое ε) может быть достигнуто ценой повышения сложности кодирования (большое n) без того, что для этого необходимо повышать отношение сигнал/шум.

Как уже упоминалось во введении, пропускная способность канала связи может быть сравнена с пропускной способностью автомагистрали, по которой в единицу времени могут проехать максимум C автомобилей. Если в единицу времени по автомагистрали проезжает меньше C автомобилей, то принципиально возможно организовать движение таким образом, чтобы вероятность аварии или пробки была сколь угодно малой. Если же в единицу времени по автомагистрали пытается проехать больше C автомобилей, то ни один инженерный гений или гений дорожной инспекции не сможет избежать аварии или пробки без перенаправления части автомобилей в объезд. Действия инженера или дорожного инспектора могут быть сравнены при этом с работой кодеров и декодеров данных источника, роль которых в нашей аналогии играет поток автомобилей.

2.6. Теорема о сжатии и передаче данных

Теперь скомбинируем два основных полученных нами результата: сжатие данных ($H_\infty(\underline{U}) < R$, теорема 2.3) и передача данных ($R < C$, теорема 2.7). Является ли условие $H_\infty(\underline{U}) < C$ необходимым и достаточным для возможности сколь угодно точной посылки данных источника по каналу?

С одной стороны, мы могли бы разработать код для преобразования данных источника напрямую во входные данные канала, а с другой стороны, мы можем сначала сжать данные источника, а затем, используя соответствующий код канала, послать их по каналу. Поскольку сжатие данных не зависит от канала, а передача данных не зависит от источника, не является очевидным, что мы ничего не теряем, используя такой двухступенчатый метод.

В данном разделе мы докажем, что этот метод так же хорош, как и любой другой метод передачи информации по каналу без памяти. С практической точки зрения это означает, что при разработке системы передачи информации мы можем раздельно рассматривать кодирование источника сообщений и кодирование канала связи. Эта комбинация будет так же эффективна, насколько может быть эффективной наша разработка при рассмотрении обеих проблем вместе.

Пусть мы имеем случайный процесс $\underline{U} = \{U_1, U_2, \ldots, U_n\}$, где каждое U_t принимает значения из конечного алфавита $\mathcal{A}_U$. Мы не будем делать никаких других предположений о роде случайного процесса $\underline{U}$, кроме того, что он обладает свойством энтропийной устойчивости:

$$\operatorname*{plim}_{n\to\infty} -\frac{1}{n}\log f(\underline{u}) = H_\infty(\underline{U}). \tag{2.41}$$

Примеры таких процессов включают в себя последовательность н.о.р. случайных величин (теорема 2.1), любые эргодические процессы (теорема 1.11), в том числе стационарные неразложимые марковские процессы.

Мы хотим послать последовательность символов $\underline{u} = \{u_1, u_2, \ldots, u_n\}$ по каналу так, чтобы получатель смог надежно восстановить эту последовательность. Отобразим последовательность $\underline{U}$ на кодовые слова $\underline{X}$ и пошлем кодовые слова по каналу. Получатель делает оценку $\widehat{\underline{u}}$ последовательности $\underline{u}$, которая была послана. Он совершает ошибку с вероятностью $P_e^{(n)} = P(\underline{u} \neq \widehat{\underline{u}})$. Система передачи информации показана на рис. 2.7.

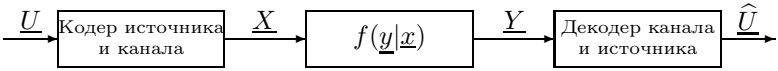


Рис. 2.7. К теореме о сжатии и передаче данных

Следующая теорема называется **теоремой Шеннона о сжатии и передаче данных**.

Теорема 2.8. *Если случайный процесс $\underline{U}$ с конечным алфавитом $\mathcal{A}_U^n$ удовлетворяет свойству энтропийной устойчивости (2.41), то существует комбинация кодов источника и канала с $P_e^{(n)} \to 0$, если абсолютная энтропия $H_\infty(\underline{U}) < C$.*

И наоборот, если $H_\infty(\underline{U}) > C$, то невозможно послать любой стационарный случайный процесс через канал связи и принять этот процесс с произвольно малой вероятностью ошибки.

Доказательство. Достижимость: Так как случайный процесс удовлетворяет свойству энтропийной устойчивости (2.41), то в соответствии с (2.9) имеется множество типичных последовательностей $\mathcal{T}_\varepsilon(\underline{U})$ с

$$\operatorname{card} \mathcal{T}_\varepsilon(\underline{U}) \le 2^{n(H_\infty(\underline{U})+\varepsilon)},$$

несущее в себе всю вероятность.

Применим раздельное, двухступенчатое кодирование. Для этого выделим множество $2^{n(H_\infty(\underline{U})+\varepsilon)}$ типичных последовательностей и закодируем только их. В соответствии с (2.27) скорость кода, состоящего только из типичных последовательностей, равна

$$R = \frac{\log_2 2^{n(H_\infty(\underline{U})+\varepsilon)}}{\log_2 2^n} = H_\infty(\underline{U}) + \varepsilon.$$

Поэтому по доказательству первой части теоремы 2.7 (о передаче данных) мы можем передать, а получатель может получить требуемую последовательность с вероятностью ошибки $P_e^{(n)} \to 0$, если $n \to \infty$ (то есть $\varepsilon \to 0$) и $R < C$, а следовательно, если $H_\infty(\underline{U}) < C$.

Обратное утверждение: Мы хотим показать, что $P_e^{(n)} \to 0$ означает, что $H_\infty(\underline{U}) \le C$ для любой комбинации кодов источника и канала с функциями кодирования $\mathsf{K}(\underline{U}) = \underline{X}$ и декодирования $\mathsf{D}(\underline{Y}) = \widehat{\underline{U}}$ (см. рис. 2.7).

Для абсолютной энтропии имеем

$$
\begin{aligned}
H_\infty(\underline{U}) &\le \frac{H(\underline{U})}{n} = && (2.42)\\
&= \frac{1}{n}H(\underline{U}|\widehat{\underline{U}}) + \frac{1}{n}I(\underline{U};\widehat{\underline{U}}) \le && (2.43)\\
&\le \frac{1}{n}(1 + P_e^{(n)} \log \operatorname{card} \mathcal{A}_U^n) + \frac{1}{n}I(\underline{U};\widehat{\underline{U}}) \le && (2.44)\\
&\le \frac{1}{n}(1 + P_e^{(n)} n \log \operatorname{card} \mathcal{A}_U) + \frac{1}{n}I(\underline{X};\underline{Y}) \le && (2.45)\\
&\le \frac{1}{n} + P_e^{(n)} \log \operatorname{card} \mathcal{A}_U + C, && (2.46)
\end{aligned}
$$

где (2.42) следует из доказательства теоремы 1.10 об энтропии стационарного случайного процесса (1.73), (2.43) следует из обобщенного на случайные процессы равенства (1.30), (2.44) следует из леммы Фано, (2.45) — из леммы об обработке данных (поскольку $\underline{U} \to \underline{X} \to \underline{Y} \to \widehat{\underline{U}}$ является цепью Маркова), а (2.46) — из отсутствия памяти канала (2.34). Теперь по условию мы имеем $P_e \to 0$ и, допуская $n \to \infty$, получаем $H_\infty(\underline{U}) \le C$.

Таким образом, мы можем передать по каналу связи стационарный случайный процесс, удовлетворяющий свойству энтропийной устойчивости (2.41), если, и только если, его абсолютная энтропия меньше, чем пропускная способность канала связи. □

Этим результатом мы объединили две основные теоремы теории информации: о сжатии данных и о передаче данных.

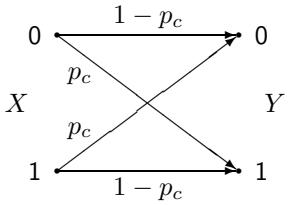
Подытожим результаты предыдущих разделов в нескольких словах. Теорема о сжатии данных — следствие понятия типичных последовательностей, которое показывает, что существует «малое» подмножество (с мощностью $2^{nH_\infty(\underline{U})}$) всех возможных исходных последовательностей, которое содержит в себе почти всю вероятность, то есть мы можем представить все последовательности источника со сколь угодно малой вероятностью ошибки, используя $H_\infty(\underline{U})$ бит на символ. Теорема о передаче данных базируется на понятии совместно-типичных последовательностей и использует тот факт, что, весьма вероятно, длинная последовательность на выходе канала связи совместно-типична с посланным кодовым словом, в то время как другое кодовое слово совместно-типично с ней с вероятностью около $2^{-nI(X;Y)}$ (теорема 2.6). Теорема о сжатии и передаче данных показывает, что мы можем конструировать код источника и код канала раздельно друг от друга.

2.7. Пропускная способность дискретных каналов связи

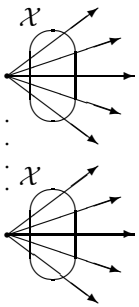
В разделе 2.5 мы дали определение дискретного канала связи без памяти как совокупности входного алфавита $\mathcal{A}_X$, выходного алфавита $\mathcal{A}_Y$ и условной функции вероятности $f(y|x)$. Дискретные каналы связи без памяти являются самыми простыми в математическом отношении каналами и при вычислении пропускной способности мы сконцентрируем внимание именно на них.

Для канала без памяти вероятность каждого конкретного символа y_t на выходе зависит только от соответствующего ему символа x_t на входе и не зависит от момента времени $t = 1, 2, \ldots, n$, поэтому на входе и выходе таких каналов достаточно рассматривать случайные величины X и Y, а не случайные процессы $\underline{X}$ и $\underline{Y}$ соответственно. Каналы без памяти удобно представлять в виде диаграммы, в которой слева вертикально расположены символы входного алфавита $\mathcal{A}_X$, а справа вертикально расположены символы выходного алфавита $\mathcal{A}_Y$. Из каждого входного символа в каждый выходной символ проводят вектор, который маркируют вероятностью появления на выходе данного выходного символа при условии, что на входе имеется данный входной. Исключение составляют только векторы с нулевыми вероятностями, которые иногда для простоты не проводятся.

Пример 2.4. Следующий канал называется **двоично-симметричным каналом** (ДСК) с вероятностью ошибки p_c, и его диаграмма имеет вид



Определение 2.6. *Дискретный канал без памяти со входным алфавитом A_X и выходным алфавитом A_Y называется* **равномерно дисперсивным**, *если множество $\mathcal{X} = \{p_1, p_2, \ldots, p_{\operatorname{card} A_Y}\}$ вероятностей для всех $\operatorname{card} A_Y$ векторов, выходящих из некоторого входного символа, инвариантно относительно входного символа:*



Пример 2.5. ДСК равномерно дисперсивен.

Лемма 2.3. *Для равномерно дисперсивного канала справедливо*

$$H(Y|X) = -\sum_{j=1}^{\operatorname{card} A_Y} p_j \log p_j, \tag{2.47}$$

где $p_1, p_2, \ldots, p_{\operatorname{card} A_Y}$ *— вероятности, соответствующие векторам, выходящим из любого входного символа.*

Доказательство. Из определения равномерной дисперсивности и энтропии следует, что

$$H(Y|X = x_i) = -\sum_{j=1}^{\operatorname{card} A_Y} p_j \log p_j$$

для $i = 1, 2, \ldots, \operatorname{card} A_X$, поэтому, используя (1.25), получаем

$$H(Y|X) = -\underbrace{\sum_{i=1}^{\operatorname{card} A_X} f(x_i)}_{1} \sum_{j=1}^{\operatorname{card} A_Y} p_j \log p_j = -\sum_{j=1}^{\operatorname{card} A_Y} p_j \log p_j.$$

□

Лемма 2.4. *Пропускная способность равномерно дисперсивного канала задается следующим выражением:*

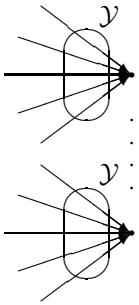
$$C = \sup_{f(x)} H(Y) + \sum_{j=1}^{\operatorname{card} A_Y} p_j \log p_j. \tag{2.48}$$

Доказательство. Используя (2.30) и (2.47), имеем

$$\begin{aligned} C = \sup_{f(x)} I(X;Y) &= \sup_{f(x)} (H(Y) - H(Y|X)) = \\ &= \sup_{f(x)} H(Y) + \sum_{j=1}^{\operatorname{card} A_Y} p_j \log p_j. \end{aligned} \tag{2.49}$$

□

Определение 2.7. *Дискретный канал без памяти называется* **равномерно фокусируемым**, *если множество* $\mathcal{Y}=\{p_1, p_2, \ldots, p_{\operatorname{card} A_X}\}$ *вероятностей всех* $\operatorname{card} A_X$ *векторов, входящих в некоторый выходной символ, инвариантно относительно выходного символа:*



Пример 2.6. ДСК является равномерно фокусируемым каналом.

Лемма 2.5. *Для равномерно фокусируемого дискретного канала без памяти выполняется равенство*

$$\sup_{f(x)} H(Y) = \log \operatorname{card} A_Y, \tag{2.50}$$

где супремум (необязательно однозначно) достигается при равномерном распределении вероятностей входных символов x_i*:*

$$f(x_i) = \frac{1}{\operatorname{card} A_X},\ i = 1, \ldots, \operatorname{card} A_X. \tag{2.51}$$

Доказательство. Покажем, что условием принятия энтропией $H(Y)$ максимального значения является равновероятность всех x_i (2.51). Действительно, тогда

$$f\left(y_j\right) = \sum_{i=1}^{\operatorname{card} A_X} f\left(y_j \,|x_i\,\right) f\left(x_i\right) = \frac{1}{\operatorname{card} A_X} \sum_{i=1}^{\operatorname{card} A_X} f\left(y_j \,|x_i\,\right),$$

где сумма в последнем равенстве принимает одинаковые значения для всех y_j в силу равномерной фокусируемости. Поэтому в этом случае все y_j тоже равновероятны, то есть $f(y_j) = \frac{1}{\operatorname{card} A_Y}$, $j = 1, \ldots, \operatorname{card} A_Y$, что является условием для принятия энтропией $H(Y)$ максимального значения $\log \operatorname{card} A_Y$ (1.19). □

Определение 2.8. *Если дискретный канал без памяти является как равномерно дисперсивным, так и равномерно фокусируемым, то он называется* **сильно симметричным**.

Теорема 2.9. *Пропускная способность сильно симметричного канала равна*

$$C = \log \operatorname{card} A_Y + \sum_{j=1}^{\operatorname{card} A_Y} p_j \log p_j, \tag{2.52}$$

где $p_1, p_2, \ldots, p_{\operatorname{card} A_Y}$ *— вероятности векторов, выходящих из любого входного символа. Данная пропускная способность достигается при равномерном распределении вероятностей входных символов.*

Доказательство. Теорема следует из лемм 2.4 и 2.5. □

Пример 2.7. Пропускная способность ДСК есть

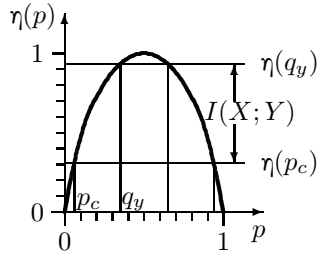
$$C = \sup_{f(x)} H(Y) + p_c \log p_c + (1-p_c)\log(1-p_c) = 1 - \eta(p_c), \tag{2.53}$$

так как ДСК одновременно равномерно дисперсивен (2.48) и фокусируем (2.50), то есть сильно симметричен (2.52).

Представляет интерес также графическая интерпретация пропускной способности ДСК. В соответствии с (2.49) и (2.53) взаимная информация между его входом и выходом равна $I(X;Y) = H(Y) - \eta(p_c)$. Если вероятность появления нуля на входе ДСК равна q_x, а вероятность появления единицы равна $1 - q_x$, то на его выходе будет ноль с вероятностью $q_y = q_x(1-p_c) + (1-q_x)p_c$, а единица — с вероятностью $1 - q_y$, поэтому $H(Y) = \eta(q_y)$ и

$$I(X;Y) = \eta(q_y) - \eta(p_c), \tag{2.54}$$

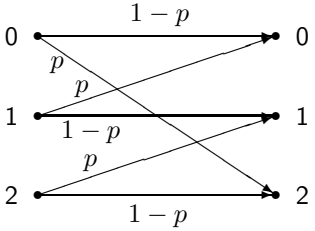
что отражено на рисунке



Заметим, что так как $q_y = q_x(1-p_c) + (1-q_x)p_c$ всегда должно лежать между p_c и $1 - p_c$, то $\eta(q_y) \geq \eta(p_c)$, и мы получаем графическое подтверждение неотрицательности взаимной информации $I(X;Y)$ (1.28).

Кроме того, для фиксированного значения p_c мы, варьируя q_x, или, что то же самое, варьируя $f(x)$, видим, что взаимная информация (2.54) достигает максимума при $q_x = 0{,}5$, то есть это значение соответствует пропускной способности ДСК (2.53), а функция вероятности $f(1) = f(0) = 0{,}5$ является согласованной с каналом связи. С другой стороны, взаимная информация между входом и выходом ДСК (2.54) равна нулю при $q_x = 0$ или $q_x = 1$. Поэтому для достижения максимальной взаимной информации (или, что то же самое, достижения пропускной способности) требуется согласование источника сообщений с каналом связи.

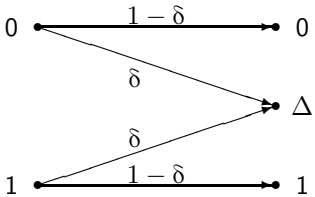
Пример 2.8. Троичный канал без памяти



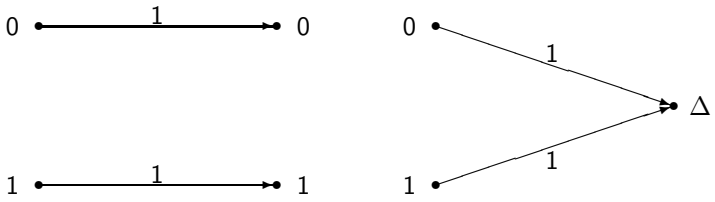
является сильно симметричным и имеет пропускную способность (2.52)

$$C = \log_2 3 + \sum_{j=1}^{3} p_j \log_2 p_j = \log_2 3 - \eta(p) \text{ бит.}$$

Пример 2.9. Рассмотрим **двоичный стирающий канал**, в котором с вероятностью δ посланный бит вытирается и принимается стирание Δ:



Данный канал эквивалентен комбинации двух следующих каналов:



С вероятностью $q_1 = 1 - \delta$ символ посылается через первый канал, имеющий пропускную способность $C_1 = 1$, а с вероятностью $q_2 = \delta$ — через второй канал, имеющий пропускную способность $C_2 = 0$.

Обобщим последний пример.

Определение 2.9. *Дискретный канал без памяти называется* **симметричным**, *если его* $\operatorname{card} A_Y$ *выходов могут быть разбиты на* n *множеств, в каждом по* $\operatorname{card} A_{Y_i}$ $(i = 1, 2, \ldots, n)$ *выходов так, что канал может быть разделен на* n *сильно симметричных компонентных каналов, каждый с* $\operatorname{card} A_X$ *входами, но* $\operatorname{card} A_{Y_i}$ *выходами. При этом каждый из каналов выбирается с вероятностью* $q_i, i = 1, 2, \ldots n$.

Большая часть встречающихся на практике каналов являются симметричными, поэтому докажем следующий результат.

Теорема 2.10. *Пропускная способность симметричного канала определяется формулой*

$$C = \sum_{i=1}^{n} q_i C_i, \tag{2.55}$$

где q_i *и* C_i *— вероятность и пропускная способность* i*-го сильно симметричного компонентного канала.*

Доказательство. Пусть случайная величина Z показывает, какой из n компонентных каналов был выбран для передачи. Тогда $P(Z = i) = q_i$ и $H(Z|Y) = 0$, так как наблюдаемое Y однозначно определяет Z. Далее

$$\begin{aligned} H(Y) &= H(Y) + 0 = H(Y) + H(Z|Y) = H(Z, Y) = \\ &= H(Z) + H(Y|Z) = H(Z) + \sum_{i=1}^{n} H(Y|Z = i) q_i, \end{aligned} \tag{2.56}$$

где в последнем равенстве использовано (1.25). Далее, так как Y однозначно определяет Z, имеем

$$\begin{aligned} H(Y|X) = H(Y, Z|X) &= H(Z|X) + H(Y|Z, X) = \\ &= H(Z) + \sum_{i=1}^{n} H(Y|X, Z = i) q_i, \end{aligned} \tag{2.57}$$

где последнее равенство следует из независимости Z и X, а также из (1.25). Теперь вычтем (2.57) из (2.56) и получим

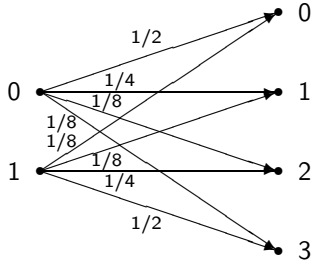
$$H(Y) - H(Y|X) = I(X;Y) = \sum_{i=1}^{n} (H(Y|Z=i) - H(Y|X,Z=i))q_i = \\ = \sum_{i=1}^{n} I_i(X;Y)q_i,$$

где $I_i(X;Y)$ — взаимная информация между входом и выходом i-го сильно симметричного компонентного канала, если этот канал используется один с тем же распределением вероятностей входных символов, что применялось и для полного канала.

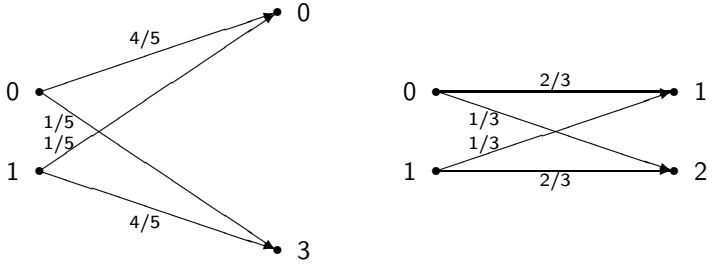
Кроме того, по теореме 2.9, равномерное распределение вероятностей входных символов максимизирует каждую взаимную информацию $I_i(X;Y)$ всех сильно симметричных компонентных каналов, поэтому теорема доказана. □

Пример 2.10. Пропускная способность двоичного стирающего канала равна (2.55) $C = 1 \cdot (1-\delta) + 0 \cdot \delta = 1 - \delta$.

Пример 2.11. Рассмотрим канал:



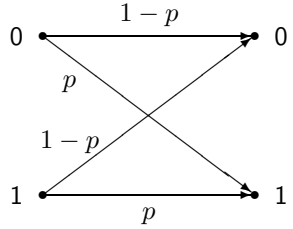
Он может быть разбит на два следующих сильно симметричных канала:



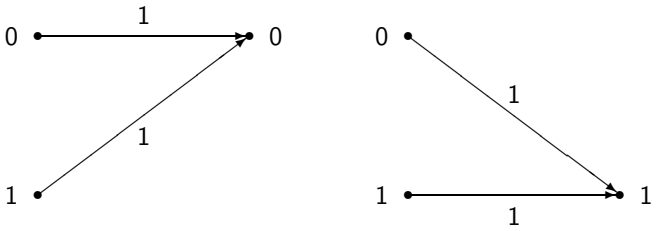
Первый из них выбирается с вероятностью $q_1 = 5/8$ и имеет пропускную способность $C_1 = 1 - \eta(1/5)$, а второй выбирается с вероятностью $q_2 = 3/8$ и имеет пропускную способность $C_2 = 1 - \eta(1/3)$. Поэтому пропускная способность исходного канала (2.55)

$$C = \frac{5}{8}\left(1 - \eta\left(1/5\right)\right) + \frac{3}{8}\left(1 - \eta\left(1/3\right)\right).$$

Пример 2.12. Следующий канал, несмотря на сходство с ДСК, имеет нулевую пропускную способность



Действительно, он может быть разбит на два следующих сильно симметричных канала:



Первый из каналов выбирается с вероятностью $q_1 = 1 - p$, а второй — с вероятностью $q_2 = p$, причем оба канала имеют пропускную способность, равную нулю.

Для исходного канала нулевая пропускная способность может быть объяснена тем, что, независимо от статистических особенностей последовательности на его входе, в последовательности на его выходе будут заведомо содержаться единицы с вероятностью p и нули с вероятностью $1 - p$. То есть знания вероятности p достаточно для описания последовательности на выходе канала связи, и никакая информация на его входе не может пройти на выход.

2.8. Гауссовский канал

В предыдущих разделах мы рассматривали дискретные каналы связи, для которых входные и выходные символы принадлежали конечным алфавитам и являлись символами дискретных по времени случайных процессов. Сейчас мы расширим ситуацию на каналы, в которых входные и выходные символы являются произвольными действительными числами, как в случае дискретного, так и в случае непрерывного времени.

Наиболее важным каналом связи с непрерывными входными (x_t) и выходными (y_t) символами является дискретный по времени гауссовский канал без памяти, описываемый формулой $y_t = x_t + z_t$ $(t = 1, 2, \ldots)$, где z_t представляют собой значения нормально распределенной случайной величины Z с математическим ожиданием $\overline{Z} = 0$ и дисперсией $\overline{Z^2} = N$. Причем значения этой случайной величины не зависят друг от друга в различные моменты времени t и никак не связаны с передаваемым сигналом x_t. Классическими примерами гауссовского канала являются каналы в дальней космической связи и телефонные каналы.

Прежде чем вычислить пропускную способность такого канала связи, заметим, что из (1.40) следует

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) = h(Y) - h(X+Z|X) = \\ &= h(Y) - h(Z|X) = h(Y) - h(Z), \end{aligned} \tag{2.58}$$

поскольку Z не зависит от X.

При заданной средней мощности входного сигнала $\overline{X^2} = E$ средняя мощность выходного сигнала вычисляется как

$$\overline{Y^2} = \overline{(X+Z)^2} = \overline{X^2} + 2\overline{X}\,\overline{Z} + \overline{Z^2} = E + N.$$

Но при фиксированной величине $\overline{Y^2} = E + N$ дифференциальная энтропия Y ограничена сверху величиной (1.36):

$$h(Y) \le \log \sqrt{2\pi e(E+N)}, \tag{2.59}$$

поскольку гауссовское распределение максимизирует дифференциальную энтропию при заданных математическом ожидании и дисперсии (см. пример 1.10).

Теперь из (1.36) мы имеем $h(Z) = \log\sqrt{2\pi eN}$, поэтому из (2.58) и (2.59) получаем

$$I(X;Y) \le \log\sqrt{2\pi e(E+N)} - \log\sqrt{2\pi eN} = \frac{1}{2}\log\left(1+\frac{E}{N}\right). \qquad (2.60)$$

Из (2.60) мы видим, что $I(X;Y)$ может быть сколь угодно большой, если сколь угодно большим будет E. Поэтому проблема определения пропускной способности $C = \sup\limits_{f(x)} I(X;Y)$ гауссовского канала связи имеет физический смысл только тогда, когда имеются некоторые ограничения на входной сигнал. Наиболее часто ограничением является средняя мощность входного сигнала $\overline{X^2} \le E$. Поэтому мы доказали следующую теорему.

Теорема 2.11. *Пропускная способность дискретного по времени гауссовского канала без памяти с дисперсией шума N и не превосходящей E мощностью входного сигнала задается формулой*

$$C = \frac{1}{2}\log\left(1+\frac{E}{N}\right). \qquad (2.61)$$

Теперь мы рассмотрим канал связи, у которого входные и выходные символы являются непрерывными по времени функциями. Причем дополнительно введем ограничение на максимальную частоту в спектре входного сигнала: $F_{\max} = W$.

Из теории сигналов известно (теорема отсчетов, называемая именем Котельникова в русскоязычной литературе и именами Найквиста-Шеннона в англоязычной), что в этом случае непрерывная функция Z может быть однозначно и без потерь восстановлена по своим дискретным отсчетам, взятым с частотой не менее чем $2W$. Или другими словами — по отсчетам, взятым с периодом не менее $\frac{1}{2W}$ за одну секунду: $Z_t = Z(\frac{t}{2W})$. То есть с помощью отсчетов мы можем свести ограниченный по частоте и непрерывный по времени канал к эквивалентному дискретному по времени каналу.

Причем в случае, когда мы таким образом дискретизируем по времени непрерывную случайную величину, имеющую гауссовское распределение, мы получим последовательность независимых и нормально распределенных случайных величин.

Теперь рассмотрим ограниченный по частоте ($F_{\max} = W$) и непрерывный по времени канал с аддитивным шумом, имеющим гауссовское

распределение и спектральную плотность мощности $N_0/2$, — ограниченный по частоте гауссовский канал. Тогда отсчеты шума являются независимыми нормально распределенными случайными величинами с дисперсией $N_0/2$, а в интервале времени, равном T секунд, имеется $n = 2WT$ отсчетов. То есть наш канал эквивалентен n параллельным дискретным по времени гауссовским каналам, каждый с пропускной способностью (2.61)

$$C_t = \frac{1}{2}\log\left(1 + \frac{2E}{N_0}\right), \tag{2.62}$$

где E — средняя мощность каждого отсчета входного сигнала.

Теперь предположим, что средняя мощность входного сигнала ограничена величиной

$$S = \frac{1}{T}\sum_{t=1}^{n}\overline{X_k^2} = \frac{nE}{T}. \tag{2.63}$$

Тогда из (2.62) и (2.63) следует, что

$$C_t = \frac{1}{2}\log\left(1 + \frac{2TS}{nN_0}\right) = \frac{1}{2}\log\left(1 + \frac{S}{WN_0}\right),$$

где учтено, что $n = 2WT$. Поэтому общая пропускная способность n параллельных каналов (задача 2.7.3 (а)) есть

$$\sum_{t=1}^{n} C_t = WT\log\left(1 + \frac{S}{WN_0}\right). \tag{2.64}$$

Определение 2.10. *Пропускная способность непрерывного по времени канала связи определяется как предел*

$$C = \lim_{T\to\infty}\frac{C_T}{T}, \tag{2.65}$$

где C_T — максимальная совместная информация, которая может быть передана за интервал времени T: $C_T = \sup\limits_{f(x)} I(X;Y)$, где супремум берется по всем плотностям распределения входного сигнала, удовлетворяющим (2.63).

Поэтому из (2.64) и (2.65) получаем следующий знаменитый результат Шеннона.

Теорема 2.12. *Пропускная способность ограниченного по частоте непрерывного гауссовского канала связи с аддитивным шумом, имеющим спектральную плотность мощности $N_0/2$, и не превосходящей S средней мощностью входного сигнала задается формулой*

$$C = W \log \left(1 + \frac{S}{WN_0}\right). \tag{2.66}$$

Пример 2.13. Типичный телефонный канал связи имеет соотношение сигнал/шум $\frac{S}{WN_0} = 25$ дБ (или $\frac{S}{WN_0} = 10^{2,5}$ раз). Для возможности передачи по одной линии нескольких разговоров телефонные сигналы ограничивают по частоте до $W = 3100$ Гц. Поэтому в предположении, что шумы в телефонной линии являются гауссовскими, пропускная способность (2.66) такого канала будет равна $3100 \log_2(1 + 10^{2,5}) \approx 25{,}7$ кбит/с. Для очень хороших телефонных каналов соотношение сигнал/шум может доходить до 40 дБ, при которых пропускная способность будет $3100 \log_2(1 + 10^4) \approx 41{,}2$ кбит/с. С примерно такими скоростями в телефонных линиях работают модемы.

Если мы зафиксируем мощность сигнала S и будем увеличивать пропускную способность, увеличивая максимальную частоту W, то получим

$$C_\infty = \lim_{W \to \infty} W \log \left(1 + \frac{S}{WN_0}\right) = \frac{S}{N_0 \ln 2} \text{ бит/с}, \tag{2.67}$$

то есть пропускная способность неограниченного по частоте гауссовского канала связи линейно растет с увеличением мощности входного сигнала.

Если суммарная мощность сигнала на интервале времени T секунд равна ST (2.62), а в сигнале содержится только k информационных бит, то мощность, приходящаяся на один информационный бит, равна $E_b = ST/k$. При этом скорость передачи данных равна $R = k/T$ бит в секунду, то есть $E_b = S/R$. Используя (2.67), получаем, что $\frac{C_\infty}{R} = \frac{E_b}{N_0 \ln 2}$.

А поскольку для надежной связи требуется $R < C_\infty$, мы получаем фундаментальную границу для отношения $\frac{E_b}{N_0}$:

$$\frac{E_b}{N_0} > \ln 2 \approx 0,69 = -1,6 \text{ дБ},$$

которая не должна нарушаться для любой системы связи с аддитивным гауссовским шумом.

2.9. Задачи

2.1.1. [2 балла] Предположим, у нас имеется урна с одним черным и двумя белыми шарами. Мы вынимаем один из шаров n раз и всякий раз кладем его обратно. Событие X заключается в том, что мы вынимаем некоторый шар.

(а) Определите энтропию $H(X)$ двоичной случайной величины X, связанной с данным событием.

(б) Сгенерируйте все возможные значения последовательности случайных величин $\{X_1, X_2, X_3, X_4, X_5\}$ для $n = 5$ и отметьте все типичные последовательности для $\varepsilon = 0{,}138$. Какая последовательность имеет наибольшую вероятность? Является ли она типичной?

(в) Подсчитайте количество всех типичных последовательностей, их долю среди всех возможных последовательностей и их суммарную вероятность P_ε. Сравните подсчитанное количество последовательностей с верхней и нижней границей (2.9) для мощности множества $\mathcal{T}_\varepsilon(X)$ всех типичных последовательностей.

(г) Повторите пункты (а)–(в) для $n = 10, 100, 1000, 2000$ и различных ε.

(д) Найдите такое n_0, начиная с которого $P_\varepsilon > 1 - \varepsilon$.

2.1.2. [2 балла] Пусть $\underline{U} = \{U_1, U_2, \ldots, U_n\}$ — случайный процесс, состоящий из н.о.р. случайных величин, принимающих значения из множества $\{0, 1\}$, каждая с функцией вероятности $f(u)$, определяемой как $f(1) = p \leq 0{,}5$ и $f(0) = 1 - p$. Оцените величину p, если известно, что количество типичных последовательностей длины n примерно равно $1{,}5^n$ для больших n.

2.1.3. [3 балла] Сформулируйте определение типичных последовательностей для непрерывных случайных величин и докажите для них теорему 2.2, пользуясь вместо понятия мощности дискретного множества понятием **объема** непрерывного n-мерного множества $\mathcal{T}_\varepsilon$:

$$\operatorname{vol} \mathcal{T}_\varepsilon = \int_{\mathcal{T}_\varepsilon} dx_1 dx_2 \ldots dx_n.$$

2.3.1. [2 балла] В теореме 2.5 дана верхняя граница для количества последовательностей $\underline{x}$ совместно-типичных с заданной последовательностью $\underline{y}$. Докажите вместо нижней границы, что для любого $\varepsilon > 0$ $(1 - \delta) 2^{n(\overline{H}(Y|X) - 2\varepsilon)} \leq \sum\limits_{\underline{y}} f\left(\underline{y}\right) \operatorname{card} \mathcal{T}_\varepsilon \left(X|\underline{y}\right)$.

2.4.1. [2 балла] Пусть $X_1, X_2, \ldots, X_n$ образуют цепь Маркова (1.79). Запишите взаимную информацию $I(X_1; X_2, \ldots, X_n)$ в ее самой простой форме.

2.5.1. [2 балла] Докажите, что пропускная способность любого канала связи без памяти лежит в пределах

$$0 \le C \le \min\{\log \operatorname{card} \mathcal{A}_X, \log \operatorname{card} \mathcal{A}_Y\}.$$

В каком случае достигается знак равенства справа?

2.5.2. [2 балла] Случайный процесс без памяти, имеющий абсолютную энтропию 15 бит на символ, передается по двоично-симметричному каналу, который имеет вероятность ошибки 0,1 и может пропустить до 1000 бит в секунду. С какой максимальной скоростью можно генерировать символы случайного процесса, чтобы иметь возможность надежно передать их через такой канал?

2.7.1. [2 балла] Объясните, почему пропускная способность двоичного симметричного канала (см. пример 2.7) при $p = 0{,}5$ равна нулю, а пропускная способность троичного симметричного канала (см. пример 2.8) при $p = 0{,}5$ не равна нулю.

2.7.2. [3 балла] Докажите, что последовательное соединение n идентичных двоично-симметричных каналов с вероятностью ошибки p эквивалентно одному двоично-симметричному каналу с вероятностью ошибки $(1 - (1 - 2p)^n)/2$. Исследуйте поведение пропускной способности нового канала при $n \to \infty$.

2.7.3. Рассмотрим n дискретных каналов без памяти, каждый с пропускной способностью C_i, $i = 1, 2, \ldots, n$.

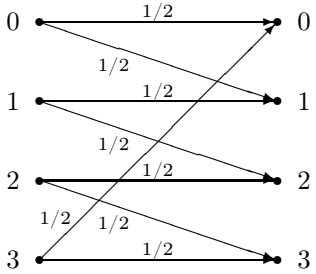
(а) [4 балла] **Произведением каналов** называется канал, входным и выходным алфавитами которого являются соответственно Декартовы произведения входных и выходных алфавитов исходных каналов. При этом в каждый момент времени по каждому из каналов передается некоторый символ. Докажите, что пропускная способность произведения каналов задается выражением $C_{prod} = \sum_{i=1}^{n} C_i$.

(б) [5 баллов] **Суммой каналов** называется канал, входным и выходным алфавитами которого являются соответственно объединения непересекающихся алфавитов исходных каналов. При этом в сумме все каналы можно использовать, но так, что в каждый момент времени можно вести передачу только по одному из каналов. Докажи-

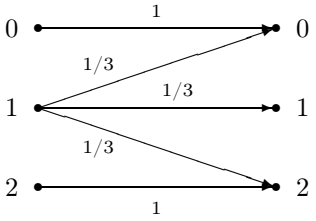
те, что пропускная способность суммы каналов задается выражением: $C_{sum} = \log \sum_{i=1}^{n} 2^{C_i}$.

2.7.4. [3 балла] Пусть $Y = X + Z$, где X — двоичная дискретная случайная величина, а Z — независимая от X случайная величина, для которой справедливо $P(Z = 0) = P(Z = a) = 0{,}5$. Найдите пропускную способность канала связи $(\mathcal{A}_X, f(y|x), \mathcal{A}_Y)$ и объясните ее зависимость от a.

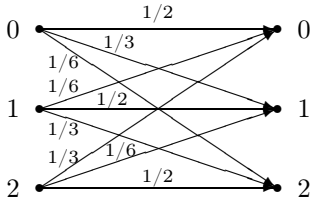
2.7.5. Определите пропускную способность следующих каналов связи и найдите согласованную с каждым из этих каналов функцию вероятности входных символов:



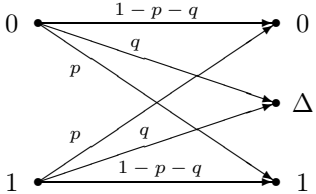
(а) [2 балла]



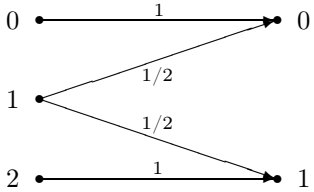
(б) [3 балла]



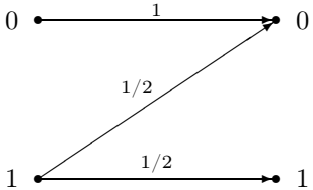
(в) [2 балла]



(г) [3 балла]



(д) [3 балла]



(е) [4 балла]

Глава 3

Кодирование источника сообщений

3.1. Классы кодов источника и неравенство Крафта

Рассмотрим теперь **источник информации** более подробно. Ограничимся при этом только **дискретными источниками**, то есть такими, которые в каждую единицу времени выбирают одно из сообщений из некоторого конечного алфавита $\mathcal{A}_U = \{a_1, a_2, \ldots, a_L\}$ и задаются многомерной совместной функцией вероятности $f(\underline{u})$, позволяющей вычислить вероятность любого отрезка сообщений. В этом смысле наше определение источника эквивалентно определению 1.9 дискретного случайного процесса $\underline{U} = \{U_1, U_2, \ldots, U_n\}$, где каждое U_t, $t = 1, \ldots, n$ принимает значения из $\mathcal{A}_U$.

Назовем дискретный источник **стационарным**, если сообщения на его выходе образуют стационарный случайный процесс. Напомним, что в соответствии с определением 1.6 все последовательности на выходе такого источника, отличающиеся только положением на оси времени, имеют одинаковые вероятностные характеристики. По аналогии определим источник **без памяти** и **эргодический** источник как имеющие на выходе соответственно стационарный случайный процесс без памяти и эргодический случайный процесс.

Определим **кодирование источника** как преобразование случайного процесса $\underline{U}$ в случайный процесс $\underline{Z}$, состоящий из следующих друг за другом в общем случае D-ичных **кодовых слов** $\underline{z}_i$, $i = 1, 2, \ldots$, выбираемых из сформированного по некоторому правилу множества $\mathcal{Z}$. Множество $\mathcal{Z}$ кодовых слов $\underline{z}_i$ называется **кодом источника**.

Для построения кода источника случайный процесс $\underline{U}$ может разбиваться на блоки постоянной или переменной длины, а на выходе кодера источника каждому такому блоку соответствуют кодовые слова, которые также могут иметь постоянную или переменную длину. Если кодовые слова имеют постоянную длину, то кодирование называется **равномерным**. В противном случае оно называется **неравномерным**.

Например, для источников без памяти $H_\infty(\underline{U}) = H(U)$ (1.67), где $H(U)$ — энтропия каждой из случайных величин, входящих в процесс $\underline{U}$, а $f(\underline{u}) = \prod_{t=1}^{n} f(u_t)$, поэтому для их кодирования достаточно построить код для отдельной случайной величины U и последовательно применять его к каждой из величин U_t, $t = 1, \ldots, n$. Для такого источника важно знать только вероятности $f(a_i)$ отдельных сообщений a_i. Примерами действующих по такому статическому принципу алгоритмов являются рассматриваемые далее алгоритмы Фано, Шеннона, Хаффмена и Танстелла.

Если источник имеет память, то применение аналогичного статического подхода может оказаться неэффективным, поскольку он не учитывает связей между величинами U_t. В данном случае можно разделить случайный процесс $\underline{U}$ на фразы, рассматривать эти фразы как отдельные сообщения и применять один из упомянутых алгоритмов (см. задачу 3.4.2).

Другой подход заключается в динамическом формировании множества кодовых слов, называемом **универсальным кодированием**. Такие алгоритмы не зависят от априорного распределения вероятностей отрезков случайного процесса $\underline{U}$, а оценивают их, самообучаясь в процессе работы. Рассматриваемые далее алгоритмы Рябко – Элайеса и Лемпеля – Зива являются примерами универсального кодирования. Промежуточное положение между статическими и универсальными методами занимает арифметическое кодирование.

Наше рассмотрение кодирования источника мы начнем с неравномерного кодирования источников без памяти.

В этом случае получателю известно, какие сообщения могли быть переданы (алфавит $\mathcal{A}_U$ и соответствующие кодовые слова), но не конкретное сообщение, переданное в конкретный момент времени. При этом он должен быть в состоянии однозначно разделить последовательность кодовых слов из кодера источника на кодовые слова и поставить им в соответствие переданные сообщения. Заметим, что в случае равномерного кодирования получатель с такой проблемой не сталкивается, так как она решается простым разделением принятой последовательности на блоки известной одинаковой длины.

Примером неравномерного кодирования может служить азбука Морзе, где наиболее частой букве английского языка «E» поставлена в соответствие одна точка «$\cdot$», а редкая буква «Q» представлена длинной комбинацией «$- - \cdot -$». Разделителем между буквами в аз-

буке Морзе служат паузы. Однако при неравномерном кодировании можно обойтись и без разделителя.

Во-первых, потребуем от нашей схемы кодирования, чтобы никакие два кодовых слова не являлись идентичными, то есть $\underline{z}_i \neq \underline{z}_j$, если $i \neq j$. Коды, удовлетворяющие этому условию, называются **инъективными**.

Во-вторых, мы хотим, чтобы без применения разделителя сообщений принятую строку из многих сообщений было возможно однозначно разделить на подстроки, соответствующие отдельным сообщениям. Коды, удовлетворяющие этому требованию, называются кодами **со свойством однозначного декодирования** или **однозначно декодируемыми** кодами. Однако для выделения из принятой строки отдельных сообщений для таких кодов в общем случае требуется знание полной принятой строки. Если же у кода со свойством однозначного декодирования ни одно кодовое слово не является началом (префиксом) никакого другого кодового слова, то в этом случае возможно однозначно распознать любое кодовое слово, как только получен его последний символ. Такой код называется **префиксным**[1].

На рис. 3.1 дана графическая интерпретация соотношения между различными классами кодов источника.

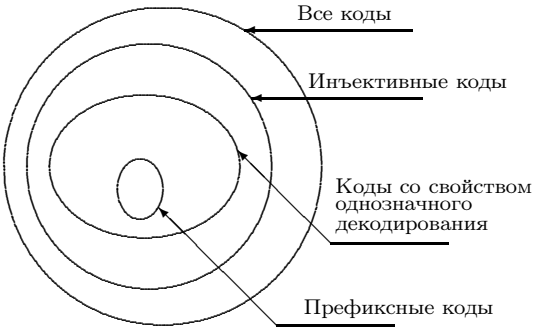


Рис. 3.1. Классы кодов источника

Пример 3.1. В таблице приведены примеры неоднозначно декодируемого инъективного кода $\mathcal{Z}_{in}$, непрефиксного кода со свойством однозначного декодирования $\mathcal{Z}_{un}$, а также префиксного кода $\mathcal{Z}_{pr}$:

[1]По сути данного свойства правильнее было бы называть такие коды беспрефиксными, однако будем следовать принятому в русскоязычной литературе термину.

$\mathcal{A}_U$	$\mathcal{Z}_{in}$	$\mathcal{Z}_{un}$	$\mathcal{Z}_{pr}$
a_1	0	10	0
a_2	01	00	10
a_3	10	11	110
a_4	010	110	111

Для инъективного кода $\mathcal{Z}_{in}$ строка 010 может соответствовать трем последовательностям сообщений: a_4, a_1a_3 или a_2a_1, поэтому такой код однозначно декодировать нельзя.

Если первые два бита для кода $\mathcal{Z}_{un}$ равны 10 или 00, то сообщения a_1 или a_2 декодируются однозначно. Если же первые два бита равны 11, то необходимо проанализировать следующие биты. Если третий бит равен 1, то первым передано сообщение a_3, если после 11 следует нечетное количество нулей, то первым передано сообщение a_4, а если четное, то первое сообщение — a_3. Повторяя приведенное рассуждение, возможно декодировать всю строку.

Для кода $\mathcal{Z}_{pr}$ мы можем в принятой строке расставить разделители между сообщениями без анализа всей строки. Например, строка 01011111010 однозначно разделяется на 0, 10, 111, 110, 10.

Для кодирования сообщений источника пригодны только коды со свойством однозначного декодирования, поэтому докажем необходимое условие того, что код удовлетворяет этому свойству.

Теорема 3.1. *Если некоторый D-ичный код источника обладает свойством однозначного декодирования, то длины $w_1, w_2, \ldots, w_L$ его кодовых слов удовлетворяют* **неравенству Крафта**

$$\sum_{i=1}^{L} D^{-w_i} \leq 1. \tag{3.1}$$

Доказательство. Пусть k — произвольное натуральное число. Рассмотрим k-ю степень суммы в левой части (3.1):

$$\left(\sum_{i=1}^{L} D^{-w_i}\right)^k = \sum_{i_1=1}^{L} \ldots \sum_{i_k=1}^{L} D^{-\left(w_{i_1}+\ldots+w_{i_k}\right)}, \tag{3.2}$$

причем каждое слагаемое в правой части (3.2) однозначно соответствует одной из возможных последовательностей из k кодовых слов, суммарная длина которой равна $w_{i_1} + \ldots + w_{i_k}$. Обозначим через $A(j)$

количество последовательностей из k кодовых слов, имеющих суммарную длину j. Тогда

$$\left(\sum_{i=1}^{L} D^{-w_i}\right)^k = \sum_{j=1}^{kw_{\max}} A(j) D^{-j}, \tag{3.3}$$

где $w_{\max} = \max\limits_i w_i$.

Так как максимальное количество различных D-ичных последовательностей длины j равно D^j, то для возможности однозначного декодирования необходимо выполнение условия

$$A(j) \le D^j. \tag{3.4}$$

Подставляя (3.4) в (3.3), получим

$$\left(\sum_{i=1}^{L} D^{-w_i}\right)^k \le kw_{\max},$$

или

$$\sum_{i=1}^{L} D^{-w_i} \le (kw_{\max})^{1/k} = 2^{\frac{\log_2 kw_{\max}}{k}}. \tag{3.5}$$

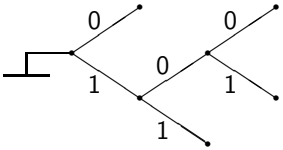
Так как неравенство в (3.5) справедливо при всех натуральных k, оно справедливо и при $k \to \infty$. А так как $\lim\limits_{k\to\infty} (\log_2 kw_{\max})/k = 0$, получаем утверждение теоремы. □

Докажем теперь, что для префиксных кодов неравенство Крафта (3.1) является и достаточным условием их существования. Для этого рассмотрим их связь со специальными графами, называемыми деревьями.

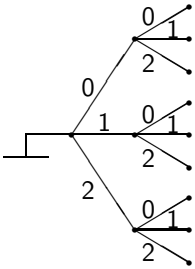
D-ичным деревом называется граф, то есть система узлов и связывающих их ветвей, в котором нет петель или замкнутых путей и в котором из каждого узла выходит не более D ветвей, а в каждый узел (кроме корня) входит точно одна ветвь. Будем обозначать выходящие из каждого узла D ветвей с помощью D различных D-ичных символов. **Полное D-ичное дерево** глубины $w_{\max}$ — это D-ичное дерево, в котором имеется в точности $D^{w_{\max}}$ конечных узлов (на глубине $w_{\max}$).

Пример 3.2.

Двоичное дерево глубины 3



Полное троичное дерево глубины 2



3^2 конечных узлов

Очевидно, что код является префиксным, если кодовые слова соответствуют только концевым узлам дерева. В противном случае код не является префиксным. Поэтому каждый префиксный код может быть идентифицирован с множеством конечных узлов некоторого D-ичного дерева, и наоборот, каждое множество конечных узлов D-ичного дерева задает префиксный код. Дерево некоторого префиксного кода становится однозначным, если его усечь, удалив все ветви, не принадлежащие кодовым словам. Неиспользуемые узлы дерева будем в дальнейшем обозначать белыми кружочками.

Теорема 3.2. *Для того чтобы существовал D-ичный префиксный код с длинами кодовых слов $w_1, w_2, \ldots, w_L$, необходимо и достаточно, чтобы выполнялось* **неравенство Крафта**

$$\sum_{i=1}^{L} D^{-w_i} \le 1. \tag{3.6}$$

Доказательство. Необходимость. Заметим, что необходимость условия (3.6) для всех однозначно декодируемых кодов была доказана в теореме 3.1, тем не менее мы приведем простое и поучительное доказательство необходимости для префиксных кодов. Предположим существование D-ичного префиксного кода, длины кодовых слов которого $w_1, w_2, \ldots, w_L$, и докажем, что при этом выполняется неравенство (3.6).

Пусть опять $w_{\max} = \max\limits_i w_i$. Используем тот факт, что в полном D-ичном дереве глубины $w_{\max}$ каждому узлу на глубине w_i всегда соответствует $D^{w_{\max}-w_i}$ конечных узлов. Теперь сконструируем одно-

значное дерево этого кода, удаляя неиспользуемые ветви и узлы полного D-ичного дерева глубины $w_{\max}$. Всякий раз для получения кодового слова $\underline{z}_i$ длины w_i мы удаляем $D^{w_{\max}-w_i}$ конечных узлов, так как ни один из этих узлов не может являться конечным для любого другого кодового слова по причине префиксности. Так как всего имеется только $D^{w_{\max}}$ узлов, которые можно удалить описанным выше способом, должно выполняться неравенство

$$D^{w_{\max}-w_1} + D^{w_{\max}-w_2} + \ldots + D^{w_{\max}-w_L} \leq D^{w_{\max}},$$

откуда получаем неравенство (3.6).

Достаточность. Предположим, $w_1, w_2, \ldots, w_L$ — положительные целые числа, такие, что выполняется (3.6). Нам необходимо показать, что в этом случае существует D-ичный префиксный код. Положим без ограничения общности, что $w_1 \leq w_2 \leq \ldots \leq w_L = w_{\max}$.

Рассмотрим полное D-ичное дерево глубины $w_{\max}$. Выберем узел кодового слова $\underline{z}_1$, соответствующий глубине w_1, и усечем полное дерево, начиная с этого узла. Для оставшегося дерева проведем ту же операцию для узла $\underline{z}_2$ глубины w_2 и покажем, что, действуя так и далее (пока не будут выбраны все L узлов), мы всегда сможем дойти до последнего L-го узла, сконструировав тем самым префиксный код.

Пусть узлы на глубинах $w_1, w_2, \ldots, w_{i-1}$ уже были таким образом выбраны. Тогда количество оставшихся конечных (с глубиной $w_{\max}$) узлов равно

$$D^{w_{\max}} - \left(D^{w_{\max}-w_1} + \ldots + D^{w_{\max}-w_{i-1}}\right) = D^{w_{\max}} \left(1 - \sum_{j=1}^{i-1} D^{-w_j}\right).$$

Причем данная величина положительна для всех $i \leq L$ в силу (3.6). Значит, если имеются узлы глубины $w_{\max}$, то должны иметься еще не использованные узлы глубины $w_i < w_{\max}$. Так как $w_1 \leq w_2 \leq \ldots \leq$ $\leq w_{i-1} \leq w_i$, ни одно из уже выбранных кодовых слов $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_{i-1}$ не может проходить через такой узел, и поэтому данный узел может быть выбран в качестве $\underline{z}_i$. □

Заметим, что доказательство достаточности в неравенстве Крафта содержит алгоритм для построения D-ичного префиксного кода с длинами кодовых слов $w_1, w_2, \ldots, w_L$ в случае, если таковой существует.

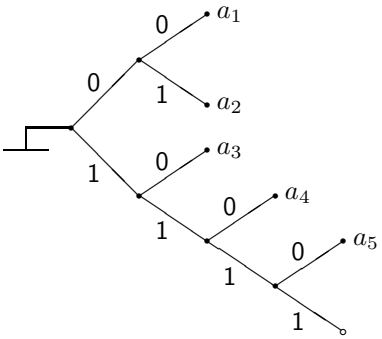
Кроме того, так как любой префиксный код является кодом со свойством однозначного декодирования, из доказательства достаточности следует, что, имея удовлетворяющий неравенству Крафта набор длин кодовых слов $w_1, w_2, \ldots, w_L$, мы можем сконструировать однозначно декодируемый код. Данный факт естественным образом дополняет теорему 3.1 и позволяет сформулировать ее по аналогии с теоремой 3.2 в форме «необходимо и достаточно».

Поэтому, несмотря на то что класс кодов со свойством однозначного декодирования шире класса префиксных кодов, он не дает нам возможности получить множества длин кодовых слов, которых бы не было в классе префиксных кодов.

Пример 3.3. Построим двоичный префиксный код с длинами кодовых слов $w_1 = 2,\ w_2 = 2,\ w_3 = 2,\ w_4 = 3,\ w_5 = 4$. Так как

$$\sum_{i=1}^{5} 2^{-w_i} = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = \frac{15}{16} \leq 1,$$

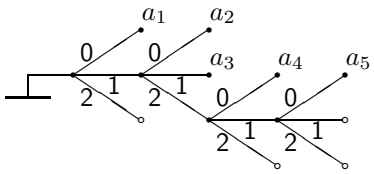
такой код существует. Построим его шаг за шагом, начиная с корня



$\mathcal{A}_U$	$\mathcal{Z}$
a_1	00
a_2	01
a_3	10
a_4	110
a_5	1110

Пример 3.4. Построим троичный префиксный код с длинами кодовых слов $w_1 = 1,\ w_2 = 2,\ w_3 = 2,\ w_4 = 3,\ w_5 = 4$. Такой код существует, так как

$$\sum_{i=1}^{5} 3^{-w_i} = \frac{1}{3} + \frac{1}{9} + \frac{1}{9} + \frac{1}{27} + \frac{1}{81} = \frac{49}{81} \leq 1.$$



$\mathcal{A}_U$	$\mathcal{Z}$
a_1	0
a_2	10
a_3	11
a_4	120
a_5	1210

Двоичный же код с теми же длинами кодовых слов построить нельзя:

$$\sum_{i=1}^{5} 2^{-w_i} = \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = \frac{19}{16} > 1.$$

3.2. Оптимальность кодов, алгоритмы Фано и Шеннона

Определим **среднюю длину кодового слова** при неравномерном кодировании как

$$\overline{W} = \sum_{i=1}^{L} f(a_i)\, w_i. \tag{3.7}$$

Она представляет собой среднее количество кодовых символов (для двоичных кодов — бит), приходящихся на одно сообщение, то есть не что иное, как определенную в разделе 2.2 скорость кодирования случайного процесса.

Средняя длина кодового слова используется как мера качества неравномерного кодирования: чем она короче, тем эффективнее кодирование.

Коды, для которых средняя длина кодового слова является наименьшей из всех возможных, называются **оптимальными**. Попытки построения оптимальных кодов породили ряд идей, первые из которых были предложены Фано и Шенноном.

Алгоритм Фано заключается в том, что вероятности появления отдельных сообщений записываются в порядке убывания, а затем делятся на D групп так, чтобы суммарные вероятности в каждой группе были как можно более близкими друг другу. Каждая группа получает при этом один из D символов в качестве первого символа кодового

слова. Затем подобное разделение повторяют внутри каждой группы до тех пор, пока в каждой группе не останется по одному сообщению.
Пример 3.5. Рассмотрим следующую случайную величину U и двоичный ($D = 2$) префиксный код $\mathcal{Z}_{pr}$ (из примера 3.1) для нее

$\mathcal{A}_U$	$f(a_i)$	$\mathcal{Z}_{pr}$
a_1	0,45	0
a_2	0,30	10
a_3	0,15	110
a_4	0,10	111

Данный код мог бы быть построен по алгоритму Фано, причем в колонке $\mathcal{Z}_{pr}$ горизонтальными линиями показано последовательное разделение сначала на две группы с вероятностями 0,45 и 0,55, затем — разделение второй группы на две подгруппы с вероятностями 0,30 и 0,25 и, наконец, последней группы — на подгруппы с вероятностями 0,15 и 0,10.

Мы имеем $H(U) = -\sum\limits_{i=1}^{4} f(a_i)\log_2 f(a_i) = 0{,}518 + 0{,}521 + 0{,}411 + 0{,}332 = 1{,}782$, а $\overline{W} = \sum\limits_{i=1}^{4} f(a_i)w_i = 0{,}45\cdot 1 + 0{,}30\cdot 2 + 0{,}15\cdot 3 + 0{,}10\cdot 3 = 1{,}80$, то есть $\overline{W} = 1{,}01H(U)$, что является очень хорошим результатом. Однако мы не можем сказать, является ли код $\mathcal{Z}_{pr}$ оптимальным. Но в любом случае наиболее вероятное сообщение a_1 кодируется этим кодом самым коротким кодовым словом, а на наименее вероятные сообщения a_3 и a_4 приходятся наиболее длинные кодовые слова.

Назовем **D-ичным деревом вероятностей** такое D-ичное дерево, в котором каждому узлу поставлены в соответствие вероятности так, что корню соответствует вероятность 1, а вероятность каждого внутреннего узла (включая корень) является суммой вероятностей узлов на глубине 1 в поддереве, выходящем из данного внутреннего узла.

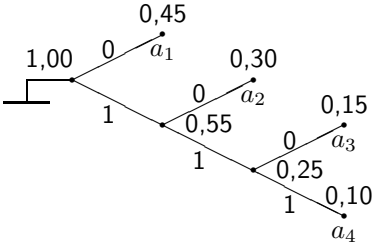
Заметим, что в соответствии с этим определением сумма вероятностей всех конечных узлов должна быть равна 1.

Теорема 3.3. *Средняя длина кодового слова может быть вычислена по дереву вероятностей как сумма вероятностей внутренних узлов, включая корень.*

Доказательство. Вероятность каждого внутреннего узла равна сумме вероятностей конечных узлов в поддереве, выходящем из данного внутреннего узла. Но конечный узел на глубине d входит в d таких

поддеревьев, соответствующих d внутренним узлам на пути от корня к этому конечному узлу. Значит, сумма вероятностей внутренних узлов равна сумме произведений вероятности каждого конечного узла и его глубины, что и представляет собой среднюю длину кодового слова. □

Пример 3.6. Дерево вероятностей для кода из примера 3.5 выглядит следующим образом:



Среднюю длину кодового слова можно вычислить по нему как $\overline{W} = \\ = 1{,}00 + 0{,}55 + 0{,}25 = 1{,}80$.

Теперь мы докажем следующий фундаментальный результат.

Теорема 3.4. *Средняя длина кодового слова оптимального кода с длинами кодовых слов, удовлетворяющих неравенству Крафта, для случайной величины U имеет границы*

$$\frac{H(U)}{\log D} \leq \overline{W} < \frac{H(U)}{\log D} + 1, \tag{3.8}$$

где знак равенства слева достигается в точности тогда, когда существуют такие целые положительные длины кодовых слов, что

$$w_i = -\log_D f(a_i), \;\; i = 1, \ldots, L. \tag{3.9}$$

Доказательство. Докажем сначала левое неравенство в (3.8). Учитывая (3.7), имеем

$$\overline{W} \log D = \sum_{i=1}^{L} f\left(a_i\right) w_i \log D = -\sum_{i=1}^{L} f\left(a_i\right) \log D^{-w_i}.$$

Рассмотрим разность

$$H(U) - \overline{W} \log D \;\; = \;\; -\sum_{i=1}^{L} f\left(a_i\right) \log f\left(a_i\right) + \sum_{i=1}^{L} f\left(a_i\right) \log D^{-w_i} =$$

$$= \sum_{i=1}^{L} f(a_i) \log \frac{D^{-w_i}}{f(a_i)} \leq$$

$$\leq \log \sum_{i=1}^{L} f(a_i) \frac{D^{-w_i}}{f(a_i)} = \log \sum_{i=1}^{L} D^{-w_i} \leq 0,$$

где первое неравенство следует из неравенства Йенсена, а второе неравенство следует из (3.1). Оба этих неравенства обращаются в точные равенства при $w_i = -\log_D f(a_i)$, $i = 1, \ldots, L$. Если хотя бы одна из величин w_i (3.9) не является целой, то нижняя граница в (3.8) не достигается. Поэтому левая часть в (3.8) справедлива. То есть в соответствии с теоремой 2.3 Шеннона о сжатии данных закодированный оптимальным кодом случайный процесс без памяти возможно полностью восстановить.

Для доказательства правой части (3.8) обозначим через $\lceil r \rceil$ наименьшее целое, большее либо равное r. Выберем длины кодовых слов как

$$w_i = \lceil -\log_D f(a_i) \rceil . \tag{3.10}$$

Мы имеем $w_i \geq -\log_D f(a_i)$, то есть $-w_i \leq \log_D f(a_i)$. Значит,

$$\sum_{i=1}^{L} D^{-w_i} \leq \sum_{i=1}^{L} D^{\log_D f(a_i)} = \sum_{i=1}^{L} f(a_i) = 1,$$

поэтому в соответствии с неравенством Крафта код с длинами слов (3.10) существует.

Нам необходимо еще доказать, что средняя длина его кодового слова $\overline{W}$ удовлетворяет неравенству в правой части (3.8).

Так как $w_i = \lceil -\log_D f(a_i) \rceil < -\log_D f(a_i) + 1$, имеем

$$\overline{W} = \sum_{i=1}^{L} f(a_i) w_i < \sum_{i=1}^{L} f(a_i) \left(-\log_D f(a_i) + 1\right) =$$

$$= \frac{-\sum_{i=1}^{L} f(a_i) \log f(a_i)}{\log D} + \sum_{i=1}^{L} f(a_i) = \frac{H(U)}{\log D} + 1. \qquad \square$$

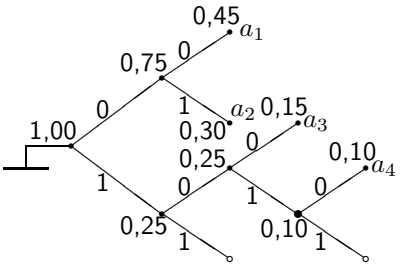
Заметим, что из (3.9) следует, что для оптимальных кодов более вероятным сообщениям (буквам алфавита) должны ставиться в соответствие менее длинные кодовые слова, а менее вероятным — более длинные.

Заметим также, что по доказательству достаточности неравенства Крафта (теорема 3.2), имея кодовые слова с длинами (3.10), мы всегда сможем построить D-ичный префиксный код, который удовлетворяет неравенству в правой части (3.8). Данный способ построения кода называется **алгоритмом Шеннона**.

Пример 3.7. Применим алгоритм Шеннона для построения двоичного префиксного кода для случайной величины U из примера 3.5:

$\mathcal{A}_U$	$f(a_i)$	$w_i = \lceil -\log_2 f(a_i) \rceil$
a_1	0,45	$\lceil -\log_2 0{,}45 \rceil = 2$
a_2	0,30	$\lceil -\log_2 0{,}30 \rceil = 2$
a_3	0,15	$\lceil -\log_2 0{,}15 \rceil = 3$
a_4	0,10	$\lceil -\log_2 0{,}10 \rceil = 4$

Построим дерево вероятностей для него следующим образом:



$\mathcal{A}_U$	$\mathcal{Z}$
a_1	00
a_2	01
a_3	100
a_4	1010

Средняя длина кодового слова для этого кода по теореме 3.3 $\overline{W} = 1 + 0{,}75 + 0{,}25 + 0{,}25 + 0{,}10 = 2{,}35$, что существенно хуже, чем для кода из примера 3.5, где $\overline{W} = 1{,}8$. Так как энтропия $H(U) = 1{,}782$, этот код тем не менее удовлетворяет неравенству в правой части (3.8)

$$\overline{W} < \frac{H(U)}{\log D} + 1 = 2{,}782.$$

Алгоритм Шеннона особенно эффективен, когда функция вероятности $f(u)$ такова, что получающаяся средняя длина кодового слова значительно больше единицы. В этом случае относительное влияние «лишней» единицы в правой части (3.8) будет невелико.

Наличие «лишней» единицы в правой части (3.8) связано с тем, что длины кодовых слов (3.9) не всегда являются целыми числами. Но мы можем уменьшить ее влияние, распределив его между многими сообщениями.

Действительно, объединим k сообщений некоторого источника в фразу и рассмотрим ее как одно сообщение из алфавита $\mathcal{A}_U^k$. Определим при этом **среднюю длину кодового слова k-го порядка** как

$$\overline{W}_k = \frac{1}{k} \sum_{\{u_1,u_2,\ldots,u_k\}} f(u_1,u_2,\ldots,u_k)\, w(u_1,u_2,\ldots,u_k), \tag{3.11}$$

где $w(u_1,u_2,\ldots,u_k)$ — длина кодового слова, соответствующего фразе $\{u_1,u_2,\ldots,u_k\}$, а $f(u_1,u_2,\ldots,u_k)$ — ее функция вероятности. Из (3.11) мы видим, что средняя длина кодового слова нового кода, аналогичная (3.7), равна

$$\overline{W} = k\overline{W}_k, \tag{3.12}$$

поэтому, применяя к новому коду (3.8), имеем

$$\frac{H(U_1,U_2,\ldots,U_k)}{\log D} \le k\overline{W}_k < \frac{H(U_1,U_2,\ldots,U_k)}{\log D} + 1. \tag{3.13}$$

Для источника без памяти величина $H(U_1,U_2,\ldots,U_k)/k$ вырождается в $H(U)$ (1.67), а для стационарного источника $\underline{U}$ она стремится к существующей абсолютной энтропии $H_\infty(\underline{U})$ (теорема 1.10). Поэтому из теоремы 3.4 и (3.13) получаем следующий результат.

Теорема 3.5. *Средняя длина кодового слова k-го порядка для оптимального кода с длинами кодовых слов, удовлетворяющих неравенству Крафта, имеет границы*

$$\frac{H_k(\underline{U})}{\log D} \le \overline{W}_k < \frac{H_k(\underline{U})}{\log D} + \frac{1}{k},$$

причем для источника без памяти

$$\frac{H(U)}{\log D} \le \overline{W}_k < \frac{H(U)}{\log D} + \frac{1}{k}, \tag{3.14}$$

а для произвольного стационарного источника

$$\lim_{k\to\infty} \overline{W}_k = \frac{H_\infty(\underline{U})}{\log D}. \tag{3.15}$$

Из (3.14) следует, что, увеличивая длину фразы, мы можем достичь сколь угодно близкой к энтропии средней длины кодового слова. Кроме того, мы таким образом учитываем влияние возможно имеющейся памяти случайного процесса $\underline{U}$ и достигаем оптимального кодирования марковских источников k-го порядка. Ценой таких достижений будет, однако, существенно более сложный анализ вероятностных характеристик источника и более сложное кодирование.

Предел в (3.15) позволяет интерпретировать абсолютную энтропию (1.63) как среднее количество D-ичных символов на выходе оптимального кодера источника, необходимое для представления одного сообщения стационарного случайного процесса $\underline{U}$.

Так же и относительной энтропии $H(f||g)$ (1.42) может быть дано толкование в терминах кодирования источника. Рассмотрим ситуацию, когда по алгоритму Шеннона сконструирован код, соответствующий некоторой функции вероятности $g(a)$, то есть с кодовыми словами длины $w_i = \lceil -\log_D g\,(a_i) \rceil$ (3.10), а используется этот код для кодирования случайной величины с некоторой другой функцией вероятности $f(a)$. Такая ситуация возможна, например, когда известна только приближенная функция вероятности подлежащих кодированию реальных данных.

Теорема 3.6. *Средняя длина кодового слова кода, сконструированного в соответствии с функцией вероятности $g(a)$ по алгоритму Шеннона и применяемого для кодирования случайной величины с функцией вероятности $f(a)$, лежит в пределах*

$$\frac{H(U) + H(f||g)}{\log D} \leq \overline{W} < \frac{H(U) + H(f||g)}{\log D} + 1. \tag{3.16}$$

Доказательство. Среднюю длину кодового слова можно вычислить как

$$\begin{aligned}
\overline{W} &= \sum_{i=1}^{L} f(a_i) \left\lceil \log_D \frac{1}{g(a_i)} \right\rceil < \sum_{i=1}^{L} f(a_i) \left(\log_D \frac{1}{g(a_i)} + 1 \right) = \\
&= \sum_{i=1}^{L} f(a_i) \log_D \left(\frac{1}{f(a_i)} \frac{f(a_i)}{g(a_i)} \right) + 1 = \\
&= \sum_{i=1}^{L} f(a_i) \log_D \frac{1}{f(a_i)} + \sum_{i=1}^{L} f(a_i) \log_D \frac{f(a_i)}{g(a_i)} + 1 = \\
&= \frac{H(U) + H(f||g)}{\log D} + 1.
\end{aligned}$$

Доказательство левой части (3.16) проводится аналогично. □

Таким образом, использование «неправильного» распределения $g(a)$ вместо $f(a)$ влечет за собой штраф в размере $H(f||g)$ D-ичных символов.

3.3. Алгоритм Хаффмена

Теперь построим метод конструирования оптимального двоичного неравномерного кода для заданной случайной величины U с алфавитом $\mathcal{A}_U = \{a_1, a_2, \ldots, a_L\}$, называемый **алгоритмом Хаффмена**. Общий D-ичный случай аналогичен двоичному (см. задачу 3.3.7).

Лемма 3.1. *Пусть отдельные сообщения упорядочены так, что их вероятности не возрастают: $f(a_1) \geq f(a_2) \geq \ldots \geq f(a_L)$. Тогда существует оптимальный двоичный префиксный код, длины кодовых слов которого не убывают, а двум наименее вероятным сообщениям соответствуют кодовые слова одинаковой длины: $w_1 \leq w_2 \leq \ldots \leq \leq w_{L-1} = w_L$.*

Доказательство. Предположим, кодовое слово для некоторого более вероятного сообщения длиннее кодового слова для некоторого менее вероятного сообщения, например $w_1 > w_2$. Тогда, приписав сообщению a_1 кодовое слово длины w_2, а сообщению a_2 — кодовое слово длины w_1, мы получим новый код с меньшей средней длиной кодового слова. Действительно, вычитая из средней длины кодового слова старого кода среднюю длину нового, мы получаем положительную величину:

$$\begin{aligned}&(f(a_1)w_1 + f(a_2)w_2) - (f(a_2)w_1 + f(a_1)w_2) = \\ &= (w_1 - w_2)\,(f(a_1) - f(a_2)) > 0,\end{aligned}$$

что невозможно в силу оптимальности.

Предположим далее, что $w_{L-1} \neq w_L$. Тогда согласно доказанному $w_{L-1} < w_L$. Следовательно, в кодовом дереве имеется единственное кодовое слово с длиной w_L. Но тогда в качестве конечного узла, соответствующего сообщению a_L, можно использовать предыдущий узел и тем самым уменьшить среднюю длину кодового слова, что снова противоречит условию оптимальности. □

Пусть теперь задан двоичный оптимальный префиксный код для случайной величины U, у которого двум наименее вероятным сообщениям соответствуют кодовые слова одинаковой длины $w_{L-1} = w_L$. По лемме 3.1 такой код существует. Построим новый код, объединив два наименее вероятных кодовых слова в одно, и образуем соответствующую ему новую случайную величину U' следующим образом:

$$f(a'_i) = \begin{cases} f(a_i), & i = 1, 2, \ldots, L-2, \\ f(a_{L-1}) + f(a_L), & i = L-1. \end{cases}$$

Используя теорему 3.3, для средней длины старого и нового кода получаем: $\overline{W} = \overline{W'} + f(a_{L-1}) + f(a_L)$, что означает, что средняя длина $\overline{W}$ старого кода минимальна в точности тогда, когда минимальна средняя длина $\overline{W'}$ нового кода. Таким образом справедлива следующая теорема.

Теорема 3.7. *Если a_{L-1} и a_L — два наименее вероятных значения U, то двоичный префиксный код для U оптимален в точности тогда, когда оптимален производный код для производной случайной величины U'.*

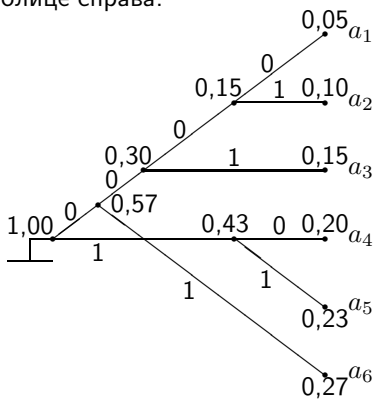
Поэтому мы доказали, что следующий рекурсивный алгоритм (алгоритм Хаффмена) конструирует оптимальный код:

Шаг 1. Пусть L конечными узлами дерева вероятностей являются сообщения $a_1, a_2, \ldots, a_L$. Поставим в соответствие каждому узлу a_i вероятность $f(a_i)$, $i = 1, 2, \ldots, L$ и будем считать все L узлов активными.

Шаг 2. Объединим два наименее вероятных активных узла в один двумя ветвями, одной из которых припишем значение 0, а другой — 1. Будем считать новый узел активным, а два его образовавших — неактивными.

Шаг 3. Если остался только один активный узел, то мы дошли до корня и на этом останавливаемся. Иначе возвращаемся к шагу 2.

Пример 3.8. Следующее дерево вероятностей соответствует построенному по алгоритму Хаффмена коду для случайной величины U, заданной в таблице справа:



$\mathcal{A}_U$	$f(a_i)$	$\mathcal{Z}$
a_1	0,05	0000
a_2	0,10	0001
a_3	0,15	001
a_4	0,20	10
a_5	0,23	11
a_6	0,27	01

Средняя длина кодового слова для этого кода в соответствии с теоремой 3.3 $\overline{W} = 1{,}00 + 0{,}57 + 0{,}43 + 0{,}30 + 0{,}15 = 2{,}45$, а энтропия $H(U) = 2{,}42$.

Отметим, что и код в примерах 3.5 и 3.6 мог бы быть построен по алгоритму Хаффмена, то есть алгоритмы Фано и Хаффмена дают в этом случае одинаковые результаты.

Как пишет Хаффмен, объединение отдельных сообщений напоминает последовательное слияние капель, ручейков и протоков в большую реку. Описанная процедура напоминает способ, которым пользуется водяное насекомое, делая отметки у каждого из этих слияний по мере движения вниз по течению. Насекомое должно запомнить код, чтобы проделать путь назад вверх по течению.

Алгоритм Хаффмена позволяет получить и оптимальную стратегию постановки вопросов, на которые можно отвечать только «да» или «нет», при поиске некоторого объекта среди многих, о которых известны вероятности их появления. Действительно, каждому такому объекту можно по алгоритму Хаффмена поставить в соответствие уникальное кодовое слово, а последовательность нулей и единиц в нем будет соответствовать последовательности ответов «да» или «нет». Среднее количество таких ответов (а значит, и вопросов, которые могут звучать, например, как: «Является такой-то бит в кодовом слове нулем или единицей?») равно средней длине кодового слова, близкой к энтропии (3.8), так как построенный по алгоритму Хаффмена код оптимален.

3.4. Алгоритм Танстелла

Неравномерное кодирование, которое мы до сих пор рассматривали, может иногда оказаться неудобным на практике. Например, если кодовые слова должны запоминаться в памяти компьютера, может быть предпочтительнее, чтобы их размер соответствовал постоянной длине компьютерного слова. Или, если передача кодовых слов должна вестись с постоянной скоростью (например, 2400 бит в секунду), при неравномерном кодировании необходимо дополнительно применять буфер, накапливающий кодовые слова.

Но как раз различие длин кодовых слов позволяло нам кодировать сообщения эффективно. Если потребовать, чтобы кодовые слова имели постоянную длину, то эффективное кодирование также возможно, но в этом случае эти кодовые слова будут соответствовать не отдельным сообщениям или фразам постоянной длины, а сообщениям переменной длины. Для того чтобы по кодовым словам было возможно однозначно реконструировать переданное сообщение, по аналогии

с неравномерным префиксным кодированием, необходимо, чтобы ни одно сообщение не было префиксом другого сообщения.

Итак, пусть все D-ичные кодовые слова имеют постоянную длину n, а сообщения источника — переменную длину, равную в среднем $\overline{K}$. Для минимизации среднего числа D-ичных символов, приходящихся на одно сообщение источника, равного $n/\overline{K}$, мы стремимся к максимизации $\overline{K}$.

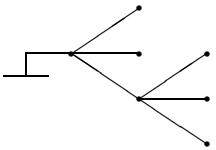
Рассмотрим L-ичный источник без памяти с алфавитом $\mathcal{A}_U = \{a_1, a_2, \ldots, a_L\}$ и вероятностями отдельных букв $f(a_i)$, $i=1,2,\ldots,L$.

Назовем L-ичное дерево, в котором из каждого узла выходят либо L, либо ноль ветвей, **комплектным деревом**, а множество сообщений, соответствующее конечным узлам этого дерева, — **комплектным множеством** сообщений.

Назовем **расщеплением** некоторого конечного узла, имеющего вероятность p в L-ичном дереве вероятностей, его превращение во внутренний узел посредством соединения с L новыми конечными узлами, имеющими вероятности $p \times f(a_i)$, $i = 1, 2, \ldots, L$. При этом сумма вероятностей новых конечных узлов будет равна вероятности вновь образовавшегося внутреннего узла, как того и требует определение дерева вероятностей.

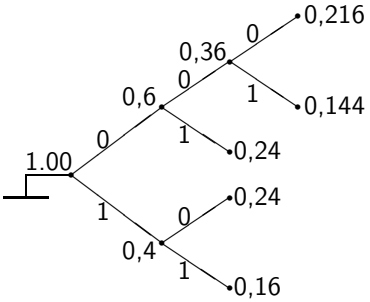
Построим теперь комплектное множество сообщений с помощью расщепления узлов L-ичного дерева. Поскольку, расщепляя некоторый конечный узел, мы удаляем из дерева один конечный узел и добавляем L конечных узлов, мы можем представить с помощью комплектного L-ичного дерева $M = L + q(L-1)$ сообщений, где q — некоторое положительное целое число.

Пример 3.9. Следующее комплектное дерево с $L = 3$ и $q = 1$ может представить $M = 5$ сообщений:



Назовем комплектное множество сообщений с $M = L + q(L-1)$ сообщениями **множеством Танстелла** для L-ичного источника без памяти, если соответствующее комплектное дерево может быть получено с помощью последовательного расщепления наиболее вероятного конечного узла.

Пример 3.10. Рассмотрим двоичный источник без памяти с вероятностью нуля 0,6 и вероятностью единицы 0,4. Проводя последовательные расщепления наиболее вероятных узлов, начиная с корневого, получаем однозначное множество Танстелла с $M = 5$ сообщениями:



При этом средняя длина сообщения в соответствии с теоремой 3.3 равна $\overline{K} = 1{,}0+0{,}4+0{,}6+0{,}36 = 2{,}36$ бит. Заметим, что в этом дереве имеется два узла с вероятностью $0{,}24$. Это значит, что имеется два множества Танстелла с $M = 6$.

Лемма 3.2. *Комплектное множество сообщений для L-ичного источника без памяти является множеством Танстелла в точности тогда, когда каждый внутренний узел в соответствующем комплектном дереве вероятностей не менее вероятен, чем каждый конечный узел.*

Доказательство. Рассмотрим комплектное множество сообщений, такое, что ни один конечный узел не вероятнее, чем внутренний. Такое дерево мы имеем всегда, когда мы удаляем L ветвей, которые исходят от наименее вероятного внутреннего узла. Если мы, действуя таким образом, дойдем до корня, а затем пойдем в обратном направлении, расщепляя на каждом шаге наиболее вероятный конечный узел, то получим множество Танстелла.

И наоборот, каждый внутренний узел как минимум так же вероятен, как и каждый конечный узел, так как множество Танстелла может быть получено последовательным расщеплением наиболее вероятных конечных узлов. □

Теорема 3.8. *Комплектное множество сообщений с M сообщениями из L-ичного источника без памяти максимизирует среднюю длину сообщения $\overline{K}$ в точности тогда, когда оно является множеством Танстелла.*

Доказательство. Комплектное множество сообщений с $M = L + q(L-1)$ сообщениями является по лемме 3.2 в точности тогда множеством Тунстелла, когда его $q+1$ внутренних узлов (включая корень) являются $q+1$ наиболее вероятными узлами в бесконечном дереве, которое бы соответствовало бесконечному потоку букв источника. Следовательно, сумма вероятностей внутренних узлов для всех множеств Тунстелла с M сообщениями будет одинаковая. Эта сумма всегда больше, чем соответствующая сумма для каждого множества, отличного от множества Тунстелла для M сообщений. Однако сумма вероятностей внутренних узлов равна по теореме 3.3 средней длине сообщения $\overline{K}$, поэтому доказательство ясно. □

Для D-ичного кода количество кодовых слов длины n равно D^n. Поэтому количество сообщений M не должно быть больше, чем D^n. Так как M возрастает на каждом шаге на $L-1$, наибольшее количество сообщений, которые мы можем закодировать, соответствует наибольшему целому числу q, такому, что

$$0 \leq D^n - M < L - 1.$$

Пусть q^* будет таким числом. Тогда

$$D^n - L = q^*(L-1) + r,$$

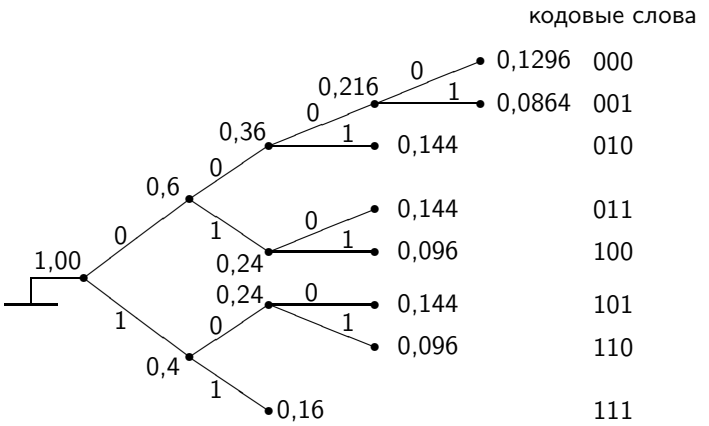
где $0 \leq r < L-1$, то есть q^* — это целая часть числа, получающегося от деления $D^n - L$ на $L-1$ (при этом предполагается, что $D^n \geq L$). Поэтому мы можем сформулировать **алгоритм Тунстелла** следующим образом:

Шаг 1. Определим $q^* = \left\lfloor \frac{D^n - L}{L-1} \right\rfloor$, где $\lfloor r \rfloor$ означает наибольшее целое, меньшее или равное r.

Шаг 2. Построить для заданного L-ичного источника без памяти множество Тунстелла с $M = L + q^*(L-1)$ элементами с помощью расщепления q^* раз наиболее вероятного узла, начиная с корня.

Шаг 3. Каждому сообщению поставить в соответствие однозначное D-ичное кодовое слово.

Пример 3.11. Пусть опять задан двоичный источник без памяти с вероятностью нуля 0,6 и вероятностью единицы 0,4, и пусть требуется построить равномерный код с длиной блока $n = 3$. По алгоритму Тунстелла получаем следующее комплектное дерево вероятностей:



Средняя длина сообщения по теореме 3.3 равна $\overline{K} = 1{,}0 + 0{,}4 + 0{,}6 + 0{,}36 + 0{,}24 + 0{,}24 + 0{,}216 = 3{,}056$ бит, что означает $n/\overline{K} = 0{,}982$ бит, приходящихся на одно сообщение источника.

Заметим, что алгоритм Танстелла базируется на комплектном множестве сообщений и в этом смысле является оптимальным. Если использовать некомплектное множество сообщений, то не исключено, что достижимы и лучшие результаты, однако для этого случая оптимальные алгоритмы не известны.

3.5. Арифметическое кодирование

Рассмотренные в предыдущих разделах алгоритмы кодирования источника являются статическими в том смысле, что оперируют фиксированными вероятностями символов источника. Такой подход может оказаться неэффективным, если алфавит источника имеет небольшую мощность. Например, для двоичного источника мы вынуждены использовать не менее одного бита на один символ, вне зависимости от энтропии источника, что заведомо плохо.

Поэтому естественной идеей является кодирование не отдельных символов, а фраз источника, что одновременно позволяет учесть и зависимости между его отдельными символами. Однако применение, например, алгоритма Хаффмена для фраз может оказаться неприемлемым из-за необходимости вычисления вероятностей всех фраз и построения соответствующего дерева, сложность чего растет экспоненциально с ростом длины фразы. Кроме того, мы в этом случае ограни-

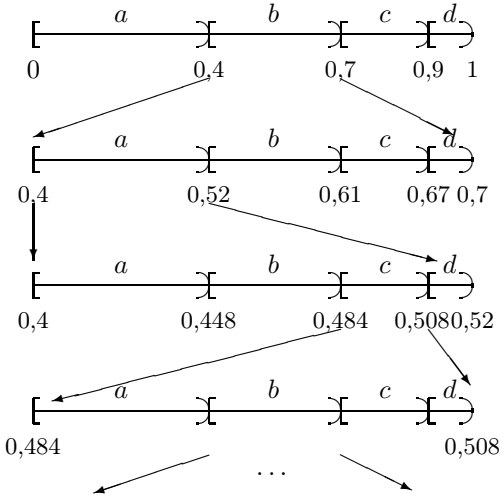


Рис. 3.2. Процесс арифметического кодирования

чены необходимостью использовать фразы только одной конкретной длины.

Описываемое в данном разделе арифметическое кодирование позволяет формировать код динамически по поступающим символам источника и является промежуточным вариантом между статическими алгоритмами и описанным в следующих разделах универсальным кодированием. Отличие от универсального кодирования состоит в том, что словарь кодовых слов остается статическим, как и в статических методах.

Поясним идею арифметического кодирования на примере. Пусть мы имеем источник с алфавитом $\mathcal{A}_U = \{a, b, c, d\}$, причем $f(a) = 0{,}4$, $f(b) = 0{,}3$, $f(c) = 0{,}2$ и $f(d) = 0{,}1$.

Предположим, нам надо закодировать последовательность символов источника $\underline{u} = \{b, a, c, b\}$. Разобьем интервал $[0; 1)$ на отрезки в соответствии с вероятностями символов источника, как показано сверху на рисунке 3.2. Поскольку первый символ, который надо закодировать, есть b, разобьем соответствующий b интервал $[0{,}4; 0{,}7)$ на подынтервалы таким же способом. При этом символу a (а значит, и последовательности $\{b, a\}$) будет соответствовать интервал $[0{,}4; 0{,}52)$. Действуя таким же способом и далее, найдем, что последовательности $\{b, a, c\}$ будет соответствовать интервал $[0{,}484; 0{,}508)$, а последовательности $\{b, a, c, b\}$ — $[0{,}4936; 0{,}5008)$.

Теперь необходимо найти эффективное представление интервала $A = [\alpha; \beta)$, соответствующего некоторой последовательности символов источника. Одним из методов может быть поиск такого наименьшего целого t, что $2^{-t} \leq l = \beta - \alpha$, а затем такого целого (четного, если есть несколько вариантов) m, что $\alpha \leq \frac{m}{2^t} < \beta$. Двоичное представление величины $r = \frac{m}{2^t}$ и будет искомым кодом данной последовательности. Разумеется, данную дробь следует предварительно сократить.

Попутно заметим, что длина интервала, соответствующего произвольной последовательности $\underline{u} = \{u_1, u_2, \ldots\}$, по нашему построению равна произведению вероятностей всех символов $\underline{u}$:

$$l = \prod_{i=1}^{n} f(u_i). \tag{3.17}$$

В нашем примере мы находим, что $t = 8$ и

$$0{,}4936 \leq \frac{m}{256} < 0{,}5008 \text{ или } 126{,}3616 \leq m < 128{,}2048.$$

Выбирая $m = 128$, получаем $r = \frac{128}{2^8} = \frac{1}{2} = (0{,}1)_2$ и код для $\{b, a, c, b\}$, равный 1, поскольку всегда $r < 1$, и первый ноль в его двоичном представлении можно отбросить.

Действуя таким же образом при кодировании, например, последовательности $\{c, c, d, a\}$, мы получим ее интервал $[0{,}876; 0{,}8776)$, $t = 10$, а $897{,}024 \leq m < 898{,}6624$ и

$$r = \frac{898}{2^{10}} = \frac{449}{512} = \frac{256 + 128 + 64 + 1}{512} = (0{,}111000001)_2.$$

Поэтому код для данной последовательности будет равен 111000001.

Теорема 3.9. *Пусть задан источник $\underline{U}$, состоящий из н.о.р. случайных величин с конечным алфавитом $\mathcal{A}_U$ и энтропией $H(U)$ каждая. Тогда для средней длины кодового слова, соответствующего последовательности символов источника длины n, при арифметическом кодировании справедливо*

$$\overline{W} \leq nH(U) + 1. \tag{3.18}$$

Доказательство. Поскольку мы ищем такое минимальное t, что $2^{-t} \leq l$, то для длины интервала справедливо $l < 2^{-(t-1)}$. Откуда имеем, что $t < \log_2 l^{-1} + 1$. Пусть $\underline{u}$ последовательность символов источника длины n. Тогда, обозначив через $t(\underline{u})$ длину кодового слова арифметического кода последовательности $\underline{u}$ и учитывая (3.17), имеем

$$\overline{W} = \sum_{\underline{u}} f(\underline{u})t(\underline{u}) < \sum_{\underline{u}} f(\underline{u}) \left(\log_2 \left(\prod_{i=1}^{n} f(u_i) \right)^{-1} + 1 \right) =$$

$$= \sum_{\underline{u}} f(\underline{u}) \sum_{i=1}^{n} \log_2 f(u_i)^{-1} + 1 = \sum_{i=1}^{n} \left(-\sum_{\underline{u}} f(\underline{u}) \log_2 f(u_i)) \right) + 1,$$

откуда следует (3.18), если учесть, что сумма всех $f(\underline{u})$ по последовательностям $\underline{u}$, не содержащим i-го элемента, равна $f(u_i)$. □

Таким образом, средняя длина кода, приходящаяся на один символ исходной последовательности при арифметическом кодировании, близка к энтропии источника, причем влияние на эту длину единицы в (3.18) минимально.

Заметим, что описанная процедура арифметического кодирования хорошо применима к источникам, у которых может быть просто вычислена функция $f(\underline{u})$. К таким относятся, например, источники, состоящие из н.о.р. случайных величин, и марковские источники. Причем данная процедура часто оказывается хорошей, даже если реальное распределение символов немного не соответствует вычисляемому (см. теорему 3.6), как это часто бывает при кодировании изображений.

3.6. Префиксное кодирование натуральных чисел

В данном разделе мы опишем **алгоритм Элайеса**, позволяющий строить префиксный код для двоичной записи натуральных чисел. Дело в том, что обычная двоичная запись $B(n)$ натуральных чисел n, представленная во втором столбце следующей таблицы, не обладает свойством префиксности.

n	$B(n)$	$\ell(n)$	$B_{Pref}(n)$	$B_{Pref}(\ell(n))$	$B_{Elias}(n)$
1	1	1	1	1	1
2	10	2	010	010	0100
3	11	2	011	010	0101
4	100	3	00100	011	01100
5	101	3	00101	011	01101
6	110	3	00110	011	01110
7	111	3	00111	011	01111
8	1000	4	0001000	00100	00100000

Например, двоичная запись числа 3 начинается с двоичной записи числа 1. Длина двоичного числа $B(n)$, равная $\ell(n) = \lfloor \log_2 n \rfloor + 1$, записана в третьем столбце таблицы.

Однако запись $B(n)$ может быть легко преобразована в префиксный код путем присоединения к ней префикса, состоящего из $\ell(n) - 1$ нулей. Получаемая в результате запись $B_{Pref}(n)$ показана в четвертом столбце таблицы. Действительно, по количеству начальных нулей можно однозначно определить длину следующей за ними двоичной записи числа n, а по ней восстановить само число. Но кодовые слова $B_{Pref}(n)$ имеют длину $\ell_{Pref}(n) = 2\lfloor \log_2 n \rfloor + 1$, что почти вдвое больше, чем $\ell(n)$.

Для уменьшения этой длины Элайес предложил скомбинировать коды $B(n)$ и $B_{Pref}(n)$ путем присоединения префикса $B_{Pref}(\ell(n))$, записанного в пятом столбце, к коду $B(n)$. Так как все кодовые слова $B(n)$ начинаются с единицы, то для сокращения общей длины она может быть опущена. Образованный таким образом код $B_{Elias}(n)$ показан в последнем столбце таблицы. Так как код $B_{Pref}(n)$ является префиксным и следующее количество бит определено однозначно, то очевидно, что $B_{Elias}(n)$ — тоже префиксный код.

На первый взгляд кажется абсурдным образовывать код $B_{Elias}(n)$ путем соединения кодовых слов $B(n)$ и $B_{Pref}(n)$, так как в результате кодовые слова становятся значительно длиннее. Но если внимательнее рассмотреть длину $\ell_{Elias}(n)$ кодовых слов $B_{Elias}(n)$, получим

$$\begin{aligned} \ell_{Elias}(n) &= \ell_{Pref}(\ell(n)) + \ell(n) - 1 = \\ &= \ell_{Pref}(\lfloor \log_2 n \rfloor + 1) + \lfloor \log_2 n \rfloor = \\ &= \lfloor \log_2 n \rfloor + 2\lfloor \log_2(\lfloor \log_2 n \rfloor + 1) \rfloor + 1. \end{aligned} \tag{3.19}$$

Так как член $2\lfloor \log_2(\lfloor \log_2 n \rfloor + 1) \rfloor$ для больших n пренебрежимо мал по сравнению с $\lfloor \log_2 n \rfloor$, то длина $\ell_{Elias}(n)$ увеличивается, по крайней мере для таких n, так же, как и $\ell(n)$, имеющая порядок $\log_2 n$. Это подтверждается рисунком 3.3, где изображены длины кодовых слов $\ell(n)$, $\ell_{Pref}(n)$ и $\ell_{Elias}(n)$ для различных натуральных чисел n. Заметим, что длина кодового слова $\ell_{Elias}(n)$ для $n \geq 16$ никогда не больше, чем $\ell_{Pref}(n)$. Кроме того, из рис. 3.3 видно, что скорость роста длины $\ell_{Pref}(n)$ существенно выше, чем у $\ell(n)$, в то время как $\ell_{Elias}(n)$ растет почти с той же скоростью, что и $\ell(n)$.

Еще один алгоритм построения двоичного префиксного кода для натуральных чисел, называемый алгоритмом Левенштейна, представлен в задаче 3.6.2.

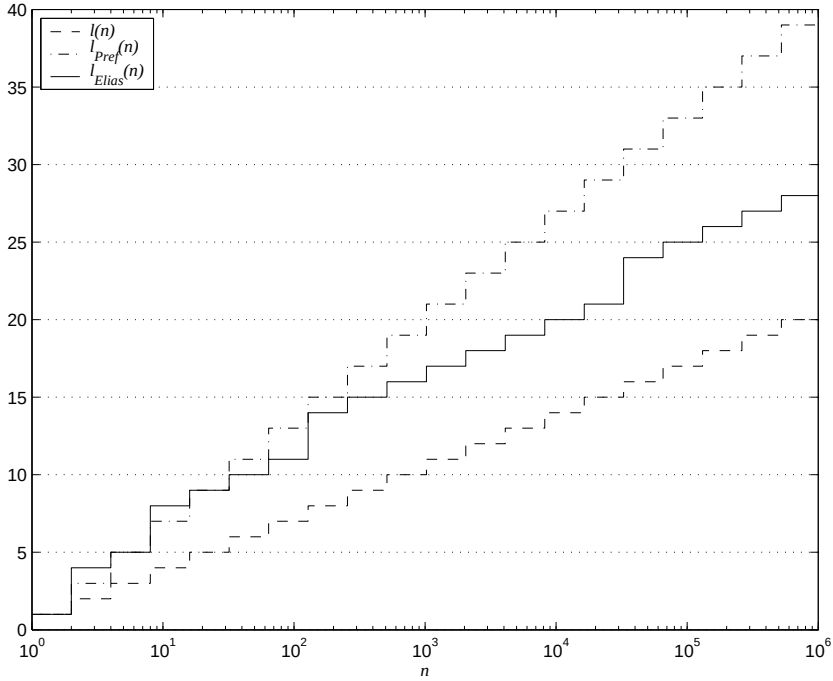


Рис. 3.3. Рост длин кодовых слов кодов $B(n)$, $B_{Pref}(n)$ и $B_{Elias}(n)$

3.7. Алгоритм Рябко – Элайеса

В этом разделе мы опишем **алгоритм Рябко – Элайеса**, который является примером универсального кодирования потока данных из дискретных стационарных источников.

Суть алгоритма Рябко – Элайеса заключается в том, что исходная (без существенного ограничения общности — двоичная) последовательность $\underline{u} = \{u_1, u_2, u_3, \ldots\}$ разделяется на фразы длины k, а затем составляется список всех фраз v_i, $i = 1, 2, \ldots, 2^k$. Кодер и декодер имеют в начальный момент времени этот список в известном им обоим порядке. При появлении в последовательности $\underline{u}$ очередной фразы v_j кодер формирует кодовое слово, равное номеру N_j этой фразы в списке, причем этот номер кодируется по алгоритму Элайеса, описанному в разделе 3.6. Вслед за этим кодер и декодер ставят фразу v_j в

начало списка, увеличивая этим на единицу номера фраз, стоящих на позициях от 1 до $j-1$. Позиции фраз, стоящих на позициях от $j+1$ и дальше, остаются неизменными.

Так происходит, например, со стопкой книг, если одну из них взять из стопки и переложить наверх. Поэтому алгоритм Рябко – Элайеса, по предложению Б. Я. Рябко в 1980 г., называется **сжатием с помощью стопки книг**. Другое название — **кодирование по степени новизны** — предложено в 1987 г. Элайесом.

При таком кодировании часто встречающиеся фразы получают в среднем малые номера и им соответствуют более короткие коды $B_{Elias}(N_j)$, в то время как редким фразам присваиваются в среднем большие номера и более длинные коды. Таким образом алгоритм Рябко – Элайеса анализирует внутреннюю статистику источника.

Пример 3.12. Предположим, что все фразы имеют длину $k = 2$ исходных символов, а исходная нумерация фраз соответствует возрастанию их значений: $[00] \to 1$, $[01] \to 2$, $[10] \to 3$, $[11] \to 4$. Тогда последовательность

$$10, 10, 01, 11, 11, ...$$

кодируется следующим образом. Вместо первой фразы 10 передается ее номер 3 (равный $B_{Elias}(3) = 0101$), а список фраз переупорядочивается: $10 \to 1$, $00 \to 2$, $01 \to 3$, $11 \to 4$. Вместо второй фразы 10 передается ее новый номер 1 (равный $B_{Elias}(1) = 1$), а список фраз остается без изменений. С приходом фразы 01 передается ее номер 3 (равный $B_{Elias}(3) = 0101$), а список фраз снова переупорядочивается: $01 \to 1$, $10 \to 2$, $00 \to 3$, $11 \to 4$. Фразе 11 соответствует номер 4 (равный $B_{Elias}(4) = 01100$) и новый список: $11 \to 1$, $01 \to 2$, $10 \to 3$, $00 \to 4$. Следующей фразе 11 соответствует уже номер 1 (равный $B_{Elias}(1) = 1$), а список остается без изменений. Таким образом, код, построенный по алгоритму Рябко – Элайеса, имеет следующий вид:

$$0101, 1, 0101, 01100, 1, ...,$$

где разделяющие запятые могут быть, конечно, опущены, так как код $B_{Elias}(n)$ является префиксным.

Чтобы оценить эффективность алгоритма Рябко – Элайеса, необходимо оценить среднюю длину $\overline{W}$ кодового слова порождаемого им кода. С целью простоты анализа предположим, что последовательность $\underline{u}$ эргодична, хотя это предположение не является необходимым и окончательный результат справедлив и для произвольных стационарных источников.

Обозначим через τ_i время, прошедшее после последнего появления фразы v_i. Так как, в отличие от определения номера фразы N_i в новом списке, для определения τ_i многократно встречающиеся одинаковые фразы учитываются так же многократно, то

$$N_i \le \tau_i. \tag{3.20}$$

Так как передаваемое i-е кодовое слово соответствует $B_{Elias}(N_i)$, то его длина равна $w_i = \ell_{Elias}(N_i)$ (3.19). Поэтому из (3.20) и (3.19) получаем

$$\begin{aligned} w_i &= \ell_{Elias}(N_i) \le \ell_{Elias}(\tau_i) = \\ &= \lfloor \log_2 \tau_i \rfloor + 2\lfloor \log_2(\lfloor \log_2 \tau_i \rfloor + 1)\rfloor + 1 \le \\ &\le \log_2 \tau_i + 2\log_2(\log_2 \tau_i + 1) + 1. \end{aligned} \tag{3.21}$$

Заметим, что если последовательность $\underline{u}$ порождена эргодическим источником, то эргодична и последовательность $\underline{v} = \{v_1, v_2, \ldots\}$. Поэтому среднее значение $\overline{\tau_i}$ может быть рассчитано по вероятности $f(v_i)$ появления фразы v_i:

$$\overline{\tau_i} = \frac{1}{f(v_i)}. \tag{3.22}$$

То есть для среднего значения длины i-го кодового слова $\overline{w_i}$ из (3.21) и неравенства Йенсена получаем следующую оценку:

$$\overline{w_i} \le \log_2 \overline{\tau_i} + 2\log_2\log_2(\overline{\tau_i} + 1) + 1,$$

откуда, учитывая (3.22), имеем

$$\overline{w_i} \le -\log_2 f(v_i) + 2\log_2(-\log_2 f(v_i) + 1) + 1. \tag{3.23}$$

Для определения $\overline{W}$ домножим (3.23) с двух сторон на $f(v_i)$ и просуммируем по всем фразам:

$$\begin{aligned} \overline{W} &= \sum_i f(v_i)\overline{w_i} \le \\ &\le H(V) + 2\sum_i f(v_i)\log_2(-\log_2 f(v_i) + 1) + 1 \le \\ &\le H(V) + 2\log_2(H(V) + 1) + 1, \end{aligned}$$

где V — случайная величина, связанная с фразами v_i, первое неравенство следует из (3.23), а второе — из неравенства Йенсена.

Наконец, учитывая, что $H(V) = H(U_1, \ldots, U_k) = kH_k(\underline{U})$ (1.62) и $\overline{W}_k = \overline{W}/k$ (3.12), получаем, что справедлива следующая теорема.

Теорема 3.10. *Средняя длина кодового слова k-го порядка для кода, построенного по алгоритму Рябко – Элайеса, имеет следующую верхнюю границу:*

$$\overline{W}_k \le H_k(\underline{U}) + \frac{2}{k}\log_2(kH_k(\underline{U}) + 1) + \frac{1}{k},$$

причем $\lim\limits_{k\to\infty} \overline{W}_k = H_\infty(\underline{U})$*, и в соответствии с теоремой 3.5 кодирование по алгоритму Рябко – Элайеса асимптотически оптимально.*

3.8. Алгоритм Лемпеля – Зива

Теперь мы опишем еще один алгоритм универсального кодирования источника, разработанный Лемпелем и Зивом в 1978 г. и называемый в их честь **алгоритмом Лемпеля – Зива**. Этот алгоритм очень прост, легко реализуем и тем не менее асимптотически оптимален для эргодических источников, так как производит их сжатие до абсолютной энтропии. Благодаря своей скорости и эффективности он включен (в различных модификациях) во многие программы сжатия компьютерных данных. Имеются также микросхемные реализации алгоритма Лемпеля – Зива.

Мы ограничимся рассмотрением только двоичного источника, поскольку результаты легко обобщаются на любой конечный алфавит.

Назовем **уникальным разделением** двоичной последовательности $\underline{u}$ ее такое разделение на фразы, отделенные друг от друга запятыми, когда нет даже пары одинаковых фраз. Например, 0,111,1 — это уникальное разделение последовательности 01111, а 0,11,11 — не уникальное.

Суть алгоритма Лемпеля – Зива заключается в том, что исходная последовательность $\underline{u} = \{u_1, u_2, \ldots, u_n\}$ разделяется на фразы, которые до сих пор не встречались. Например, если дана последовательность

$$1011010100010\ldots,$$

то после разделения мы получим

$$\overset{1}{1},\overset{2}{0},\overset{3}{11},\overset{4}{01},\overset{5}{010},\overset{6}{00},\overset{7}{10},\ldots,$$

где каждая новая фраза имеет свой однозначный порядковый номер. Такое разделение является уникальным и называется **разделением**

Лемпеля – Зива. В соответствии с этой конструкцией каждая последующая фраза состоит из некоторой (уже встречавшейся) предыдущей фразы как префикса и еще одного отличительного конечного бита. Поэтому закодировать фразы можно, указывая номер префикса и конечный бит, а восстановить — добавляя конечный бит к известному префиксу.

1	0	11	01	010	00	10	...	фразы после разделения
↓	↓	↓	↓	↓	↓	↓		
(0,1)	(0,0)	(1,1)	(2,1)	(4,0)	(2,0)	(1,0)	...	кодовые слова
↓	↓	↓	↓	↓	↓	↓		
1	0	11	01	010	00	10	...	фразы после реконструкции

Обозначим через $c(n)$ количество фраз в разделении Лемпеля – Зива входной последовательности $\underline{u}$ длины n. Для задания префикса его двоичной записью нам потребуется $\lfloor \log c(n) \rfloor + 1$ бит, для каждой фразы потребуется $\lfloor \log c(n) \rfloor + 2$ бит, а суммарная длина сжатой последовательности в битах будет равна

$$\ell_{LZ}(\underline{u}) = c(n)(\lfloor \log c(n) \rfloor + 2). \tag{3.24}$$

Рассмотренный алгоритм требует двух проходов по последовательности, поскольку сначала необходимо определить число фраз $c(n)$ и узнать количество бит, необходимое для задания префикса. При втором проходе мы непосредственно создаем кодовые слова, затрачивая на каждый префикс одинаковое количество бит $\lfloor \log c(n) \rfloor + 1$. Поэтому выполнения свойства префиксности в данном случае не требуется. Если размер префикса сделать меньшим в начале последовательности и динамически увеличивать по мере надобности, то потребуется только один проход по последовательности и будет использовано меньше бит для их задания. Кроме того, для записи префикса можно использовать алгоритм Элайеса из раздела 3.6. Получаемый при этом код станет префиксным. Однако такие модификации не отразятся на асимптотической эффективности алгоритма Лемпеля – Зива.

Лемма 3.3. **(Неравенство Лемпеля – Зива):** *Число фраз $c(n)$ в уникальном разделении двоичной последовательности $\underline{u}$ длины n удовлетворяет неравенству*

$$c(n) \le \frac{n}{(1-\varepsilon_n)\log n}, \tag{3.25}$$

где $\varepsilon_n \to 0$ при $n \to \infty$.

Доказательство. Пусть n_k — это сумма длин всех различных фраз с длиной, не большей k, то есть

$$n_k = \sum_{j=1}^{k} j2^j = (k-1)2^{k+1} + 2.$$

Если $n = n_k$, то число фраз $c(n_k)$ в уникальном разделении последовательности с длиной n_k не превзойдет общего количества фраз, то есть

$$c(n_k) \le \sum_{j=1}^{k} 2^j = 2^{k+1} - 2 < 2^{k+1} < \frac{n_k}{k-1}.$$

Если $n_k \le n < n_{k+1}$, то пусть $n = n_k + \delta$, где $\delta < n_{k+1} - n_k = (k+1)2^{k+1}$. Тогда разделение Лемпеля – Зива дает $c(n_k)$ фраз с длиной не больше k и $\delta/(k+1)$ фраз с длиной $k+1$. Таким образом,

$$c(n) \le \frac{n_k}{k-1} + \frac{\delta}{k+1} \le \frac{n_k + \delta}{k-1} = \frac{n}{k-1}. \tag{3.26}$$

Теперь оценим величину k для заданного n. Пусть опять $n_k \le n < n_{k+1}$. Тогда $n \ge n_k = (k-1)2^{k+1} + 2 \ge 2^k$, то есть

$$k \le \log n. \tag{3.27}$$

Более того, $n < n_{k+1} = k2^{k+2} + 2 < (k+2)2^{k+2} \le (\log n + 2)2^{k+2}$, где в последнем неравенстве учтено (3.27). Поэтому

$$k + 2 \ge \log \frac{n}{\log n + 2},$$

откуда для всех $n \ge 4$:

$$\begin{aligned} k - 1 &\ge \log n - \log(\log n + 2) - 3 = \\ &= \left(1 - \frac{\log(\log n + 2) + 3}{\log n}\right) \log n = \\ &= (1 - \varepsilon_n) \log n, \end{aligned} \tag{3.28}$$

где $\varepsilon_n \to 0$ при $n \to \infty$. Комбинируя (3.28) с (3.26), мы получаем результат леммы. □

Пусть $\underline{u} = \{u_{1-k}, u_{2-k}, \ldots, u_0, u_1, u_2, \ldots, u_n\}$ является реализацией случайного процесса $\underline{U}$, а последовательность $\{u_1, u_2, \ldots, u_n\}$ разделена на $c(n)$ различных фраз $\{\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_{c(n)}\}$.

Пусть далее ν_i будет индексом начала i-й фразы ($i = 1, 2, \ldots, c(n)$) в разделении $\underline{u}$, то есть $\underline{v}_i = \{u_{\nu_i}, u_{\nu_i+1}, \ldots, u_{\nu_{i+1}-1}\}$. Для каждого i

обозначим k бит $\underline{u}$, предшествующих $\underline{v}_i$, через $\underline{s}_i = \{u_{\nu_i - k}, u_{\nu_i - k + 1}, \ldots, u_{\nu_i - 1}\}$. При этом $\underline{s}_1 = \{u_{1-k}, u_{2-k}, \ldots, u_0\}$, а $\underline{u} = \{\underline{s}_1, \underline{v}_1, \underline{v}_2, \ldots, \underline{v}_{c(n)}\}$.

Обозначим через $c_{l,\underline{s}}$ число фраз $\underline{v}_i$ с длиной l $(l = 1, 2, \ldots)$ и предшествующими k битами $\underline{s}_i = \underline{s} \in \mathcal{A}_U^k$. Заметим, что

$$\sum_{l,\underline{s}} c_{l,\underline{s}} = c(n) \quad \text{и} \quad \sum_{l,\underline{s}} l c_{l,\underline{s}} = n. \tag{3.29}$$

Лемма 3.4. *Для некоторого уникального разделения (в частности, разделения Лемпеля – Зива) последовательности $\underline{u}$ справедливо* **неравенство Зива:**

$$\log Q_k(\underline{u}) - \log f(\underline{s}_1) \le -\sum_{l,\underline{s}} c_{l,\underline{s}} \log c_{l,\underline{s}}, \tag{3.30}$$

где $Q_k(\underline{u})$ является марковской аппроксимацией k-го порядка (1.80) для функции вероятности $f(\underline{u})$.

Доказательство. По определению $Q_k(\underline{u})$ (1.80) имеем

$$Q_k(\underline{u}) = Q_k(\underline{s}_1, \underline{v}_1, \underline{v}_2, \ldots, \underline{v}_{c(n)}) = f(\underline{s}_1) \prod_{i=1}^{c(n)} f(\underline{v}_i | \underline{s}_i),$$

поэтому

$$\begin{aligned}
\log Q_k(\underline{u}) - \log f(\underline{s}_1) &= \sum_{i=1}^{c(n)} \log f(\underline{v}_i | \underline{s}_i) = \\
&= \sum_{l,\underline{s}} \sum_{i:\ \ell(\underline{v}_i)=l, \underline{s}_i = \underline{s}} \log f(\underline{v}_i | \underline{s}_i) = \\
&= \sum_{l,\underline{s}} c_{l,\underline{s}} \sum_{i:\ \ell(\underline{v}_i)=l, \underline{s}_i = \underline{s}} \frac{1}{c_{l,\underline{s}}} \log f(\underline{v}_i | \underline{s}_i) \le \\
&\le \sum_{l,\underline{s}} c_{l,\underline{s}} \log \left(\sum_{i:\ \ell(\underline{v}_i)=l, \underline{s}_i = \underline{s}} \frac{1}{c_{l,\underline{s}}} f(\underline{v}_i | \underline{s}_i) \right),
\end{aligned}$$

где неравенство следует из неравенства Йенсена и выпуклости логарифма.

Так как все $\underline{v}_i$ различны, мы имеем $\sum_{i:\ \ell(\underline{v}_i)=l, \underline{s}_i = \underline{s}} f(\underline{v}_i | \underline{s}_i) \le 1$. Таким образом,

$$\log Q_k(\underline{u}) - \log f(\underline{s}_1) \le \sum_{l,\underline{s}} c_{l,\underline{s}} \log \frac{1}{c_{l,\underline{s}}},$$

что доказывает неравенство Зива. □

Теорема 3.11. *Пусть $\underline{U}$ — стационарный эргодический случайный процесс с абсолютной энтропией $H_\infty(\underline{U})$, и пусть $c(n)$ является числом фраз в уникальном разделении реализации $\underline{u}$ длины n этого процесса. Тогда*

$$\lim_{n\to\infty} \sup \frac{\ell_{LZ}(\underline{u})}{n} = \lim_{n\to\infty} \sup \frac{c(n)\log c(n)}{n}, \tag{3.31}$$

а

$$\operatorname*{plim}_{n\to\infty} \sup \frac{c(n)\log c(n)}{n} \le H_\infty(\underline{U}). \tag{3.32}$$

Доказательство. Равенство (3.31) является простым следствием (3.24) и неравенства Лемпеля – Зива (3.25). Действительно, из (3.24) имеем

$$\lim_{n\to\infty} \sup \frac{\ell_{LZ}(\underline{u})}{n} = \lim_{n\to\infty} \sup \left(\frac{c(n)\log c(n)}{n} + \frac{c(n)}{n} \right),$$

а $\lim\limits_{n\to\infty} \sup c(n)/n = 0$ по (3.25).

Для доказательства неравенства (3.32) перепишем неравенство Зива как

$$\begin{aligned} \log Q_k(\underline{u}) - \log f(\underline{s}_1) &\le -\sum_{l,\underline{s}} c_{l,\underline{s}} \log \frac{c_{l,\underline{s}} c(n)}{c(n)} = \\ &= -c(n)\log c(n) - c(n) \sum_{l,\underline{s}} \frac{c_{l,\underline{s}}}{c(n)} \log \frac{c_{l,\underline{s}}}{c(n)}. \end{aligned} \tag{3.33}$$

Обозначим $\pi_{l,\underline{s}} = \frac{c_{l,\underline{s}}}{c(n)}$. Тогда из (3.29) получаем

$$\sum_{l,\underline{s}} \pi_{l,\underline{s}} = 1 \quad \text{и} \quad \sum_{l,\underline{s}} l\pi_{l,\underline{s}} = \frac{n}{c(n)}.$$

Теперь определим случайные величины V и S так, что $f(V = l, S = \underline{s}) = \pi_{l,\underline{s}}$. Таким образом, из $\overline{V} = \frac{n}{c(n)}$ и (3.33) имеем

$$\log Q_k(\underline{u}) - \log f(\underline{s}_1) \le c(n)H(V,S) - c(n)\log c(n)$$

или

$$-\frac{1}{n}\log Q_k(\underline{u}) + \frac{1}{n}\log f(\underline{s}_1) \ge \frac{c(n)\log c(n)}{n} - \frac{c(n)}{n} H(V,S). \tag{3.34}$$

Покажем, что $\frac{c(n)}{n}H(V,S) \to 0$ при $n \to \infty$. В соответствии с диаграммой Эйлера (рис. 1.2)

$$H(V,S) \le H(V) + H(S) \le H(V) + k,$$

где второе неравенство следует из того, что последовательность s состоит из k бит.

Так как нам известно, что V — неотрицательная случайная величина с заданным математическим ожиданием $\overline{V} = \frac{n}{c(n)}$, то в соответствии с результатами раздела 1.5 (пример 1.9, равенство (1.55)) ее максимальная энтропия, достигаемая на геометрическом распределении, равна $(\overline{V}+1)\log(\overline{V}+1) - \overline{V}\log\overline{V}$, поэтому

$$\begin{aligned} H(V) &\le \left(\frac{n}{c(n)}+1\right)\log\left(\frac{n}{c(n)}+1\right) - \frac{n}{c(n)}\log\frac{n}{c(n)} = \\ &= \log\frac{n}{c(n)} + \left(\frac{n}{c(n)}+1\right)\log\left(\frac{c(n)}{n}+1\right). \end{aligned}$$

Таким образом,

$$\frac{c(n)}{n}H(V,S) \le \frac{c(n)}{n}k + \frac{c(n)}{n}\log\frac{n}{c(n)} + \left(\frac{c(n)}{n}+1\right)\log\left(\frac{c(n)}{n}+1\right). \tag{3.35}$$

Однако из неравенства Лемпеля – Зива (лемма 3.3) в форме $\frac{c(n)}{n} \le \frac{1}{(1-\varepsilon_n)\log n}$ следует, что $\frac{c(n)}{n} \to 0$ при $n \to \infty$, поэтому каждое слагаемое в (3.35) стремится к нулю при $n \to \infty$. То есть из (3.34) имеем

$$\frac{c(n)\log c(n)}{n} \le -\frac{1}{n}\log Q_k(\underline{u}) + \frac{1}{n}\log f(\underline{s}_1) + \varepsilon_k(n),$$

где $\varepsilon_k(n) \to 0$ и $\frac{1}{n}\log f(\underline{s}_1) \to 0$ при $n \to \infty$. Отсюда

$$\operatorname*{plim\,sup}_{n\to\infty} \frac{c(n)\log c(n)}{n} \le \operatorname*{plim}_{n\to\infty} -\frac{1}{n}\log Q_k(\underline{u}) = H_\infty(\underline{U}),$$

что для достаточно большого k следует из эргодичности и стационарности случайного процесса $\underline{U}$ (лемма 1.1, равенство (1.83)).

Таким образом, средняя длина кодового слова сжатых по алгоритму Лемпеля – Зива реализаций эргодического случайного процесса асимптотически не больше, чем абсолютная энтропия этого процесса. То есть алгоритм Лемпеля – Зива асимптотически оптимален. □

3.9. Задачи

3.1.1. [3 балла] Определите, какие из перечисленных кодов

U	$\mathcal{Z}_1$	$\mathcal{Z}_2$	$\mathcal{Z}_3$	$\mathcal{Z}_4$	$\mathcal{Z}_5$	$\mathcal{Z}_6$	$\mathcal{Z}_7$
a_1	0	0	0	0	10	00	00
a_2	01	10	01	01	11	01	10
a_3	–	11	11	10	110	10	11
a_4	–	–	–	–	–	11	100
a_5	–	–	–	–	–	–	110

обладают свойством однозначного декодирования и для каждого такого кода сконструируйте, если возможно, бесконечную последовательность кодовых слов, которую можно разбить на исходные сообщения двумя (!) различными способами (что не противоречит свойству однозначной декодируемости). Докажите, что подобные последовательности невозможны в префиксных кодах.

3.1.2. [1 балл] При каких условиях достигается знак равенства в неравенстве Крафта?

3.1.3. [2 балла] Какое максимальное количество сообщений может быть передано с помощью двоичного префиксного кода с максимальной длиной кодового слова $w_{\max}$?

3.2.1. (а) [2 балла] Сообщения источника a_1, a_2, a_3, a_4, a_5 имеют вероятности 0,3, 0,2, 0,2, 0,15, 0,15 соответственно. Эти сообщения должны быть закодированы двоичным префиксным кодом, причем известно, что передача нуля занимает 1 секунду, а передача единицы — 3 секунды. Найдите код, который минимизирует среднее время, требуемое для передачи одного сообщения, и вычислите это время.

(б) [3 балла] Любой такой код может быть представлен деревом, в котором длина ребра пропорциональна времени, необходимому для передачи соответствующего сообщения. Покажите, что у дерева кода, минимизирующего среднее время передачи, вероятности, соответствующие промежуточным и концевым узлам, не должны увеличиваться с увеличением длины.

3.2.2. [3 балла] Найдите пример такой случайной величины U, что для любого $\varepsilon > 0$ оптимальный двоичный код имеет среднюю длину $\overline{W} > H(U) + 1 - \varepsilon$.

3.2.3. [4 балла] Рассмотрите проблему поиска оптимального однозначно декодируемого кода как стандартную проблему минимизации

его средней длины по всем длинам кодовых слов, удовлетворяющих неравенству Крафта. Выведите из этого рассмотрения алгоритм Шеннона.

3.2.4. [3 балла] Докажите, что для кодов, построенных по алгоритму Фано, справедливо $\overline{W} \le \frac{H(U)}{\log D} + 2$.

3.2.5. [5 баллов] Докажите, что средняя длина кодового слова произвольного двоичного инъективного кода для случайной величины U с мощностью алфавита L удовлетворяет неравенству $\overline{W} \ge H(U) - \log\log L - C$, где C — не зависящая от L константа. Данный результат означает, что отказ от требования однозначной декодируемости приводит лишь к выигрышу порядка $\log\log L$ по средней длине кодового слова.

3.2.6. [5 баллов] Обозначим через $[x]$ расстояние от числа x до ближайшего к нему целого числа. Докажите, что средняя длина кодового слова однозначно декодируемого двоичного кода для случайной величины U с алфавитом $\{a_1, a_2, \ldots, a_L\}$ и функцией вероятности $f(a_i)$, $i = 1, 2, \ldots, L$ удовлетворяет неравенству

$$\overline{W} \ge H(U) + \frac{\ln 2}{2\sqrt{2}} \sum_{i=1}^{L} f(a_i) \left[\log \frac{1}{f(a_i)}\right]^2 .$$

Данный результат усиливает левое неравенство в (3.8).

3.3.1. [2 балла] Постройте по алгоритму Хаффмена код для источника, генерирующего сообщения a_1, a_2, a_3, a_4, a_5 с вероятностями $1/3$, $1/5$, $1/5$, $2/15$, $2/15$ соответственно. Обоснуйте, что этот код оптимален и для источника с множеством вероятностей $1/5$, $1/5$, $1/5$, $1/5$, $1/5$.

3.3.2. [2 балла] Постройте по алгоритму Хаффмена код для источника, генерирующего сообщения a_1, a_2, a_3, a_4 с вероятностями $1/3$, $1/3$, $1/4$, $1/12$ соответственно. Покажите, что имеется два различных множества длин кодовых слов $\{1, 2, 3, 3\}$ и $\{2, 2, 2, 2\}$, для которых существуют оптимальные коды. Покажите, что существуют оптимальные коды, длины некоторых кодовых слов которых превосходят длины кодовых слов (3.10) кодов, конструируемых по алгоритму Шеннона.

3.3.3. [2 балла] Какие из следующих кодов не могут быть построены по алгоритму Хаффмена ни для каких источников?

(а) $\{0, 10, 11\}$

(б) $\{00, 01, 10, 110\}$

(в) $\{01, 10\}$

3.3.4. [3 балла] Найдите такое распределение вероятностей $f(a)$, при котором алгоритм Фано строит код с большей средней длиной кодового слова, чем алгоритм Хаффмена.

3.3.5. [3 балла] Покажите, что количество всех операций (сложения, объединения и шагов при переупорядочивании) в алгоритме Хаффмена не более чем $M \cdot L^2$, где M — не зависящая от мощности алфавита L константа.

3.3.6. [2 балла] Пусть игрок А выбирает некоторый объект во Вселенной, а игрок Б пытается идентифицировать этот объект с помощью серии вопросов, на которые даются ответы «да» или «нет». Предположим, игрок Б достаточно умен, чтобы использовать оптимальный код, соответствующий распределению объектов. Мы наблюдаем, что игроку Б требуется в среднем 38,5 вопросов для идентификации объекта. Найдите нижнюю границу для количества объектов во Вселенной.

3.3.7. [4 балла] Обобщите алгоритм Хаффмена на D-ичный случай и докажите его оптимальность.

3.3.8. [4 балла] Рассмотрите двоичный код, построенный по алгоритму Хаффмена для источника, генерирующего m равновероятных сообщений. Определите длины кодовых слов, количество кодовых слов с различными длинами и среднюю длину кодового слова. Для каких значений m средняя длина кодового слова равна энтропии?

3.3.9. [4 балла] Докажите, что сумма длин L кодовых слов двоичного кода, построенного по алгоритму Хаффмена, лежит в пределах

$$L \log L \leq \sum_{i=1}^{L} w_i \leq (L^2 + L - 2)/2.$$

3.4.1. [5 баллов] Докажите, что отношение длины n D-ичных кодовых слов к средней длине сообщения $\overline{K}$ в комплектном множестве сообщений при оптимальном равномерном кодировании L-ичного источника без памяти, описываемого случайной величиной U с алфавитом $\mathcal{A}_U = \{a_1, a_2, \ldots, a_L\}$, лежит в пределах

$$\frac{H(U)}{\log D} \leq \frac{n}{\overline{K}} < \frac{H(U)}{\log D} + \frac{\log\left(2/p_{\min}\right) \log L}{(n \log D - \log L) \log D},$$

где $p_{\min} = \min\limits_i f(a_i) > 0$.

3.4.2. [3 балла] Рассмотрим источник без памяти, на выходе которого имеется единица с вероятностью 0,3, а ноль — с вероятностью 0,7. Объедините тройки последовательных сообщений в новое сообщение и постройте код по алгоритму Хаффмена для новых сообщений. Определите среднюю длину кодового слова третьего порядка для такого кода. Постройте также код с длинами кодовых слов, равными трем по алгоритму Танстелла для исходного источника, и определите среднее количество кодовых слов, приходящихся на одно сообщение источника. Сравните результаты.

3.5.1. [2 балла] Закодируйте с помощью арифметического кода последовательность $\{b, c, b, a, b\}$, если $f(a) = 0{,}1$, $f(b) = 0{,}6$ и $f(c) = 0{,}3$.

3.5.2. [5 баллов] Разработайте метод декодирования арифметического кода.

3.6.1. [2 балла] Проверьте выполнение неравенства Крафта для кодов $B_{Pref}(n)$ и $B_{Elias}(n)$.

3.6.2. [5 баллов] При построении префиксного кода для натурального числа n по алгоритму Элайеса мы к двоичной записи (без первой единицы) этого числа приписывали код $B_{Pref}(\ell(n))$ от его длины. Переоткройте другой алгоритм префиксного кодирования натуральных чисел, называемый **алгоритмом Левенштейна**, основываясь на следующей идее. Припишем к двоичной записи без первой единицы $B'(n)$ числа n слева запись $B'(\lfloor \log_2 n \rfloor)$, затем еще левее запись $B'(\lfloor \log_2 \log_2 n \rfloor)$ и так далее всего k раз. Затем поставим 0 и k единиц. Чему должно быть равно k, чтобы код был префиксным? Докажите, что для длины построенного по алгоритму Левенштейна кода $B_{Lev}(n)$ справедлива оценка $\ell_{Lev}(n) = \log_2 n + \log_2 \log_2 n(1 + O(1))$. Проверьте выполнение неравенства Крафта для $B_{Lev}(n)$. Какому натуральному числу соответствует $B_{Lev}(n) = 111100100010111101001$? Начиная с какого n алгоритм Левенштейна дает код меньшей длины, чем алгоритм Элайеса?

Указание: $B_{Lev}(1) = 10$, $B_{Lev}(2) = 1100$, $B_{Lev}(4) = 1110000$, $B_{Lev}(8) = 11101000$, $B_{Lev}(16) = 111100000000$.

3.7.1. [2 балла] Закодируйте с помощью алгоритма Рябко – Элайеса последовательность 0000001101010000011010100110 1 при длинах фраз 2 и 3. Сравните результаты.

3.8.1. [2 балла] Закодируйте с помощью алгоритма Лемпеля – Зива последовательность 000000110101000001101010011 01.

Глава 4

Теоретико-информационные основы криптологии

4.1. Классические системы шифрования

Криптологию можно с одинаковым правом назвать как наукой, так и искусством конструирования и взламывания систем защиты информации. Как искусство она уже несколько тысяч лет применяется военными, дипломатами, шпионами. Как наука она находится в достаточно молодом возрасте и берет начало от статьи Шеннона 1949 г. [2].

Криптология подразделяется на **криптографию**, занимающуюся тем, как зашифровать сообщение, **криптоанализ**, занимающийся тем, как взломать (атаковать) зашифрованное, и **теорию аутентичности**, занимающуюся тем, как защитить передаваемое сообщение от несанкционированных изменений.

На рис. 4.1 изображена структурная схема классической криптосистемы. Для предотвращения несанкционированного доступа к сообщению $\underline{M}$ его шифруют, получая **криптограмму** $\underline{C} = E_{\underline{K}}(\underline{M})$. Данная криптограмма посылается получателю или же запоминается. Способ шифрования $E_{\underline{K}}$ называется **шифром**. Шифр зависит от **ключа** $\underline{K}$, который известен только авторизированному пользователю. Криптоаналитику, или же злоумышленнику, ключ не известен. Для упрощения анализа обычно предполагают, что криптоаналитик знает применяемый шифр $E_{\underline{K}}$. Единственная информация, недоступная криптоаналитику — ключ. Очевидно, что система шифрования должна быть построена таким образом, чтобы была равна нулю вероятность равенства сообщения $\underline{M}$ и его оценки криптоаналитиком $\widehat{\underline{M}}$.

Поставим в соответствие 26 буквам английского алфавита числа от 0 до 25: $A \to 0,\ B \to 1,\ C \to 2,\ \ldots,\ Z \to 25$. **Шифр Цезаря**, который он использовал 2000 лет назад при переписке, состоял из одного

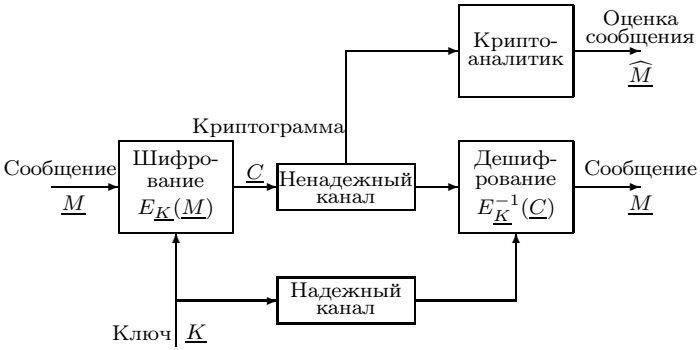


Рис. 4.1. Классическая криптосистема

ключа со следующим однозначным отображением

$$E_3(M) \equiv (M+3) \bmod 26, \;\; M = 0, 1, \ldots, 25,$$

то есть каждая буква сообщения просто заменялась отстоящей от нее третьей буквой в алфавите. Данный шифр имеет всего 26 ключей $0 \leq K \leq 25$ и может быть обобщен как

$$E_K(M) \equiv (M+K) \bmod 26, \;\; M = 0, 1, \ldots, 25.$$

Цезарь применял всегда $K = 3$, император Август $K = 4$. Причем имеются исторические ссылки на то, что Брут взломал шифр Цезаря. Сделать это было, зная систему шифрования, легко: стоило только перебрать все возможные 26 ключей и выделить имеющее смысл, то есть наиболее вероятное, сообщение из 26 возможных.

Шифр Цезаря является циклической перестановкой букв и может быть обобщен на **шифр перестановки**, в котором ключом шифра является некоторая произвольная перестановка букв. Например, пусть

исходный алфавит	`ABCDEFGHIJKLMNOPQRSTUVWXYZ`
переставленный алфавит	`XGBCAMNEFKDQTZOHRIVLPWSUJY`.

Тогда

сообщение	`INFORMATIONTHEORY`
криптограмма	`FZMOITXLFOZLEAOIJ`.

Количество ключей составляет в этом случае уже $26! > 4 \cdot 10^{26}$, и систематический их перебор не особенно привлекателен. Однако криптоанализ может быть проведен на основе частотного анализа, поскольку характерные для языка частоты определенных букв или их комбинаций в криптограмме не будут отличаться от их частот в сообщении. Чем короче криптограмма, тем сложнее проводить статистический анализ. Однако один из ведущих криптоаналитиков времен Второй мировой войны Уильям Фридман считал, что поиск наиболее вероятного сообщения или ключа можно начинать, зная криптограмму из 25 букв. Правильность данного утверждения мы обоснуем теоретически в разделе 4.2.

Одним из методов скрыть статистическую структуру сообщения является **полиалфавитный шифр**, использующий n алфавитов для перестановки. Одна и та же буква в криптограмме может в таком случае представлять разные буквы сообщения, что существенно увеличивает надежность.

Полиалфавитный шифр с $n = 2$ называется **шифром Виженера** (1523 – 1596) и использует для шифрования таблицу на рис. 4.2. Самая верхняя строка в таблице Виженера — это алфавит сообщения. Далее имеется 26 «алфавитов Цезаря», ключом каждого из которых является его первая буква. Левый столбец является алфавитом ключа, в качестве которого используется слово или фраза. Для шифрования ключ периодически записывается под сообщением, а буква криптограммы считывается в таблице Виженера с пересечения строки, соответствующей букве ключа, и столбца, соответствующего букве сообщения. Таким образом, для получения криптограммы ключ «накладывается» на сообщение. Например, сообщение «INFORMATIONTHEORY» шифруется с помощью ключа «CRYPTO» как

сообщение	`INFORMATIONTHEORY`
ключ	`CRYPTOCRYPTOCRYPT`
криптограмма	`KEDDKACKGDGHJVMGR`

Такой способ шифрования считался полностью надежным до 1863 г., когда Фридрих Касиски заметил, что если в сообщении встречаются одинаковые слова (последовательности букв), то они из-за периодичности применения ключа часто (но случайным образом) шифруются в одинаковые последовательности букв криптограммы. Такие повторения позволяют при ее достаточной длине установить длину ключа. После этого криптоаналитик разделяет криптограмму на соот-

	ABCDEFGHIJKLMNOPQRSTUVWXYZ
A	ABCDEFGHIJKLMNOPQRSTUVWXYZ
B	BCDEFGHIJKLMNOPQRSTUVWXYZA
C	CDEFGHIJKLMNOPQRSTUVWXYZAB
D	DEFGHIJKLMNOPQRSTUVWXYZABC
E	EFGHIJKLMNOPQRSTUVWXYZABCD
F	FGHIJKLMNOPQRSTUVWXYZABCDE
G	GHIJKLMNOPQRSTUVWXYZABCDEF
H	HIJKLMNOPQRSTUVWXYZABCDEFG
I	IJKLMNOPQRSTUVWXYZABCDEFGH
J	JKLMNOPQRSTUVWXYZABCDEFGHI
K	KLMNOPQRSTUVWXYZABCDEFGHIJ
L	LMNOPQRSTUVWXYZABCDEFGHIJK
M	MNOPQRSTUVWXYZABCDEFGHIJKL
N	NOPQRSTUVWXYZABCDEFGHIJKLM
O	OPQRSTUVWXYZABCDEFGHIJKLMN
P	PQRSTUVWXYZABCDEFGHIJKLMNO
Q	QRSTUVWXYZABCDEFGHIJKLMNOP
R	RSTUVWXYZABCDEFGHIJKLMNOPQ
S	STUVWXYZABCDEFGHIJKLMNOPQR
T	TUVWXYZABCDEFGHIJKLMNOPQRS
U	UVWXYZABCDEFGHIJKLMNOPQRST
V	VWXYZABCDEFGHIJKLMNOPQRSTU
W	WXYZABCDEFGHIJKLMNOPQRSTUV
X	XYZABCDEFGHIJKLMNOPQRSTUVW
Y	YZABCDEFGHIJKLMNOPQRSTUVWX
Z	ZABCDEFGHIJKLMNOPQRSTUVWXY

Рис. 4.2. Таблица Виженера

ветствующее число шифров Цезаря, которые «ломаются» достаточно просто.

В 1918 г. американский инженер Гильберт Вернам предложил самую важную модификацию шифра Виженера. Он произвел следующее наблюдение: если случайная двоичная последовательность суммируется по модулю 2 с исходной двоичной последовательностью (в терминологии шифра Виженера — накладывается на нее), то вся статистика исходной последовательности, на которой основывались преж-

ние криптоаналитические атаки, теряется. Правильность интуитивного заключения Вернама доказал Шеннон 30 лет позднее. Единственным недостатком **шифра Вернама** является то, что длина ключа (случайной двоичной последовательности) должна быть равна длине сообщения. Тем не менее впервые в истории была построена надежная система шифрования и надежная не потому, что криптоаналитик не располагал знаниями и техникой для взлома, а потому, что невозможно определить наиболее вероятный ключ, так как все они равновероятны. Такая система шифрования часто используется шпионами и была, например, найдена у Рудольфа Абеля во время ареста в 1957 г. в Нью-Йорке.

4.2. Основы теории Шеннона о защите информации

Модель Шеннона для криптосистемы состоит в общем случае из отображения $E_{\underline{K}}(\underline{M}) = \underline{C}$ сообщений $\underline{m} = \{m_1, m_2, \ldots, m_n\}$ на криптограммы $\underline{c} = \{c_1, c_2, \ldots, c_n\}$ с помощью ключей $\underline{k} = \{k_1, k_2, \ldots, k_n\}$, где $\underline{m}$, $\underline{c}$ и $\underline{k}$ рассматриваются как реализации случайных процессов с конечным алфавитом.

Шеннон предполагал, что каждое конкретное сообщение $\underline{m}$ шифруется с помощью конкретного ключа $\underline{k}$ только один раз, что криптоаналитик имеет доступ к бесконечным вычислительным ресурсам, знает отображение $E_{\underline{K}}$ для каждого ключа $\underline{k}$, знает априорные вероятности $f(\underline{m})$ и $f(\underline{k})$ и имеет в распоряжении единственную криптограмму $\underline{c}$. Задачей криптоаналитика является вычисление условных вероятностей $f(\underline{m}|\underline{c})$ для всех $\underline{m}$ или $f(\underline{k}|\underline{c})$ для всех $\underline{k}$ и определение либо наиболее вероятного сообщения $\underline{m}$, либо наиболее вероятного ключа $\underline{k}$.

Пусть $H(\underline{K})$ и $H(\underline{M})$ — соответственно энтропии ключа и сообщения (1.57).

Определение 4.1. *Условная энтропия ключа* $H(\underline{K}|\underline{C})$ *называется* **неопределенностью ключа**, *а условная энтропия сообщения* $H(\underline{M}|\underline{C})$ *называется* **неопределенностью сообщения**.

Эти величины являются неопределенностью, перед которой стоит криптоаналитик, пытающийся определить ключ или сообщение по криптограмме.

Так как условная энтропия никогда не больше безусловной (1.27),

имеем

$$H(\underline{K}|\underline{C}) \le H(\underline{K}) \tag{4.1}$$

и $H(\underline{M}|\underline{C}) \le H(\underline{M})$, то есть знание криптограммы не может (в среднем) увеличить неопределенность ключа или сообщения.

Теорема 4.1. *Неопределенность сообщения не может быть больше неопределенности ключа:*

$$H(\underline{M}|\underline{C}) \le H(\underline{K}|\underline{C}). \tag{4.2}$$

Доказательство. В соответствии с диаграммой Эйлера (с. 28), распространенной на условные энтропии, имеем

$$H(\underline{K}, \underline{M}|\underline{C}) = H(\underline{K}|\underline{C}) + H(\underline{M}|\underline{K}, \underline{C}).$$

Так как знание ключа и криптограммы должно быть достаточно для знания сообщения, то $H(\underline{M}|\underline{K}, \underline{C}) = 0$, а так как совместная неопределенность ключа и сообщения не меньше неопределенности только сообщения $H(\underline{K}, \underline{M}|\underline{C}) \ge H(\underline{M}|\underline{C})$, то мы имеем результат теоремы. □

Избыточность n-го порядка $\varrho_n(\underline{M}) = H_0 - H_n(\underline{M})$ (2.10) имеет центральное значение в исследовании системы шифрования, и именно благодаря ей возможен успех криптоанализа.

Теорема 4.2. *Неопределенность ключа ограничена снизу*

$$H(\underline{K}|\underline{C}) \ge H(\underline{K}) - n\varrho_n(\underline{M}). \tag{4.3}$$

Доказательство. Для всех систем шифрования $H(\underline{M}, \underline{K}) = H(\underline{C}, \underline{K})$. А так как $\underline{M}$ и $\underline{K}$ независимы друг от друга, то

$$H(\underline{M}, \underline{K}) = H(\underline{M}) + H(\underline{K}) = nH_n(\underline{M}) + H(\underline{K}).$$

Далее $H(\underline{C}, \underline{K}) = H(\underline{C}) + H(\underline{K}|\underline{C}) \le nH_0 + H(\underline{K}|\underline{C})$. То есть $nH_n(\underline{M}) + H(\underline{K}) \le nH_0 + H(\underline{K}|\underline{C})$, откуда и следует (4.3). □

Значит, если избыточность $\varrho_n(\underline{M}) = 0$, то из (4.1) и (4.3) следует, что $H(\underline{K}|\underline{C}) = H(\underline{K})$, то есть знание криптограммы ничем не облегчает поиск ключа. Поэтому сжатие данных (удаление из них избыточности) является важным дополнением к системе шифрования, так как повышает ее надежность. Особенно актуален данный факт для шифрования текстов реального языка: убрав избыточность, мы оставим криптоаналитику существенно меньше шансов расшифровать текст.

Если длина n криптограммы превосходит $L_n = H(\underline{K})/\varrho_n(\underline{M})$, то нижняя граница для неопределенности ключа $H(\underline{K}|\underline{C})$ из (4.3) равна

нулю, что означает наш риск, что аналитику может повезти с оценкой текущего ключа. Поэтому Шеннон ввел следующее определение.

Определение 4.2. *Наименьшая средняя длина криптограммы, такая, что существует в точности одна пара сообщение – ключ, дающая данную криптограмму, называется* **расстоянием единственности** *и обозначается* L_∞.

Из определения и (4.3) следует, что

$$\mathsf{L}_\infty = \frac{H(\underline{K})}{\rho_\infty(\underline{M})}, \tag{4.4}$$

то есть система шифрования с расстоянием единственности L_∞ никогда не должна применяться для шифрования сообщений длины больше либо равной L_∞.

Пример 4.1. Для шифра перестановки $H(\underline{K}) = \log_2(27!) \approx 93{,}14$ бит; для английского языка $\rho_\infty(\underline{M}) \approx 4{,}75 - 1{,}34 = 3{,}41$ (1.88), то есть модель Шеннона предсказывает $\mathsf{L}_\infty = \frac{93{,}14}{3{,}41} \approx 27$ букв (чуть больше, чем граница в 25 букв, которую дает Фридман).

Следует, однако, иметь в виду, что при попытке вскрыть шифр перестановки, зная только частоты появления n-буквенных сочетаний, для успешного криптоанализа может потребоваться криптограмма существенно большей длины, чем L_∞. Эту длину можно оценить, подставив в формулу (4.4) вместо абсолютной избыточности избыточность n-го порядка $\rho_n(\underline{M})$ (2.10), вычисленную с учетом (1.88). Тогда получаем, что знание только частот отдельных букв потребует криптограмму длины $L_1 = \frac{93{,}14}{4{,}75-4{,}03} \approx 129$ букв, знание частот пар букв потребует криптограмму длины $L_2 = \frac{93{,}14}{4{,}75-3{,}32} \approx 65$, троек — $L_3 = \frac{93{,}14}{4{,}75-3{,}10} \approx 56$, четверок — $L_4 = \frac{93{,}14}{4{,}75-2{,}80} \approx 48$ букв и т. д. То есть чем меньше мы знаем о передаваемом сообщении, тем большей длины криптограмма требуется для успешного анализа и тем он сложнее.

Пример 4.2. Для шифра Вернама $H(\underline{K}) = nH_0$, так как $\rho_\infty(\underline{M}) < H_0$, то $n < \mathsf{L}$, то есть расстояние единственности превосходит длину криптограммы.

Определение 4.3. *Система шифрования дает оптимальную защиту, если* $I(\underline{M};\underline{C}) = 0$, *то есть если сообщение и криптограмма статистически независимы.*

Теорема 4.3. *Для оптимальной системы шифрования энтропия ключа не меньше энтропии сообщения:*

$$H(\underline{M}) \le H(\underline{K}). \tag{4.5}$$

Доказательство. Для любой оптимальной системы шифрования имеет место $H(\underline{M}) = H(\underline{M}|\underline{C}) \leq H(\underline{K}|\underline{C}) \leq H(\underline{K})$, где первое равенство следует из определения оптимальной системы шифрования, а первое неравенство есть (4.2). □

Неравенство (4.5) является фундаментальной границей Шеннона для оптимального, то есть теоретически надежного, шифрования.

Теорема 4.4. *Шифр Вернама является оптимальной системой шифрования.*

Доказательство. Для шифра Вернама криптограмма $\underline{c}$ получается путем посимвольного сложения по модулю два сообщения $\underline{m}$ и ключа $\underline{k}$. Причем ключ $\underline{k}$ мы можем трактовать как последовательность ошибок в двоично-симметричном канале с $p_c = 1/2$. Так как при этой вероятности его пропускная способность равна нулю (2.53), то $I(\underline{M};\underline{C}) = 0$.□

4.3. Задачи

4.2.1. [2 балла] Докажите, что для любой нетривиальной системы шифрования

$$I(\underline{M};\underline{C}) \geq H(\underline{M}) - H(\underline{K}),$$

то есть в системе с малым количеством ключей криптограмма несет много информации о сообщении.

4.2.2. [2 балла] Пусть сообщения $a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8$ источника без памяти имеют вероятности 1/2, 1/4, 1/8, 1/16, 1/32, 1/64, 1/128, 1/128 соответственно. Найдите неопределенность ключа и сообщения, а также расстояние единственности, если фразы длины n, порождаемые данным источником, шифруются

(а) шифром Цезаря,

(б) шифром перестановки.

4.2.3. [2 балла] Определите расстояние единственности для шифра Виженера.

4.2.4. [3 балла] Дайте верхнюю оценку абсолютной энтропии английского языка, исходя из предположения, что зашифрованное шифром перестановки сообщение длины n или более букв может быть однозначно расшифровано.

Библиографические ссылки

1. *Shannon C. E.* A Mathematical Theory of Communication. Bell Sys. Tech. J. 27: 379 – 423 (Part I), 623 – 656 (Part II); рус. пер.: *Шеннон К.* Математическая теория связи // Работы по теории информации и кибернетике. М.: ИЛ, 1963.

2. *Shannon C. E.* Communication Theory of Secrecy Systems. Bell Sys. Tech. J. 28: 656 – 715; рус. пер.: *Шеннон К.* Теория связи в секретных системах // Работы по теории информации и кибернетике. М.: ИЛ, 1963.

3. *Стратонович Р. Л.* Теория информации. М.: Советское радио, 1975. 424 с.

4. *Коган И. М.* Прикладная теория информации. М.: Радио и связь, 1981. 216 с.

5. *Колесник В. Д., Полтырев Г. Ш.* Курс теории информации. М.: Наука, 1982. 416 с.

6. *Дмитриев В. И.* Прикладная теория информации: учебник для вузов. М.: Высш. шк., 1989. 320 с.

7. *Cover T. M., Thomas J. A.* Elements of Information Theory. John Wiley & Sons, 1991. 542 p.

8. *Johannesson R.* Informatonstheorie — Grundlage der (Tele-)Kommunikation. Addison-Wesley, 1992. 302 S.

9. *Verdú S.* Fifty Years of Shannon Theory // IEEE Transactions on Information Theory. 1998. Vol. 44. N. 6. P. 2057 – 2078.

10. *Вентцель Е. С., Овчаров Л. А.* Теория случайных процессов и ее инженерные приложения: учеб. пособие для втузов. 2-е изд. М.: Высш. шк., 2000. 383 с.

11. *Алферов А. П., Зубов А. Ю., Кузьмин А. С., Черемушкин А. В.* Основы криптографии: учеб. пособие. М.: Гелиос АРВ, 2001. 480 с.

12. *Вентцель Е. С.* Теория вероятностей: учебник для вузов. 7-е изд. М.: Высш. шк., 2001. 575 с.

13. *Иванов М. А.* Криптографические методы защиты информации в компьютерных системах и сетях. М.: КУДИЦ-ОБРАЗ, 2001. 368 с.

14. *Lyre H.* Informationstheorie: eine philosophisch-naturwissenschaftliche Einführung. München: Fink, 2002. 279 S.

15. *Hankerson D., Harris G., Johnson P.* Introduction to information theory and data compression. CRC Press LLC, 2003. 366 p.

16. Математические и компьютерные основы криптологии: учеб. пособие / Ю. С. Харин, В. И. Берник, Г. В. Матвеев, С. В. Агиевич. Минск: Новое знание, 2003. 382 с.

17. *Чернавский Д. С.* Синергетика и информация (динамическая теория информации). 2-е изд., испр. и доп. М.: Едиториал УРСС, 2004. 288 с.

18. *Эвери Д.* Теория информации и эволюция. М.; Ижевск: НИЦ «Регулярная и хаотическая динамика»: Ин-т компьютерных исследований, 2006. 252 с.

19. *Демин А. И.* Парадигма дуализма: пространство – время, информация – энергия. М.: Изд-во ЛКИ, 2007. 320 с.

20. *Духин А. А.* Теория информации: учеб. пособие. М.: Гелиос АРВ, 2007. 248 с.

21. *Панин В. В.* Основы теории информации: учеб. пособие для вузов. М.: БИНОМ. Лаборатория знаний, 2007. 436 с.

22. *Desurvire E.* Classical and Quantum Information Theory: An Introduction for the Telecom Scientist. Cambridge University Press, 2009. 714 p.

23. *Sommaruga G. (Ed.)* Formal Theories of Information: From Shannon to Semantic Information Theory and General Concepts of Information. Berlin: Springer, 2009. 269 p.

24. *Кудряшов Б. Д.* Теория информации: учеб. для вузов. СПб.: Питер, 2009. 320 с.

Предметный указатель

Оглавление

Учебное издание

Свирид Юрий Владимирович

ОСНОВЫ ТЕОРИИ ИНФОРМАЦИИ

КУРС ЛЕКЦИЙ

В авторской редакции

Технический редактор *Г. М. Романчук*
Корректор *С. П. Гринкевич*

Ответственный за выпуск *Т. М. Турчиняк*

Подписано в печать 19.08.2010. Формат 60×84/16. Бумага офсетная.
Гарнитура Roman. Печать офсетная. Усл. печ. л. 8,84. Уч.-изд. л. 7,98.
Тираж 100 экз. Зак. 854

Белорусский государственный университет.
Лицензия ЛИ N 02330/0494425 от 08.04.2009.
Пр. Независимости, 4, 220030, Минск.

Отпечатано с оригинала-макета заказчика.
Республиканское унитарное предприятие
«Издательский центр Белорусского государственного университета»
ЛП N 02330/0494178 от 03.04.2009.
Ул. Красноармейская, 6, 220030, Минск.