Министерство образования и науки Российской Федерации

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Приоритетный национальный проект «Образование»
Национальный исследовательский университет

А.Л. ГЕЛЬГОР ***Е.А. ПОПОВ***

ТЕОРЕТИКО-ИНФОРМАЦИОННЫЕ ОСНОВЫ ТЕЛЕКОММУНИКАЦИОННЫХ СИСТЕМ

Рекомендовано Учебно-методическим объединением по университетскому политехническому образованию в качестве учебного пособия для студентов высших учебных заведений, обучающихся по направлению подготовки магистров "Техническая физика"

Санкт-Петербург
Издательство политехнического университета
2013

УДК 621.391:005(075.8)
ББК 32.811:32.97я73
Г 32

Рецензенты:

Кафедра радиопередающих устройств и средств подвижной связи СПбГУТ им. проф. М.А. Бонч-Бруевича, зав. кафедрой д.т.н., профессор М. Сиверс

Кафедра военных телекоммуникационных систем Военной академии связи им. С.М. Будённого, нач. кафедры к.в.н., доц. А. Боговик

Гельгор А.Л. **Теоретико-информационные основы телекоммуникационных систем**: учеб. пособие / Гельгор А.Л., Попов Е.А. — СПб.: Изд-во Политехн. ун-та, 2013. — 288 с.

Формулируются основные понятия, касающиеся определения количества собственной и взаимной информации, относящиеся к дискретным и непрерывным источниками сообщений: энтропия, избыточность, взаимная информация и др. Обсуждаются информационные характеристики случайных процессов, обладающих различными вероятностными свойствами. Большое внимание уделено моделям дискретных каналов: идеальные дискретные каналы, симметричные и несимметричные каналы, каналы с памятью (модель Джильберта). Рассматриваются основные идеи сжатия дискретной информации. Излагаются основные принципы передачи информации по дискретным каналам с шумами.

Пособие предназначено для студентов, обучающихся по направлению подготовки “Техническая физика”. Оно может быть также использовано при обучении студентов направлений подготовки “ Инфокоммуникационные технологии и системы связи”, “Радиотехника”.

Работа выполнена в рамках реализации программы развития национального исследовательского университета “Модернизация и развитие политехнического университета как университета нового типа, интегрирующего мультидисциплинарные научные исследования и надотраслевые технологии мирового уровня с целью повышения конкурентоспособности национальной экономики”

Печатается по решению редакционно-издательского совета Санкт-Петербургского государственного политехнического университета.

ISBN

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Развитие современного общества немыслимо без широкого исследования разнообразных средств передачи информации. Совершенствование этих средств, включая увеличение объёмов передаваемой информации, повышение качества (верности) передачи, рост скорости доставки сообщений были и остаются важнейшими задачами, решаемыми разработчиками вычислительных и телекоммуникационных систем и сетей связи.

Несмотря на то, что в основополагающих работах Клода Элвуда Шеннона (Claude Elwood Shannon), с появлением которых обычно связывают рождение теории передачи информации, речь шла о потенциальных возможностях *произвольной* системы, инвариантно к физической природе источника сообщений и канала передачи, исторически развитие теории и практики построения систем можно условно отнести к двум различным направлениям.

Первое из них было обусловлено созданием в середине XX века цифровых электронных вычислительных машин и последующим их развитием. Успехи построения все более надёжных и быстродействующих вычислительных комплексов сделали возможным решение многих задач, связанных с обработкой и хранением информации. При этом все возрастающие объёмы обрабатываемой информации породили проблему её *сжатия*, потенциальная возможность и сравнительная эффективность которого является одним из выводов теории Шеннона. Некоторые результаты решения этой проблемы в её классической постановке ныне известны каждому компьютерному пользователю, работающему с программными архиваторами, которые прошли эволюцию от двухпроходного алгоритма Хаффмана до алгоритмов словарного сжатия Зива–Лемпеля и продолжают развиваться и сейчас.

При решении большинства задач, связанных с вопросами сжатия информации, явно или неявно предполагается высокое качество канала передачи — настолько высокое, что проблема *искажения* информации не является актуальной, что, например, имеет место при передаче информации по внутренним магистралям какого-либо процессора. Однако ситуация резко меняется, если в канале действуют помехи, интенсивность которых спо-

собна привести к частичной, а то и полной потери информации. В этих условиях на первый план выходят задачи поиска таких методов передачи и приёма сообщений, которые бы обеспечивали минимальную потерю информации при прохождении сообщений по каналу с помехами. Решение таких задач и составило второе направление.

Исторически сложилось так, что специалисты, занимающиеся решением какой-либо из указанных задач даже в рамках одного направления, зачастую были в значительной степени обособлены друг от друга. Так, радиоинженеры, как правило, были недостаточно квалифицированными в программировании; программисты, занимаясь разработкой очередного программного обеспечения, имели весьма смутное представление о системах передачи информации; и те и другие в большинстве своём не владели в должной степени математическим аппаратом; математики, в свою очередь, имели сложности в инженерной постановке задач и практической интерпретации результатов.

Такая ситуация с узкоспециализированной подготовкой неожиданно оказалась тревожной к началу 1980-х годов, когда чётко обозначилась тенденция к глобализации телекоммуникационных систем и выявилась необходимость (и не в последнюю очередь — экономическая) в более подготовленных специалистах. Создание систем, предназначенных для передачи мультиплексированных, компрессированных и различных по своей природе данных (текст, звук, видео, управление, телеметрия), содержащих в качестве составных элементов локальные, а порой и глобальные компьютерные сети и работающих в условиях целого ряда технико-экономических ограничений, требует от разработчиков глубокой подготовки, опирающейся на фундамент теории информации.

К сожалению, существующая на сегодняшний момент отечественная научная и методическая литература не удовлетворяет в должной степени условиям подготовки таких специалистов. Современные издания в большинстве своём ориентированы на специалистов, занимающихся разработкой в узкой области, и носят прикладной характер. В этих условиях представляется целесообразным рассмотрение теории передачи информации, ориентированное на современное состояние средств телекоммуникаций и возможность практического применения теоретических положений. При этом оказывается необходимым использование как математического аппа-

рата теории информации, так и многих аспектов современной физики, радиотехники, вычислительной техники и связи.

Заметим, что многие выводы классической теории информации получены, исходя из упрощающих предположений, таких, как идеальность канала передачи с прямоугольной частотной характеристикой или финитность спектра передаваемых сигналов, в то время как реальные телекоммуникационные каналы значительно далеки от таких моделей. Указанное обстоятельство не позволяет сравнивать между собой эффективность различных систем и, как следствие, оценивать их стоимость, возможности и перспективность.

В то же время, корректное использование современных результатов теории информации позволяет в ряде случаев получить полезные количественные и качественные результаты, например, оценить величину пропускной способности реальных непрерывных каналов передачи, сравнить между собой по пропускной способности различные телекоммуникационные и информационные системы, компьютерные сети. Кроме того, практическое использование результатов расчётов пропускной способности реальных непрерывных каналов передачи сообщений (проводных, спутниковых, радиорелейных, подвижной связи, коротковолновых каналов передачи, волоконно-оптических каналов, цифровых каналов мониторинга и др.) позволяет оценить, насколько полно используются возможности каналов передачи и определить теоретические резервы увеличения объёмов передаваемой информации в рассматриваемой информационной системе при сохранении требуемого качества приёма сообщений.

История развития теории информации с самого рождения была весьма бурной: новизна тематики, оригинальность и общность подхода, а также само название теории — может быть чересчур общее (многие исследователи предлагали и предлагают более “узкие” названия: “теория передачи информации” или “теория связи”, как это называлось в самих основополагающих работах Шеннона) — произвели в середине XX века настоящий научный бум. В эти годы теория информации привлекла пристальное внимание многих исследователей, что не замедлило сказаться на её достижениях и в теоретическом плане, и в практическом применении.

Вместе с тем, такое бурное развитие теории сделало её очень “модной” наукой, что не замедлило сказаться на появлении многочисленных попыток,

связанных с привлечением теоретико-информационных концепций к решению задач самых разнообразных научных направлений: физики, психологии, лингвистики и др. В ряде случаев дело сводилось к механическому переносу терминов теории информации, иногда без ясного понимания их смысла. Как известно, первым, забившим тревогу по этому поводу, был сам Шеннон, выступивший со статьёй “The Bandwagon”. После его выступления поток публикаций, в которых понятия теории информации используются некорректно, несколько поредел; произошёл заметный “отлив” интереса от теории информации, но это послужило ей только на пользу, так как ушли большинство тех, кто либо разочаровался в возможности автоматического перенесения результатов, либо почувствовал себя несостоятельным для преодоления обнажившихся после первых успехов трудностей.

Шли годы, и спокойное, “убаюкивающее” развитие теории информации породило другую крайность — мнение о том, что методы построения систем передачи информации, такие, как, например, фильтрация, корреляционный приём, разнесение, достаточно изучены и основываются на простых и интуитивно понятных соображениях. Согласно этому мнению, теория информации даёт лишь математически более сложный аппарат, подтверждающий или уточняющий границы применения тех или иных методов. Однако если посмотреть на достижения теории за крайние 20–30 лет: новые классы каскадных и турбокодов и методы их декодирования; сигнально-кодовые конструкции на основе спектрально-эффективных сигналов; параллельная пространственная обработка сигналов и др., то нетрудно понять, что эти достижения возможны лишь на основе глубоких исследований, когда требуется знание идей, постановок и решений задач теории информации наряду с владением рядом разделов математики. Техническая интуиция здесь уже не является адекватной моделью анализа.

В предлагаемом учебном пособии, формулируются основные понятия, касающиеся определения количества собственной и взаимной информации, относящие к дискретным и непрерывным источникам сообщений, обсуждаются информационные характеристики случайных процессов. Анализируются классические опыты К. Шеннона и А. Колмогорова, посвящённые исследованию информационных характеристик литературных источников, проведён сравнительный анализ энтропии письменной и устной речи. Большое внимание уделено моделям дискретных каналов: иде-

альные дискретные каналы, симметричные и несимметричные каналы, каналы с памятью (модель Джильберта). Рассматриваются основные идеи сжатия дискретной информации (хаффмановское кодирование). Излагаются основные принципы передачи информации по дискретным каналам с помехами.

В конце каждой главы приведён список вопросов и задний, среди которых имеются как простые, предназначенные для проверки того, что студент правильно понял формулировки определений и свойств, так и достаточно сложные, в том числе, расчётного характера, требующие весьма глубокого понимания материала.

Учебное пособие составлено преподавателями кафедры “Радиоэлектронные средства защиты информации” по материалам курсов лекций, читаемых авторами на физических и технических факультетах Санкт-Петербургского государственного политехнического университета и предназначено для студентов, обучающихся по направлениям подготовки “Техническая физика”, “Инфокоммуникационные технологии и системы связи”, “Радиотехника”, “Информатика и вычислительная техника”.

1. КОЛИЧЕСТВО ИНФОРМАЦИИ, СОДЕРЖАЩЕЙСЯ В СООБЩЕНИИ

Для того, чтобы иметь возможность сравнивать различные источники сообщений и различные каналы связи, необходимо ввести некоторую количественную меру, позволяющую оценивать содержащуюся в сообщении информацию. Такая мера в виде количества информации была введена К. Шенноном [1], что позволило ему построить общую математическую теорию связи. Рассмотрим основные идеи этой теории применительно к *дискретному* и *непрерывному* источникам и попытаемся найти удобную меру количества информации, заключённой в некотором сообщении.

1.1. КОЛИЧЕСТВО ИНФОРМАЦИИ, СОДЕРЖАЩЕЙСЯ В ИСТОЧНИКЕ ДИСКРЕТНЫХ СООБЩЕНИЙ

Рассмотрим дискретный источник, представляющий собой l-элементное множество $X = \{x_k\}$, $k = 1, \ldots, l$. Будем называть *сообщением* событие, состоящее в появлении любой конечной последовательности элементов этого множества. При этом сами элементы x_k будем называть *элементарными сообщениями*, а их количество l — *объёмом источника*.

Основная идея, лежащая в основе получения количественной меры информации применительно к источнику дискретных сообщений, заключается в том, что такая мера определяется не конкретным смысловым содержанием данного сообщения, а тем фактом, что источник выбирает данное элементарное сообщение x_k из конечного множества X с определённой вероятностью $p(x_k)$ ($k = 1, \ldots, l$). Если выбираемый элемент сообщения заранее известен (например, всегда посылается один и тот же сигнал), то естественно полагать, что заключающаяся в нем информация равна нулю. Поэтому считается, что выбор элемента x_k происходит с некоторой отличной от единицы вероятностью $p_k = p(x_k)$, зависящей в общем случае от того, какая последовательность элементов предшествовала данному. Примем, что количество информации, заключённое в элементарном сообщении x_k, является непрерывной дифференцируемой функцией этой вероятности $I[p(x_k)]$. Определим вид этой функции так, чтобы получающаяся при

этом мера количества информации удовлетворяла условию аддитивности, т. е. если вероятность какого-либо сообщения факторизуется, то количество информации, содержащейся в таком сообщении, равняется сумме количеств информации, содержащейся в каждом из элементов, образующих сообщение. При этом будем считать отличными от нуля все вероятности p_k выбора элементов, что имеет место для всех практически используемых источников.

Для определения вида функции количества информации проведём следующие рассуждения. Пусть $p(x_i, x_k)$ — вероятность того, что источник произведёт последовательный выбор элементов x_i и x_k, а $I[p(x_i, x_k)]$ — количество информации, соответствующее этому выбору. При этом количество информации, содержащейся в x_i, равно $I[p(x_i)]$. Для определения количества информации, содержащейся в x_k, нужно учесть условную вероятность $p(x_k / x_i)$ выбора x_k после x_i. По правилу умножения вероятностей $p(x_i, x_k) = p(x_k / x_i)\, p(x_i)$, тогда, учитывая предположение об аддитивности меры количества информации, имеет место соотношение:

$$I\left[p(x_i, x_k)\right] = I\left[p(x_i)\right] + I\left[p(x_k / x_i)\right],$$

или, ещё раз использовав правило умножения вероятностей:

$$I\left[p(x_k / x_i)\, p(x_i)\right] = I\left[p(x_i)\right] + I\left[p(x_k / x_i)\right]. \qquad (1.1.1)$$

Для определения явного вида функции I в полученном функциональном уравнении возьмём частные производные по $p(x_i)$ от обеих частей. Тогда, поскольку

$$\frac{\partial I\left[p(x_k / x_i)\, p(x_i)\right]}{p(x_i)} = \frac{\partial I\left[p(x_k / x_i)\, p(x_i)\right]}{\partial\left[p(x_k / x_i)\, p(x_i)\right]} \frac{\partial\left[p(x_k / x_i)\, p(x_i)\right]}{p(x_i)} =$$

$$= p(x_k / x_i) \frac{\partial I\left[p(x_k / x_i)\, p(x_i)\right]}{\partial\left[p(x_k / x_i)\, p(x_i)\right]},$$

умножив обе части на $p(x_i)$[1] и введя для простоты записи обозначения $p(x_i) = \xi$, $p(x_k / x_i) = \eta$, $\varphi = \xi\eta$, получим простое дифференциальное уравнение:

$$\xi I'(\xi) = \varphi I'(\varphi). \qquad (1.1.2)$$

Справедливость этого уравнения для произвольных величин ξ и φ

[1] По предположению все $p(x_k)$ отличны от нуля.

($0 < \xi \le 1,\ 0 < \varphi \le 1$) возможна лишь в том случае, если обе его части представляют некоторую постоянную величину α, не зависящую ни от ξ, ни от φ:

$$\xi I'(\xi) = \alpha .$$

Тогда, интегрируя полученное уравнение, находим

$$I(\xi) = \alpha \ln |\xi| + c ,$$

или, возвращаясь к прежним обозначениям,

$$I[p(x_k)] = \alpha \ln p(x_k) + c . \qquad (1.1.3)$$

Для определения константы c следует воспользоваться высказанным выше условием, по которому заранее предопределённый элемент сообщения, т. е. имеющий вероятность $p = 1$, содержит нулевое количество информации. Следовательно, $I(1) = 0$, откуда $c = 0$. Что касается коэффициента α, то его можно выбрать произвольно, так как он определяет лишь систему единиц, в которых измеряется информация. Поскольку $\ln p \le 0$, разумно выбрать α отрицательным, для того, чтобы количество информации было положительным.

Наиболее простым является выбор $\alpha = -1$. Тогда

$$I(p) = -\ln p = \ln(1/p) . \qquad (1.1.4a)$$

При этом единица количества информации равна той информации, которая содержится в элементарном сообщении, имеющем вероятность появления $1/e$ ($e = 2{,}71828\ldots$ — основание натуральных логарифмов). Такую единицу количества информации называют натуральной единицей (иногда используются термины *нит* или *нат*), и она используется при аналитических расчётах информационных характеристик. Однако чаще всего выбирают значение $\alpha = -1/\ln 2$, и в этом случае

$$I(p) = -\log_2 p = \log_2(1/p) . \qquad (1.1.4б)$$

При таком выборе α единица информации называется двоичной или *битом* (bit, binary digit). Она равна информации, содержащейся в сообщении о том, что наступило событие, вероятность появления которого равна 1/2. В принципе возможен выбор и других единиц измерения количества информации, например, десятичных, что имело место на заре развития теории информации. В тех случаях, когда выбор единиц не играет роли, пишут

$$I(p) = -\log p , \qquad (1.1.4)$$

считая, что логарифм берется по любому основанию, больше единицы.

Заметим, что в современной научно-технической литературе понятие бита трактуется чрезвычайно широко; зачастую под ним понимается любой двоичный разряд при цифровом представлении информации. С теоретико-информационных позиций такой подход представляется не совсем корректным, поскольку в этом случае, например, идущие друг за другом 8 двоичных разрядов “содержат” 8 бит информации. В действительности количество информации, заключенное в этих разрядах, определяется вероятностными характеристиками источника.

1.2. ЭНТРОПИЯ ИСТОЧНИКА ДИСКРЕТНЫХ СООБЩЕНИЙ

Для теории связи основное значение имеет не количество информации $I(p_k)$, содержащейся в одном элементе x_k, а средняя по всем элементам источника X величина количества информации $H(X)$, называемая *энтропией* источника:

$$H(X) = \mathbf{E}\left\{ I\left[p(x_k) \right] \right\}. \qquad (1.2.1)$$

Здесь, как и всюду в дальнейшем, символ **E** обозначает математическое ожидание случайной величины.

В случае источника с независимыми элементами, т. е. когда вероятность $p(x_k)$ выбора элемента x_k не зависит от ранее выбранных элементов, энтропия источника равна

$$H(X) = -\sum_{k=1}^{l} p(x_k) \log p(x_k). \qquad (1.2.2)$$

Рассмотрим основные свойства энтропии источника, у которого элементы выбираются независимо друг от друга.

Для того чтобы энтропия равнялась нулю, необходимо и достаточно, чтобы одна из вероятностей $p(x_k)$ равнялась единице, а все остальные — нулю.

Докажем необходимость. Пусть

$$H(X) = -\sum_{k=1}^{l} p(x_k) \log p(x_k) = 0.$$

Если $p \in (0;1)$, то $\log p \in (-\infty;0)$, и ненулевые слагаемые одного знака вида $p \log p$ не могут дать в сумме нуль, поэтому остаётся $p(x_i) = 1$, $p(x_k) = 1$ $(k \neq i)$.

Для доказательства достаточности положим $p(x_i)=1,\ p(x_k)=0\ (k\neq i)$, тогда, учитывая, что

$$\lim_{p\to 0} p\ln p=\lim_{p\to 0}\frac{\ln p}{1/p}=\lim_{p\to 0}\frac{1/p}{\left(-1/p^2\right)}=-\lim_{p\to 0} p=0,$$

получаем:

$$1\cdot 0+\lim_{p\to 0}\sum_{k\neq i} p(x_k)\log p(x_k)=0.$$

При фиксированном объёме l источника энтропия принимает максимальное значение $H_{\max}=\log l$ при равновероятных значениях символов.

Как известно [2], необходимым условием экстремума функции нескольких переменных $H(p_1, \ldots, p_l)$ при наличии дополнительного ограничения вида $\varphi(p_1,\ldots,p_l)=0$ является система уравнений

$$\frac{\partial L(p_1,\ldots,p_l)}{\partial p_k}=0\ \ (k=1,\ldots,l), \qquad (1.2.3)$$

где

$$L(p_1,\ldots,p_l)=H(p_1,\ldots,p_l)+\lambda\varphi(p_1,\ldots,p_l)$$

есть функция Лагранжа, а λ — пока неизвестный множитель, значение которого находится из дополнительного ограничения.

Для удобства записи введем вектор градиента

$$\nabla=\left(\frac{\partial}{\partial p_1}\quad\frac{\partial}{\partial p_2}\quad\cdots\quad\frac{\partial}{\partial p_l}\right)^T=\begin{pmatrix}\dfrac{\partial}{\partial p_1}\\ \cdots\\ \dfrac{\partial}{\partial p_l}\end{pmatrix}$$

и вероятностный вектор

$$\mathbf{p}=\left(p_1\quad p_2\quad\ldots\quad p_l\right),$$

тогда указанную систему (1.2.3) можно записать в компактном виде

$$\nabla L=0. \qquad (1.2.3a)$$

Полагая

$$H(\mathbf{p})=-\sum_{k=1}^{l} p_k\log p_k,\qquad \varphi(\mathbf{p})=\sum_{k=1}^{l} p_k-1,$$

имеем:

$$-\left(\log p_k + p_k \frac{1}{p_k}\log e\right)+\lambda = 0\,,\ (k=1,...,l)$$

откуда следует, что величины $\log p_k$ и, следовательно, p_k одинаковы для всех значений *k*. Найти их можно из условия $\sum_{k=1}^{l} p_k = 1$, что даёт $p_1 = p_2 = ... = p_l = 1/l$. При этом, как нетрудно видеть, $H = H_{\max} = \log l$.

Условия (1.2.3) существования экстремума (для данной задачи — максимума) для произвольной функции $L(\mathbf{p})$ являются необходимыми; их достаточность в данном случае очевидна. Вообще же достаточным условием существования локального экстремума наряду с (1.2.3) является знакоопределённость выражения $\mathbf{p}\nabla^2 L(\mathbf{p})\mathbf{p}^T$, где $\nabla^2 L(\mathbf{p})$ — матрица смешанных производных

$$\nabla^2 L(\mathbf{p}) = \begin{pmatrix} \frac{\partial^2 L(\mathbf{p})}{\partial p_1^2} & \frac{\partial^2 L(\mathbf{p})}{\partial p_1 \partial p_2} & \cdots & \frac{\partial^2 L(\mathbf{p})}{\partial p_1 \partial p_l} \\ \frac{\partial^2 L(\mathbf{p})}{\partial p_2 \partial p_1} & \frac{\partial^2 L(\mathbf{p})}{\partial p_2^2} & \cdots & \frac{\partial^2 L(\mathbf{p})}{\partial p_2 \partial p_l} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 L(\mathbf{p})}{\partial p_l \partial p_1} & \frac{\partial^2 L(\mathbf{p})}{\partial p_l \partial p_2} & \cdots & \frac{\partial^2 L(\mathbf{p})}{\partial p_l^2} \end{pmatrix},$$

называемая *матрицей Гессе*. Тогда для локального максимума

$$\mathbf{p}\nabla^2 L(\mathbf{p})\mathbf{p}^T < 0 \qquad (1.2.4a)$$

и

$$\mathbf{p}\nabla^2 L(\mathbf{p})\mathbf{p}^T > 0 \qquad (1.2.4б)$$

для локального минимума.

Для рассматриваемой задачи

$$\frac{\partial^2 L(\mathbf{p})}{\partial p_i \partial p_j} = \begin{cases} -1/p_i, & i = j \\ 0 & i \neq j \end{cases}, (i, j = 1, \ldots, l),$$

так что

$$\mathbf{p}\nabla^2 L(\mathbf{p})\mathbf{p}^T = \begin{pmatrix} p_1 \cdots p_l \end{pmatrix} \begin{pmatrix} -1/p_1 & 0 & \cdots & 0 \\ 0 & -1/p_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & -1/p_l \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ \cdots \\ p_l \end{pmatrix} = -1,$$

и в точке $\mathbf{p}_0 = (1/l\ 1/l\ \ldots\ 1/l)$ имеет место и локальный, и глобальный максимум.

Проиллюстрируем доказанное свойство на примере двоичного источника $X = \{x_1, x_2\}$, вероятности элементов которого равны $p(x_1) = p$ и $p(x_2) = 1 - p$. В этом случае энтропия источника $H(X)$ равна

$$H = -\left[p \log_2 p + (1-p) \log_2 (1-p) \right] \text{ бит на символ,} \qquad (1.2.5)$$

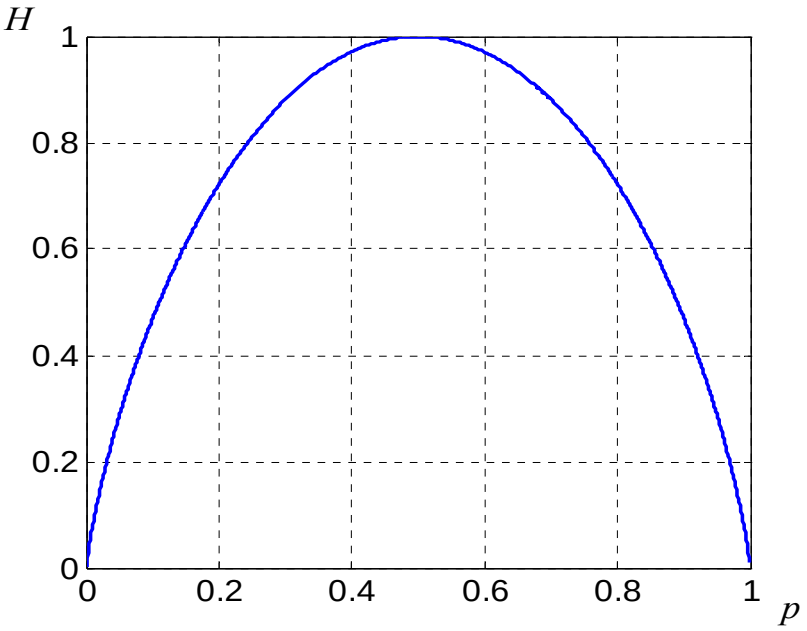
и её вид как функции от p приведён на рис. 1.1, *а*.

На рис. 1.1, *б* показаны линии уровня для поверхности $H(p_1, p_2)$, а также пунктиром нанесена линия $p_1 + p_2 = 1$ — условие нормировки, являющееся ограничивающим условием. Видно, что безусловный максимум, равный $(l/e)\log e = 1{,}062$, достигаемый в точке $\mathbf{p} = (1/e\ 1/e)$, превышает значение $H_{\max} = \log l = 1$, получающееся как пересечение (в данном случае — касание) линии уровня поверхности $H(p_1, p_2)$ и прямой $p_1 + p_2 = 1$.

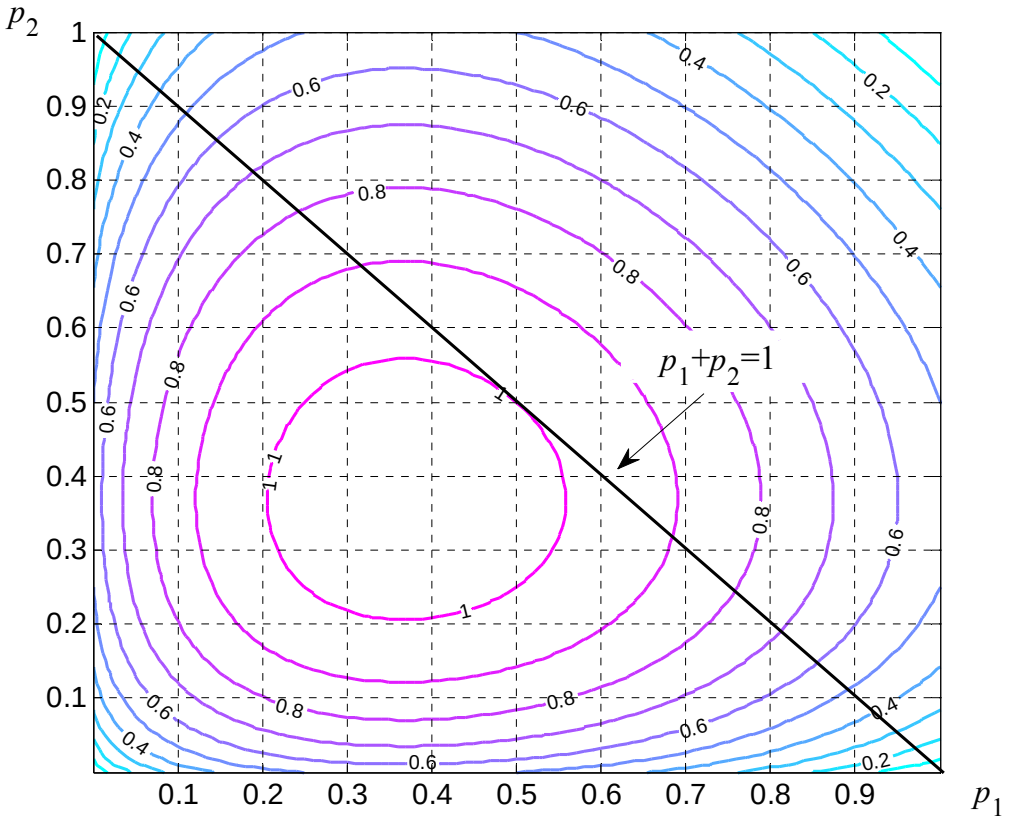
В дальнейшем при указании единиц измерения энтропии будем опускать составляющую “на символ”, измеряя энтропию просто в битах, но не забывая при этом, что речь идёт об усреднённом по всему источнику значении количества информации.

Итак, энтропия принимает наименьшее значение, равное нулю для источника, в котором все сообщения заранее предопределены: одно из них появляется всегда, а остальные — никогда. С другой стороны, энтропия принимает наибольшее значение $\log l$ в том случае, когда все сообщения независимы и одинаково вероятны. Эти свойства часто используют для того, чтобы толковать энтропию как меру неопределенности в эксперименте с получением сообщений от источника. Если сообщения заранее предопределены, и наблюдатель знает, какое он получит сообщение, то неопределенность эксперимента равна нулю. Если ни одно сообщение не имеет преимуществ по сравнению с другими, то наблюдатель с одинаковыми шансами может получить любое сообщение — неопределённость максимальна.

Такое толкование энтропии является полезным, но, скорее, качественным подходом, поскольку наряду с неравновероятностью выбора элементов источники характеризуются ещё одним свойством: *памятью*. Обратимся теперь к изучению свойств источников с памятью, в которых вероятность выбора данного элемента зависит от того, какие элементы источник генерировал на предыдущих шагах.



а)



б)

Рис. 1.1. Энтропия двоичного источника: зависимость H как функция от p (*а*) и линии уровня поверхности $H(p_1, p_2)$ (*б*)

1.3. ЭНТРОПИЯ ИСТОЧНИКА С ПАМЯТЬЮ

Математической моделью источников с памятью являются *цепи Маркова*. Цепью Маркова *n*-го порядка называется последовательность зависимых испытаний, при которой для условных вероятностей справедливо соотношение

$$p\left(x_k^{(i)} / x^{(j_1)}, x^{(j_2)}, ..., x^{(j_n)}\right) = p\left(x_k^{(i)} / x^{(j_1)}, x^{(j_2)}, ..., x^{(j_{n+1})}\right), \quad (1.3.1)$$

где $i > j_1 > j_2 > \ldots > j_n > j_{n+1}$.

Другими словами, условная вероятность некоторого исхода x_k в *i*-м испытании, когда известны исходы в предыдущих *n* испытаниях, не зависит от более ранних исходов.

Далее, в Главе 2 будут достаточно подробно изложены основные результаты, относящиеся к теории марковских цепей. В данный момент только отметим, что в марковском источнике последовательность, состоящая из *n* элементов $\left(x^{(i)}, x^{(i-1)}, \ldots, x^{(i-n+1)}\right)$, определяет некоторое *состояние* S_q источника $\left(q = 1, \ldots, l^n\right)$, в котором вероятность выбора *k*-го элемента равна $p\left(x_k / S_q\right)$. При известных вероятностях $p\left(x_k / S_q\right)$ могут быть вычислены вероятности $P(S_q)$ каждого из состояний S_q. Тогда безусловные вероятности $p(x_k)$, равны[1]

$$p\left(x_k\right) = \sum_{q=1}^{l^n} p\left(S_q\right) p\left(x_k / S_q\right) \quad (1.3.2)$$

Величина

$$H\left(X / S_q\right) = -\sum_{k=1}^{l} p\left(x_k / S_q\right) \log p\left(x_k / S_q\right) \quad (1.3.3)$$

представляет из себя энтропию источника, находящегося в S_q-м состоянии, поэтому безусловная энтропия источника $H(X)$ может быть найдена путём усреднения $H\left(X / S_q\right)$ по всем возможным состояниям:

$$H\left(X\right) = -\sum_{q=1}^{l^n} \sum_{k=1}^{l} P\left(S_q\right) p\left(x_k / S_q\right) \log p\left(x_k / S_q\right) =$$

[1] Безусловные вероятности $p(x_k)$ существуют при условии *эргодичности* источника, выполняемом для всех источников, представляющих практический интерес.

$$= -\sum_{q=1}^{l^n}\sum_{k=1}^{l} p\left(x_k, S_q\right)\log p\left(x_k / S_q\right), \qquad (1.3.4)$$

где $p\left(x_k, S_q\right)$ — вероятность совместного события x_k и S_q.

Покажем, что наличие вероятностных связей между элементами источника сообщений уменьшает энтропию такого источника.

Пусть $H(X)$ — энтропия источника с памятью, определяемая согласно (1.3.4), а $H^*(X)$ — энтропия аналогичного источника, в котором элементы x_k ($k = 1, \ldots, l$) выбираются независимо друг от друга. Рассмотрим разность

$$H(X) - H^*(X) = -\sum_{q=1}^{l^n}\sum_{k=1}^{l} p(x_k, S_q)\log p\left(x_k / S_q\right) + \sum_{k=1}^{l} p(x_k)\log p(x_k) =$$

$$= \sum_{q=1}^{l^n}\sum_{k=1}^{l} p(x_k, S_q)\log\frac{p(x_k)}{p(x_k / S_q)} = \log e\sum_{q=1}^{l^n}\sum_{k=1}^{l} p(x_k, S_q)\ln\frac{p(x_k)}{p(x_k / S_q)}.$$

Используя неравенство $\ln x \le x - 1$, получаем

$$H(X) - H^*(X) \le \log e\sum_{q=1}^{l^n}\sum_{k=1}^{l} p\left(x_k, S_q\right)\left[\frac{p\left(x_k\right)}{p\left(x_k / S_q\right)} - 1\right] = 0.$$

Итак, $H(X) \le H^*(x)$, причём равенство достигается только при условии $p(x_k) = p(x_k / S_q)$, т. е. когда элементы выбираются независимо друг от друга.

Для характеристики источника сообщений X представляет интерес сравнение энтропии $H(X)$ этого источника с максимально возможным значением $H_{\max} = \log l$ при заданном объёме l канального алфавита. С этой целью вводится понятие *избыточности источника* $\rho(X)$:

$$\rho(X) = 1 - H(X) / H_{\max} = 1 - H(X) / \log l. \qquad (1.3.5)$$

Из рассмотренных выше свойств энтропии источника следует, что причиной избыточности могут являться неодинаковые вероятности выбора элементов сообщения, а также наличие вероятностных связей между элементами.

Для источника с фиксированной скоростью важной характеристикой является *производительность*, т. е. среднее количество информации в единицу времени. Если в среднем выбор каждого сообщения занимает время T, то производительность источника

$$H'(X) = \frac{H(X)}{T}. \qquad (1.3.6)$$

Наконец, ещё одной важной характеристикой является *объем передаваемой информации Q*, определяемый как произведение количества *N* сгенерированных источником символов сообщения на энтропию источника *H*(*X*) (с учетом всех вероятностных связей):

$$Q = NH(X). \qquad (1.3.7)$$

Рассмотрим примеры определения информационных характеристик источников дискретных сообщений.

Простейшим примером источника является источник без памяти, аналогичный тому, что представлен в табл. 1.1.

Таблица 1.1

Пример дискретного источника без памяти

x_k	x_1	x_2	x_3	x_4
$p(x_k)$	0,2	0,3	0,4	0,1

Как нетрудно убедиться, энтропия такого четырехэлементного источника составляет 1,847 бит, а его избыточность равна 0,076, что, как будет видно из дальнейшего, является достаточно малой величиной.

Пусть теперь x_k — значения случайной величины, распределённой по биномиальному закону:

$$p(x_k) = p_L(k) = C_L^k p^k (1-p)^{L-k},\ k = 0,1,\ldots,L, \qquad (1.3.8)$$

где p — вероятность единичного исхода.

Заметим, прежде всего, что в данном случае алфавит источника состоит из $L = l + 1$ элемента. Далее, поскольку элементы источника независимы, его энтропия равна

$$H = -\sum_{k=0}^{l} p_l(k) \log p_l(k) = -\sum_{k=0}^{L} C_L^k p^k (1-p)^{L-k} \log C_L^k -$$

$$-\sum_{k=0}^{L} C_L^k p^k (1-p)^{L-k} \log p^k - \sum_{k=0}^{L} C_L^k p^k (1-p)^{L-k} \log (1-p)^{L-k}.$$

Так как $\sum_{k=0}^{L} k C_L^k p^k (1-p)^{L-k} = Lp$ (математическое ожидание числа исходов) и $\sum_{k=0}^{L} C_L^k p^k (1-p)^{L-k} = 1$ (сумма вероятностей всех исходов),

$$H = -\sum_{k=0}^{L} C_L^k p^k (1-p)^{L-k} \log C_L^k - L\left[p \log p + (1-p)\log(1-p) \right].$$

В приведённой сумме первое и последнее слагаемые обнуляются, поскольку при $k = 0$ и $k = L$ значение $\log C_L^k = 0$, поэтому энтропия источника окончательно имеет следующий вид:

$$H = -\sum_{k=1}^{L-1} C_L^k p^k (1-p)^{L-k} \log C_L^k - L\left[p \log p + (1-p)\log(1-p) \right]. \qquad (1.3.9)$$

Например, для $p = 0{,}5$ и $L = 5$ энтропия равна $H = 2{,}20$ бит. Тогда избыточность источника составляет

$$\rho = 1 - H / \log_2 (L+1) = 0{,}15 .$$

Влияние на информационные характеристики источников статистической зависимости его элементов вначале проиллюстрируем на простейшем примере. Рассмотрим двоичный источник X с элементами A и B, память которого простирается лишь на один соседний символ, т. е. источник описывается цепью Маркова первого порядка с матрицей $\boldsymbol{\pi}$ переходных вероятностей вида

$$\boldsymbol{\pi} = \begin{bmatrix} P(A/A) & P(B/A) \\ P(A/B) & P(B/B) \end{bmatrix}.$$

Если на предыдущем шаге выбранным элементом был A, то источник с вероятностью $P(A/A)$ повторяет на своём выходе этот же элемент и с вероятностью $P(B/A) = 1 - P(A/A)$ изменяет значение выхода на противоположное. Аналогичное поведение имеет место, если на предыдущем шаге выбранным элементом был B. Заметим, что матрица, у которой сумма элементов в каждой строке равна единице, называется *вероятностной* или *стохастической*. Если к тому же единице равна сумма элементов в каждой строке, то такая матрица называется *бистохастической*.

Описанный источник X имеет два состояния, определяемые последним выбранным элементом: $S_1 = (x_k, A)$ и $S_2 = (x_k, B)$, где x_k — текущий выбираемый элемент ($k = 1, 2$).

Найдём безусловные вероятности $P(A)$ и $P(B)$ выбора элементов A и B. Из уравнения

$$P(A) = P(A/A)P(A) + P(A/B)\left(1 - P(A)\right)$$

находим

$$P(A)=\frac{P(A/B)}{1-P(A/A)+P(A/B)},\quad P(B)=1-P(A). \qquad (1.3.10)$$

Из соотношений (1.3.10) следует, что для “симметричного” источника, т. е. источника, для которого $P(A/A) = P(B/B)$ и $P(A/B) = P(B/A)$, безусловные вероятности одинаковы и равны 0,5. Пусть, например,

$$P(A/A) = 0{,}8;\ P(A/B) = 0{,}2;\ P(B/A) = 0{,}2 \text{ и } P(B/B) = 0{,}8.$$

Тогда

$$P(A) = P(B) = 0{,}5,$$

и такие же вероятности, определяемые как

$$P(S_1) = P\left(x_k, A\right) = P(A/A)P(A) + P(B/A)P(A);$$

$$P(S_2) = 1 - P(S_1),$$

имеют состояния S_1 и S_2. Поэтому энтропия такого источника равна $H = 0{,}72$ бит, а избыточность, вызванная вероятностными связями между элементами, составляет $\rho = 0{,}28$. В то же время, энтропия двоичного источника с независимыми элементами, вероятности появления которых равны найденным $P(A)$ и $P(B)$, достигает своего максимального значения, равного единице.

Пусть теперь источник не является “симметричным”, т. е. описывается несимметричной матрицей $\boldsymbol{\pi}$. Например, для значений переходных вероятностей

$$P(A/A) = 0{,}3;\ P(A/B) = 0{,}1;\ P(B/A) = 0{,}7 \text{ и } P(B/B) = 0{,}9$$

безусловные вероятности появления элементов A и B оказываются равны соответственно $P(A) = 0{,}125$ и $P(B) = 0{,}875$. Такими же оказываются вероятности двух возможных состояний источника, что приводит к значениям энтропии $H = 0{,}52$ бита и избыточности $\rho = 0{,}48$.

Рассмотренные примеры касались модельных источников, для которых задавались теоретические вероятности появления символов. Проиллюстрируем теперь рассмотренные выше определения и свойства на более конкретном примере, где вместо теоретических вероятностей будут фигурировать эмпирические частоты.

Пусть источник сгенерировал следующее сообщение:

КОЛОКОЛ_ОКОЛО_КОЛОКОЛА

Разумеется, с точки зрения статистики русского языка (см. разд. 2.9) данный пример не является типичным, прежде всего, из-за малого объема вы-

борки. Тем не менее, на нем можно вполне отчетливо проследить информационные характеристики источников осмысленной письменной речи.

Для начала, предположим, что источник генерирует элементы независимым образом. Тогда, для пятиэлементного источника ($l = 5$) на основании анализа $N = 22$ сгенерированных элементов имеет место следующая статистика (табл. 1.2).

Таблица 1.2

Статистика встречаемости символов источника

x_1 = ‘О’	$P(x_1) = 9/22 = 0{,}409$
x_2 = ‘Л’	$P(x_2) = 5/22 = 0{,}227$
x_3 = ‘К’	$P(x_3) = 5/22 = 0{,}227$
x_4 = ‘_’ (пробел)	$P(x_4) = 2/22 = 0{,}091$
x_5 = ‘А’	$P(x_5) = 1/22 = 0{,}046$

Заметим, что при переходе от обыкновенных дробей к десятичным вследствие погрешностей округления последнего значащего разряда может случиться так, что сумма всех вероятностей (точнее, статистических частот) оказывается отличной от единицы. Такой эффект особенно заметен при многошаговых расчетах с использованием средств вычислительной техники. Одним из возможных способов решения этой проблемы является изменение на каком-либо этапе вычислений значения одной из p_k в “нужную” сторону[1]. При этом проверка условия нормировки $\sum_{(k)} p_k = 1$ должна быть достаточно частой.

Полученный в табл. 1.2 источник X характеризуется энтропией

$$H(X) = -\big(0{,}409\log 0{,}409 + 2\cdot 0{,}227\log 0{,}227 +$$
$$+0{,}091\log 0{,}091 + 0{,}049\log 0{,}049\big) = 2{,}03 \text{ бит}$$

и избыточностью

$$\rho = 1 - \frac{2{,}03}{\log 5} = 0{,}13.$$

Общий объем информации при наблюдении за таким источником составляет

$$Q = 22 \cdot 2{,}03 = 44{,}66 \text{ бит}.$$

[1] При этом, однако, существует опасность замены малых, но ненулевых значений нулевыми, что может привести к гораздо более серьёзным последствиям.

Откажемся теперь от предположения о независимости источника и будем считать, что генерируются зависимые элементы. Простейшей моделью такого источника с памятью является учет только одного предшествующего символа. В этом случае при $l = 5$ и $n = 1$, как легко видеть, источник может находиться в одном из $l^n = 5$ состояний $S_q = (x_k, x_q)$, заключающихся в том, что на каждом текущем шаге генерируется произвольный элемент x_k, а на предыдущем шаге был сгенерирован определенный элемент x_q, $q = 1, \ldots, 5$. Конкретно, имеют место следующие состояния:

$S_1 = (x_k, \text{‘О’}), S_2 = (x_k, \text{‘Л’}), S_3 = (x_k, \text{‘К’}), S_4 = (x_k, \text{‘\_’}), S_5 = (x_k, \text{‘А’})$.

Оценка вероятностей $p(S_q)$ состояний источника, а также условных вероятностей $p(x_k/S_q)$ появления символов основана на подсчете статистики всевозможных двухбуквенных сочетаний, или *диграмм* (табл. 1.3)[1].

Таблица 1.3

Статистика встречаемости двухбуквенных сочетаний

Состояние	Сочетания	Встречаемость
S_1	ОЛ	5
	ОК	3
	О_	1
S_2	ЛО	3
	Л_	1
	ЛА	1
S_3	КО	5
S_4	_О	1
	_К	1
S_5	–	–

Обратим внимание на наличие своего рода “концевых эффектов”, проявляющихся, в частности, в возникновении “пустого” состояния S_5. Их влияние можно снять, например, путём введения в начале и конце заранее предопределённых элементов, как это делается при формировании сигналов при пакетной передаче, либо вообще не принимая во внимание соот-

[1] Поскольку по правилам русского языка чтение букв происходит слева направо, в диграммах текущий символ стоит на последнем по порядку следования месте.

ветствующие элементы. Однако, сохраняя “чистоту эксперимента”, в данном случае не будем этого делать. В Главе 2 при изучении свойств марковских цепей этому вопросу будет уделено большее внимание.

Итак, общее количество двухбуквенных сочетаний равно $M = 21$, тогда вероятности состояний имеют следующие значения:

$p(S_1) = 9/21$; $p(S_2) = 5/21$; $p(S_3) = 5/21$; $p(S_4) = 2/21$; $p(S_5) = 0$.

Далее, имея значения вероятностей состояний, можно приступить к нахождению условных энтропий. Для этого из представленных в табл. 1.3 данных сформируем переходную матрицу $\boldsymbol{\pi}$ условных вероятностей

$$\boldsymbol{\pi} = \begin{pmatrix} p(x_1/S_1) & p(x_2/S_1) & p(x_3/S_1) & p(x_4/S_1) & p(x_5/S_1) \\ p(x_1/S_2) & p(x_2/S_2) & p(x_3/S_2) & p(x_4/S_2) & p(x_5/S_2) \\ p(x_1/S_3) & p(x_2/S_3) & p(x_3/S_3) & p(x_4/S_3) & p(x_5/S_3) \\ p(x_1/S_4) & p(x_2/S_4) & p(x_3/S_4) & p(x_4/S_4) & p(x_5/S_4) \\ p(x_1/S_5) & p(x_2/S_5) & p(x_3/S_5) & p(x_4/S_5) & p(x_5/S_5) \end{pmatrix}.$$

Поскольку, находясь в каком-либо состоянии S_q, источник с единичной вероятностью генерирует какой-нибудь элемент x_k, во всех строках кроме той, что соответствует “пустому” состоянию, вызванному концевыми эффектами, сумма элементов равна единице. Следовательно, матрица $\boldsymbol{\pi}$ принимает следующий числовой вид:

$$\boldsymbol{\pi} = \begin{pmatrix} 0 & 5/9 & 3/9 & 1/9 & 0 \\ 3/5 & 0 & 0 & 1/5 & 1/5 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Прежде чем приступить к нахождению условных энтропий, найдем значения безусловных вероятностей $p(x_k)$ появления элементов для источника с памятью и сравним полученные значения с аналогичными значениями для источника без памяти. Усредняя матрицу $\boldsymbol{\pi}$ по каждому столбцу, получаем

$$p(x_1) = \sum_{q=1}^{5} p(x_1/S_q)\, p(S_q) =$$

$$= 0 \cdot \frac{9}{21} + \frac{3}{5} \cdot \frac{5}{21} + 1 \cdot \frac{5}{21} + \frac{1}{2} \cdot \frac{2}{21} + 0 \cdot 0 = 0,409$$

и, аналогично,

$$p(x_2) = 0{,}238;\ p(x_3) = 0{,}191;\ p(x_4) = 0{,}096;\ p(x_5) = 0{,}048.$$

Как показывает сравнение полученных вероятностей со значениями, представленными в табл. 1.2, безусловные значения вероятностей появления элементов рассматриваемого источника с зависимыми элементами практически совпадают с аналогичными значениями источника без памяти.

Теперь, имея матрицу условных вероятностей и используя соотношения (1.3.3) и (1.3.4), получим набор условных энтропий $H(X / S_q)$ и безусловную энтропию $H(X)$ источника с памятью. Последовательно находя, имеем:

$$H(X / S_1) = -\sum_{k=1}^{5} p\left(x_k / S_1\right) \log p\left(x_k / S_1\right) =$$

$$= -\left(\frac{5}{9}\log\frac{5}{9} + \frac{3}{9}\log\frac{3}{9} + \frac{1}{9}\log\frac{1}{9}\right) = 1{,}351 \text{ бит};$$

$$H(X / S_2) = 1{,}351 \text{ бит};\ H(X / S_3) = 0{,}000 \text{ бит};$$

$$H(X / S_4) = 1{,}000 \text{ бит};\ H(X / S_5) = 0{,}000 \text{ бит}.$$

Отсюда, безусловная энтропия равна

$$H(X_1) = -\sum_{q=1}^{5} H\left(H / S_q\right) P\left(S_q\right) =$$

$$= 1{,}351 \cdot 0{,}429 + 1{,}324 \cdot 0{,}238 + 0 \cdot 0{,}238 + 1 \cdot 0{,}095 + 0 \cdot 0 = 0{,}977 \text{ бит}.$$

Общий объем информации при наблюдении за источником составляет

$$Q = 22 \cdot 0{,}977 = 21{,}494 \text{ бит},$$

а его избыточность равна

$$\rho = 1 - \frac{0{,}977}{\log 5} = 0{,}58.$$

Итак, переход к источнику с памятью с учетом даже одного предыдущего символа приводит, как это следует из приведенного анализа, к уменьшению энтропии и существенному возрастанию избыточности. Еще меньшей оказывается энтропия при дальнейшем увеличении “длины” памяти n, т. е. при дальнейшем учете предыдущих символов. Однако с учетом каждого последующего символа значение энтропии снижается на все меньшую величину. Более того, как будет рассмотрено далее, для реальных источников, начиная с некоторого момента, энтропия вообще перестает зависеть от числа учитываемых символов.

1.4. КОЛИЧЕСТВО ИНФОРМАЦИИ, СОДЕРЖАЩЕЙСЯ В НЕПРЕРЫВНОМ СООБЩЕНИИ

Непрерывные сообщения источника и соответствующие им сигналы характеризуются тем, что они могут принимать бесконечное число возможных форм. Попробуем получить меру количества информации для непрерывного сообщения, исходя из уже известной характеристики для дискретного распределения, используя механизм дискретизации с последующим предельным переходом.

Рассмотрим случайную величину ξ, принимающую непрерывное множество (континуум) значений X. Пусть $w_\xi(x)$ — плотность распределения вероятностей значений величины ξ. Разобьем множество значений X на дискретное число интервалов шириной Δx. Вероятность p_k того, что значение величины ξ находится в пределах интервала $[k\Delta x; (k+1)\Delta x]$ ($k = \ldots, -1, 0, 1, \ldots$) равна

$$p_k = P\left(x \in \left[k\Delta x;\ (k+1)\Delta x\right]\right) = \int_{k\Delta x}^{(k+1)\Delta x} w_\xi(x)\,dx .$$

Полагая, что плотность вероятности $w_\xi(x)$ является достаточно гладкой функцией, заменим последний интеграл приближенным значением $w_\xi\left(x_k^*\right)\Delta x$, где $x_k^* \in \left[k\Delta x;(k+1)\Delta x\right]$. Тогда в соответствии с (1.3) можно записать выражение для энтропии $\tilde{H}(X)$ дискретизированной величины, или, как ее часто называют, *энтропии на один отсчет*:

$$\tilde{H}(X) = -\sum_{(k)} p_k \log p_k \approx -\sum_{(k)} w_\xi\left(x_k^*\right)\Delta x \log\left[w_\xi\left(x_k^*\right)\Delta x\right] =$$

$$= -\sum_{(k)} w_\xi\left(x_k^*\right)\Delta x \log w_\xi\left(x_k^*\right) - \log\Delta x \sum_{(k)} w_\xi\left(x_k^*\right)\Delta x .$$

Устремим $\Delta x \to 0$, при этом по определению интеграла Римана имеем

$$\lim_{\Delta x\to 0}\sum_{(k)} w_\xi\left(x_k^*\right)\Delta x = \int_X w_\xi(x)\,dx = 1$$

и

$$\lim_{\Delta x\to 0}\sum_{(k)} w_\xi\left(x_k^*\right)\log w_\xi\left(x_k^*\right)\Delta x = \int_X w_\xi(x)\log w_\xi(x)\,dx .$$

Тогда энтропия непрерывного сообщения $H(X) = \lim\limits_{\Delta x\to 0}\tilde{H}(X)$ имеет вид

$$H(X) = -\int_{-\infty}^{\infty} w_\xi(x) \log w_\xi(x)\, dx - \lim_{\Delta x \to 0} \log \Delta x . \qquad (1.4.1)$$

Из последнего выражения видно, что энтропия, т. е. степень неопределенности появления любого значения непрерывного сообщения бесконечно велика, ибо величина $-\lim_{\Delta x \to 0} \log \Delta x$ равна бесконечности, поэтому (1.4.1) не может иметь практического применения. Этого и следовало ожидать: неопределенность реализации одного из бесконечного и несчетного множества состояний бесконечно велика. Тем не менее, полученное выражение несет в себе важную составляющую

$$h(X) = -\int_{-\infty}^{\infty} w_\xi(x) \log w_\xi(x)\, dx , \qquad (1.4.2)$$

которая называется *дифференциальной энтропией*.

Попробуем придать этой характеристике информационный смысл. Заметим, что невозможность практического применения соотношения (1.4.1) означает невозможность введения для непрерывных сообщений *абсолютной* информационной характеристики, однако можно попытаться ввести *относительную* меру, т. е. сравнивать неопределенность выбора реализации непрерывного сообщения с некоторым стандартом. В качестве такого стандарта можно взять неопределенность какого-либо простого распределения, например, равномерного распределения в интервале шириной c. Произведя для такого распределения описанный выше процесс квантования, получим энтропию на один отсчёт $\tilde{H}_c(X)$, равную

$$\tilde{H}_c(X) = \log \frac{c}{\Delta x} = \log c - \log \Delta x .$$

Будем характеризовать неопределенность выбора непрерывной величины ξ величиной $h_c(X)$, к которой в пределе стремится разность энтропий $\tilde{H}(X)$ и $\tilde{H}_c(X)$:

$$h_c(X) = \lim_{\Delta x \to 0} \left[\tilde{H}(X) - \tilde{H}_c(X) \right] =$$

$$= -\int_{-\infty}^{\infty} w_\xi(x) \log w_\xi(x)\, dx - \log c = -\int_{-\infty}^{\infty} w_{\boldsymbol{\xi}}(x) \log c w_\xi(x)\, dx .$$

В частности, полагая $c = 1$, т. е. рассматривая в качестве стандарта равномерное распределение в единичном интервале, получим выражение (1.4.2).

Итак, информационной характеристикой непрерывного сообщения является дифференциальная энтропия — относительная величина, показывающая насколько неопределенность появления рассматриваемого сообщения больше или меньше неопределенности появления сообщения, значения которого равновероятны на единичном интервале.

Разумеется, предложенный подход содержит изрядную долю условности, и может показаться чересчур искусственным с практической точки зрения. Однако, как будет показано далее, на определенном этапе построения конструкции информационного описания источников сообщений от введения вспомогательного равновероятного источника можно будет отказаться.

Дифференциальная энтропия обладает свойствами, во многом аналогичными, но в ряде случаев и отличными от свойств энтропии дискретных величин. В частности, она может принимать отрицательные значения.

Заметим, что все сказанное допускает обобщение и на случай многомерных непрерывных распределений. В этом случае под ξ и η следует понимать многомерные векторы $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ и $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$. При этом все интегралы соответственно становятся кратными (n-мерными), а стандартом сравнения является неопределенность распределения, равномерного в многомерном единичном объёме $V = c^n$. Подробнее об этом будет сказано в следующем разделе.

1.5. ПРИНЦИП ЭКСТРЕМУМА ДИФФЕРЕНЦИАЛЬНОЙ ЭНТРОПИИ И ЭКСТРЕМАЛЬНЫЕ РАСПРЕДЕЛЕНИЯ

Часто при решении различных теоретических и практических задач возникает ситуация, когда известны лишь некоторые ограничения, накладываемые на случайную величину ξ. Чаще всего такими ограничениями являются значения моментов распределений, но могут быть и другие условия, например, ограничение области возможных значений случайной величины. При этом требуется, не привлекая дополнительных сведений, задаться некоторым модельным распределением вероятностей, обладающим заданными свойствами. Такая задача возникает, например, при выборе “наилучшего” по тому или иному критерию распределения значений искусственно создаваемой помехи при заданной мощности сигналов.

Разумеется, заданным ограничениям могут удовлетворять различные распределения (более того, нетрудно понять, что таких распределений бесконечное множество), поэтому целесообразно ставить вопрос о нахождении такого распределения, которое помимо того, что удовлетворяет заданным ограничениям, является наилучшим с точки зрения некоторого критерия оптимальности. В рамках теории информации в качестве такого критерия предлагается *принцип экстремума энтропии*, согласно которому из всех подходящих распределений следует выбирать то, которое обладает экстремальной, например, максимальной энтропией.

Заметим, что в такой постановке задача принципиально отличается от задач, связанных с поиском экстремума функции нескольких переменных (например, задачи нахождения дискретного распределения, имеющего максимальную энтропию), а является вариационной задачей по оптимизации функционалов.

Пусть известно, что область возможных значений величины ξ ограничена интервалом $[x_1; x_2]$. Найдем распределение, обладающее при этом условии максимальной энтропией. Для этого необходимо решить следующую вариационную задачу: найти функцию $w(x)$, обеспечивающую максимум функционала

$$h(X) = -\int_{x_1}^{x_2} w(x) \log w(x) dx \qquad (1.5.1)$$

при дополнительном условии нормировки плотности вероятности:

$$\int_{x_1}^{x_2} w(x) dx = 1 .$$

Как известно [3], если требуется найти экстремум функционала вида

$$J\left[w(x)\right] = \int_{x_1}^{x_2} F\left[x, w(x)\right] dx \qquad (1.5.2)$$

при дополнительных ограничениях

$$\int_{x_1}^{x_2} ц_1\left[x, w(x)\right] dx = c_1 , \ \ldots , \ \int_{x_1}^{x_2} ц_n\left[x, w(x)\right] dx = c_n , \qquad (1.5.3)$$

то функция, обеспечивающая экстремум, является решением дифференциального уравнения

$$\frac{\partial F}{\partial w}+\lambda_1\frac{\partial \varphi_1}{\partial w}+\ldots+\lambda_n\frac{\partial \varphi_n}{\partial w}=0\,, \qquad (1.5.4)$$

где $\lambda_1, \ldots, \lambda_n$ — множители Лагранжа, значения которых определяются подстановкой $w(x)$, являющейся решением уравнения (1.5.4), в систему (1.5.3). Применительно к данной задаче (a — произвольное основание)

$$F\left[x, w(x)\right]dx=-w(x)\log w(x)\equiv -w(x)\log_a w(x)\,;$$

$$\varphi_1\left[x, w(x)\right]=w(x)\,,$$

поэтому

$$\frac{\partial F\left[x, w(x)\right]}{\partial w(x)}=-\frac{1}{\ln a}\left[\log w(x)+1\right]=-\log e-\log w(x)\,;$$

$$\frac{\partial \varphi_1}{\partial w(x)}=1\,.$$

Тогда, подставляя эти результаты в уравнение (1.5.4), имеем:

$$-\log e-\log w\left(x\right)+\lambda=0\,,$$

откуда

$$w(x)=\exp\left(\lambda-\log e\right).$$

Пользуясь условием нормировки, определим неизвестную константу λ:

$$1=\int_{x_1}^{x_2}\exp\left(\lambda-\log e\right)dx=\left(x_2-x_1\right)\exp\left(\lambda-\log e\right).$$

Итак,

$$w\left(x\right)=\begin{cases}\dfrac{1}{x_2-x_1}, & x\in\left[x_1;x_2\right]\\ 0, & x\notin\left[x_1;x_2\right]\end{cases}, \qquad (1.5.5)$$

т. е. максимальной энтропией при ограниченности возможных значений обладает равномерное распределение вероятностей.

Этот пример подчёркивает общность свойств дифференциальной энтропии и энтропии дискретного распределения, имеющего максимальную энтропию при равновероятных элементах. Кроме того, данный пример ещё раз оправдывает то, что в качестве стандарта неопределённости при введении дифференциальной энтропии была выбрана неопределённость равномерного распределения.

Пусть теперь область возможных значений случайной величины ξ неограниченна, и заданы значения её математического ожидания $\mathbf{E}[\xi] = a$ и дисперсии $\mathbf{D}[\xi] = \sigma^2$. При этом задача сводится к нахождению максимума функционала

$$h(X) = -\int_{-\infty}^{\infty} w(x) \log w(x) dx \tag{1.5.6}$$

при условии нормировки

$$\int_{-\infty}^{\infty} w(x) dx = 1, \tag{1.5.7}$$

а также дополнительных условиях

$$\int_{-\infty}^{\infty} x w(x) dx = a \tag{1.5.8}$$

и

$$\int_{-\infty}^{\infty} (x-a)^2 w(x) dx = \sigma^2 . \tag{1.5.9}$$

В этом случае функции F, φ_1, φ_2, и φ_3 имеют следующий вид:

$$F\left[x, w(x)\right] dx = -w(x) \log w(x);$$

$$\varphi_1\left[x, w(x)\right] = w(x);$$

$$\varphi_2\left[x, w(x)\right] = x w(x);$$

$$\varphi_3\left[x, w(x)\right] = (x-a)^2 w(x),$$

следовательно,

$$\frac{\partial F\left[x, w(x)\right]}{\partial w(x)} = -\left[\log e + \log w(x)\right];$$

$$\frac{\partial \varphi_1}{\partial w(x)} = 1;$$

$$\frac{\partial \varphi_2}{\partial w(x)} = x;$$

$$\frac{\partial \varphi_3}{\partial w(x)} = (x-a)^2 .$$

Подставляя эти выражения в (1.5.4), имеем:

$$-\log e - \log w(x) + \lambda_1 + \lambda_2 x + \lambda_3 (x-a)^2 = 0,$$

откуда

$$w(x)=\exp\left(\lambda_1+\lambda_2 x+\lambda_3\left(x-a\right)^2-\log e\right).$$

Использование условия нормировки (1.5.7) даёт:

$$1=\exp\left(\lambda_1-\log e\right)\int_{-\infty}^{\infty}\exp\left[\lambda_2 x+\lambda_3\left(x-a\right)^2\right]dx,$$

причём для сходимости интеграла константа λ_3 должна быть отрицательной, а поскольку

$$\int_{-\infty}^{\infty}\exp\left[\lambda_2 x+\lambda_3\left(x-a\right)^2\right]dx=\exp\left(\lambda_2 a\right)\int_{-\infty}^{\infty}\exp\left[\lambda_2 y+\lambda_3 y^2\right]dy=$$

$$=\exp\left(\lambda_2 a\right)\exp\left(-\frac{\lambda_2^2}{4\lambda_3}\right)\sqrt{\frac{\pi}{-\lambda_3}},$$

получаем

$$\exp\left(\lambda_1-\log e\right)=\sqrt{\frac{-\lambda_3}{\pi}}\exp\left(\frac{\lambda_2^2}{4\lambda_3}-\lambda_2 a\right). \qquad (1.5.10)$$

Использование (1.5.9) после несложных преобразований приводит к выражению вида

$$\frac{1}{-\lambda_3\sqrt{\pi}}\left[\frac{\sqrt{\pi}}{2}+\frac{\lambda_2^2\sqrt{\pi}}{4\left(-\lambda_3\right)}+\frac{\lambda_2}{2\sqrt{-\lambda_3}}\int_{-\infty}^{\infty}\exp\left(-z\right)dz\right]=\sigma^2, \qquad (1.5.11)$$

из которого следует, что константа λ_2 должна быть равна нулю, ибо в противном случае левая часть (1.5.11) оказывается бесконечной. Подставляя (1.5.11) в (1.5.10), получаем

$$\lambda_3=-\frac{1}{2\sigma^2},$$

т. е. в рассматриваемом случае экстремальное распределение

$$w\left(x\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left(x-a\right)^2}{2\sigma^2}\right] \qquad (1.5.12)$$

является нормальным. Заметим, что условие (1.5.8) не позволяет получить дополнительных связей между неизвестными константами: при его использовании образуется тождество. Разумеется, это вовсе не означает независимость результата от математического ожидания, поскольку диспер-

сия является *центральным* моментом, т. е. моментом относительно среднего значения.

Найдём дифференциальную энтропию полученного экстремального распределения (1.5.12). Имеем:

$$h(X)=-\frac{\log e}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}\exp\left[-(x-a)^2/2\sigma^2\right]\left[\ln\left(2\pi\sigma^2\right)^{-1/2}-(x-a)^2/\left(2\sigma^2\right)\right]dx,$$

а поскольку

$$\int_{-\infty}^{\infty}\exp\left(-b^2x^2\right)dx=\frac{\sqrt{\pi}}{|b|};$$

$$\int_{-\infty}^{\infty}x^2\exp\left(-b^2x^2\right)dx=\frac{\sqrt{\pi}}{2|b|^3},$$

получаем

$$h(X)=\frac{1}{2}\left[\ln\left(2\pi\sigma^2\right)+1\right]\log e=\frac{1}{2}\ln\left(2\pi\sigma^2e\right)\log e=\log\sqrt{2\pi e\sigma^2}\ . \quad (1.5.13)$$

Таким образом, дифференциальная энтропия нормального распределения зависит от его дисперсии и не зависит от его математического ожидания. Можно показать (см. задачу 1.14), что это свойство является общим для всех распределений.

Экстремальность нормального распределения для одномерных случайных величин наводит на мысль о возможности выполнения такого же свойства и для многомерных распределений.

Пусть $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_n)$ — совокупность случайных величин, описываемых совместной плотностью вероятности $w_{\xi}^{(n)}(\mathbf{x})=w_{\xi}^{(n)}\left(x^{(1)},\ldots,x^{(n)}\right)$, так что по ней могут быть определены собственные и смешанные моменты различных порядков, в частности, математические ожидания

$$m_k\equiv\mathbf{E}\left[\xi_k\right]=\int_{X^n}x^{(k)}w_{\xi}^{(n)}(\mathbf{x})d\mathbf{x}=\int_{X}xw_{\xi_k}(x)dx\,,\ k=1,\ldots,n, \quad (1.5.14)$$

и ковариации

$$b_{kj}\equiv\mathrm{Cov}\left[\xi_k,\xi_j\right]=\int_{X^n}\left(x^{(k)}-\mathbf{E}[\xi_k]\right)\left(x^{(j)}-\mathbf{E}[\xi_j]\right)w_{\xi}^{(n)}(\mathbf{x})d\mathbf{x}=$$

$$=\int_{X^2}\left(x^{(k)}-\mathbf{E}[\xi_k]\right)\left(x^{(j)}-\mathbf{E}[\xi_j]\right)w_{\xi}^{(2)}\left(x^{(k)},x^{(j)}\right)dx^{(k)}dx^{(j)}\,. \quad (1.5.15)$$

Матрица

$$\mathbf{K} = \left\| b_{kj} \right\| = \left\| \mathrm{Cov}\left[\xi_k, \xi_j \right] \right\|, \; k, j = 1, \ldots, n,$$

элементами которой являются ковариации пар, называется *корреляционной матрицей* совокупности случайных величин. Обозначим через $\mathbf{K}^{-1}$ матрицу, обратную матрице $\mathbf{K}$, т. е. такую, для которой матричное произведение $\mathbf{K}^{-1}\mathbf{K}$ равно единичной матрице n-го порядка $\mathbf{I}_n$.

Обратная матрица $\mathbf{A}^{-1}$ может быть определена для любой исходной квадратной $n \times n$-матрицы $\mathbf{A}$, если она *невырождена*, т. е. если её определитель $\det \mathbf{A}$ отличен от нуля. При этом элементы α_{kj} $(k, j = 1, \ldots, n)$ обратной матрицы равны

$$\alpha_{kj} = \frac{A_{jk}}{\det \mathbf{A}},$$

где $A_{kj} = (-1)^{k+j} \overline{M}_j^k$ — алгебраическое дополнение элемента a_{kj} исходной матрицы, получающееся (с учётом знака) посредством вычёркивания в определителе $\det \mathbf{A}$ k-й строки и j-го столбца.

Пусть β_{kj} $(k, j = 1, \ldots, n)$ — элементы матрицы $\mathbf{K}^{-1}$. Тогда определим n-мерное нормальное распределение посредством задания многомерной плотности $w_n(\mathbf{x})$ следующего вида:

$$w_{\mathrm{norm}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \mathbf{K}}} \exp\left[-\frac{1}{2} \sum_{k=1}^{n} \sum_{j=1}^{n} \left(x^{(k)} - m_k \right) \left(x^{(j)} - m_j \right) \beta_{kj} \right], \quad (1.5.16)$$

где $\det \mathbf{K}$ — определитель матрицы $\mathbf{K}$.

Запись многомерного нармального распределения удобна в матричной форме:

$$w_{\mathrm{norm}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \mathbf{K}}} \exp\left[-\frac{1}{2} (\mathbf{x} - \mathbf{m}_\xi)^T \mathbf{K}^{-1} (\mathbf{x} - \mathbf{m}_\xi) \right], \quad (1.5.16a)$$

где $\mathbf{m}_\xi$ — вектор математических ожиданий.

Для произвольного многомерного распределения $w_\xi^{(n)}(\mathbf{x})$ можно определить дифференциальную энтропию

$$h(\xi_1, \ldots, \xi_n) = -\int_{X^n} w_\xi^{(n)}(\mathbf{x}) \log w_\xi^{(n)}(\mathbf{x}) d\mathbf{x}, \quad (1.5.17)$$

где интегрирование ведётся по всему n-мерному распределению.

Выделим класс функций $w_\xi^{(n)}(\mathbf{x})$ с заданными математическими ожиданиями (1.5.14) и ковариациями (1.5.15). Тогда, исходя из аналогии с одномерным случаем, сформулируем и докажем следующее свойство.

При заданной корреляционной матрице $\mathbf{K}$ *для любой плотности вероятности выполняется соотношение*

$$h(\xi_1,\ldots,\xi_n) \le \frac{n}{2}\log 2\pi e + \frac{1}{2}\log\det\mathbf{K}\,, \qquad (1.5.18)$$

причём равенство достигается лишь в том случае, когда $w_{\xi}^{(n)}(\mathbf{x})$ *является плотностью вероятности n-мерного нормального распределения.*

Заметим, что, как и в одномерном случае, дифференциальная энтропия (1.5.17) зависит от математических ожиданий только опосредованно через ковариации (см. задачу 1.14).

Докажем справедливость (1.5.18), используя неравенство для логарифма. Введём функцию

$$h_{\text{norm}}(\xi_1,\ldots,\xi_n) = \frac{n}{2}\log 2\pi e + \frac{1}{2}\log\det\mathbf{K}\,, \qquad (1.5.19)$$

которая представляет собой дифференциальную энтропию многомерного нормального распределения. Действительно,

$$h_{\text{norm}}(\xi_1,\ldots,\xi_n) = -\int_{X^n} w_{\text{norm}}(\mathbf{x})\log w_{\text{norm}}(\mathbf{x})d\mathbf{x} = \frac{n}{2}\log 2\pi + \frac{1}{2}\log\det\mathbf{K} +$$

$$+\frac{\log e}{2}\sum_{k=1}^{n}\sum_{j=1}^{n}\beta_{kj}\int_{X^n} w_{\text{norm}}(\mathbf{x})\left(x^{(k)}-m_k\right)\left(x^{(j)}-m_j\right)d\mathbf{x} =$$

$$= \frac{n}{2}\log 2\pi + \frac{1}{2}\log\det\mathbf{K} + \frac{\log e}{2}\sum_{k=1}^{n}\sum_{j=1}^{n}\beta_{kj}b_{kj}\,.$$

Из определения обратной матрицы и свойства симметричности корреляционной матрицы $\mathbf{K}$ нетрудно видеть, что

$$\sum_{j=1}^{n}\beta_{kj}b_{kj} = \begin{cases}1, & k=j\\ 0, & k\ne j\end{cases},$$

поэтому

$$\sum_{k=1}^{n}\sum_{j=1}^{n}\beta_{kj}b_{kj} = n\,,$$

и h_{norm} равно (1.5.19).

Теперь, используя неравенство для логарифма, имеем:

$$h(\xi_1,\ldots,\xi_n) - h_{\text{norm}}(\xi_1,\ldots,\xi_n) = -\int_{X^n} w(\mathbf{x})\log w(\mathbf{x})d\mathbf{x} + \int_{X^n} w(\mathbf{x})\log w_{\text{norm}}(\mathbf{x})d\mathbf{x} =$$

$$= \int_{X^n} w(\mathbf{x}) \log \frac{w_{\text{norm}}(\mathbf{x})}{w(\mathbf{x})} d\mathbf{x} \le \int_{X^n} w(\mathbf{x}) \left[\frac{w_{\text{norm}}(\mathbf{x})}{w(\mathbf{x})} - 1 \right] d\mathbf{x} = 0,$$

причём равенство достигается лишь в том случае, когда $w(\mathbf{x}) = w_{\text{norm}}(\mathbf{x})$.

Вычисление дифференциальной энтропии для многомерных распределений в общем случае представляет собой достаточно сложную задачу, если только, конечно, входящие в $\boldsymbol{\xi}$ случайные величины не являются независимыми. Даже при подсчёте $h(\boldsymbol{\xi})$ для нормального распределения согласно (1.5.19) необходимо знать ковариации между всеми компонентами, составляющими случайный вектор $(\xi_1, \ldots, \xi_n)$, что для больших значений n требует трудоёмких вычислений. Ещё более задача может усложниться при нахождении характеристик взаимной информации между многомерными случайными последовательностями.

Однако существует возможность преодоления многих трудностей с помощью специального *ортогонального* преобразования [11] вектора $\boldsymbol{\xi}$, при котором происходит *декорреляция* отдельных компонент случайного вектора, т. е. обнуление всех элементов корреляционной матрицы, не стоящих на главной диагонали. Это, разумеется, не приводит к полной взаимной *независимости* компонент (кроме нормального распределения, для которого некоррелированность и независимость эквивалентны), но позволяет хотя бы приближённо рассчитывать информационные характеристики на основе факторизации многомерного распределения.

Ортогональное преобразование является *обратимым детерминированным преобразованием*, а следовательно, оно не изменяет информационные характеристики, в частности, количество информации, содержащееся в многомерной случайной величине.

Возможность ортогонального преобразования базируется на *симметричности и неотрицательной знакоопределённости* корреляционной матрицы $\mathbf{K}$, означающее, что для любого ненулевого вектора $\mathbf{v}$ матричное произведение

$$\mathbf{v}\mathbf{K}\mathbf{v}^T \ge 0, \qquad (1.5.20)$$

где символ "T" означает транспонирование, т. е. перестановку местами строк и столбцов матрицы.

Справедливость (1.5.20) можно показать, используя перестановочность операций суммирования и взятия математического ожидания. Действительно, представляя $\mathbf{K}$ в виде

$$\mathbf{K} = \mathbf{E}\left[\left(\boldsymbol{\xi} - \mathbf{m}_{\xi}\right)^{T}\left(\boldsymbol{\xi} - \mathbf{m}_{\xi}\right)\right],$$

где $\mathbf{m}_{\xi}$ — вектор математических ожиданий, можно записать

$$\mathbf{v}\mathbf{K}\mathbf{v}^{T} = \mathbf{v}\left(\mathbf{E}\left[\left(\boldsymbol{\xi} - \mathbf{m}_{\xi}\right)^{T}\left(\boldsymbol{\xi} - \mathbf{m}_{\xi}\right)\right]\right)\mathbf{v}^{T} =$$

$$= \mathbf{E}\left[\mathbf{v}\left(\boldsymbol{\xi} - \mathbf{m}_{\xi}\right)^{T}\left(\boldsymbol{\xi} - \mathbf{m}_{\xi}\right)\mathbf{v}^{T}\right] = \mathbf{E}\left[\left(\mathbf{v}\left(\boldsymbol{\xi} - \mathbf{m}_{\xi}\right)^{T}\right)^{2}\right] \geq 0.$$

Ортогонализация симметричной и знакоопределённой матрицы означает существование *матрицы ортогонального преобразования* $\mathbf{Q}$, такой, что матричное произведение $\mathbf{Q}\mathbf{K}\mathbf{Q}^{T}$ даёт в итоге диагональную матрицу $\boldsymbol{\Lambda}$, у которой ненулевыми являются только элементы, стоящие на главной диагонали:

$$\mathbf{Q}\mathbf{K}\mathbf{Q}^{T} = \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}.$$

При этом произведение $\mathbf{Q}\mathbf{Q}^{T}$ есть единичная матрица $\mathbf{U}$ n-го порядка, а для определителей матриц $\mathbf{Q}$ и $\boldsymbol{\Lambda}$ справедливы соотношения

$$\left|\det \mathbf{Q}\right| = 1,$$

$$\det \mathbf{K} = \det \boldsymbol{\Lambda} = \lambda_1 \lambda_2 \cdots \lambda_n.$$

Нахождение ортогональной матрицы сводится к решению задачи о собственных числах и собственных векторах: *найти ортонормальный набор векторов* $\mathbf{q}_1, \dots, \mathbf{q}_n$ *и набор чисел* $\lambda_1, \dots, \lambda_n$, *удовлетворяющих системе уравнений*

$$\mathbf{q}_k \mathbf{K} = \lambda_k \mathbf{q}_k, \; k = 1, \dots, n, \qquad (1.5.21)$$

и тогда матрица $\mathbf{Q}$ — это матрица, строками которой являются собственные векторы матрицы $\mathbf{K}$.

Легко видеть, что числа $\lambda_1, \dots, \lambda_n$ (их обычно называют *спектром матрицы* $\mathbf{K}$) являются решением уравнения

$$\det\left(\mathbf{K} - \lambda \mathbf{U}\right) = 0, \qquad (1.5.22)$$

после чего собственные векторы находятся из системы

$$\left(\mathbf{K}-\lambda_k\mathbf{U}\right)\mathbf{q}^T=0,\qquad k=1,\ \ldots,\ n. \tag{1.5.23}$$

При этом, поскольку определитель такой системы есть нуль (ранг её матрицы равен $n-1$), находятся не сами компоненты вектора $\mathbf{q}$, а связь между ними ($n-1$ компонентов выражаются через какой-либо n-й компонент), и для их однозначного нахождения необходимо использовать условие единичной нормы:

$$\|\mathbf{q}\|=\sqrt{\mathbf{q}\mathbf{q}^T}=\sqrt{\sum_{k=1}^{n}q_k^2}=1. \tag{1.5.24}$$

Например, пусть $\boldsymbol{\xi}=(\xi_1,\ \xi_2)$ — двумерный нормальный вектор, задаваемый корреляционной матрицей

$$\mathbf{K}=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}, \tag{1.5.25}$$

определитель которой равен

$$\det\mathbf{K}=\sigma_1^2\sigma_2^2-\rho^2\sigma_1^2\sigma_2^2=\sigma_1^2\sigma_2^2\left(1-\rho^2\right),$$

так что обратная матрица $\mathbf{K}^{-1}$ имеет следующий вид:

$$\mathbf{K}^{-1}=\frac{1}{\sigma_1^2\sigma_2^2\left(1-\rho^2\right)}\begin{pmatrix}\sigma_2^2 & -\rho\sigma_1\sigma_2\\ -\rho\sigma_1\sigma_2 & \sigma_1^2\end{pmatrix}=$$

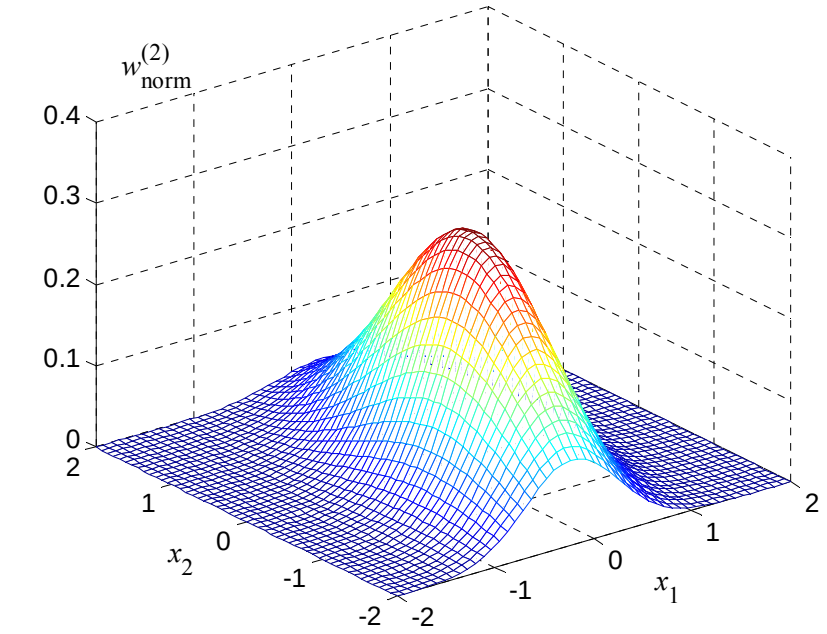
$$=\frac{1}{\left(1-\rho^2\right)}\begin{pmatrix}\dfrac{1}{\sigma_1^2} & -\dfrac{\rho}{\sigma_1\sigma_2}\\ -\dfrac{\rho}{\sigma_1\sigma_2} & \dfrac{1}{\sigma_2^2}\end{pmatrix}.$$

Таким образом, выражение для плотности вероятности двумерного гауссовского вектора с математическими ожиданиями m_1, m_2 и дисперсиями σ_1^2, σ_2^2 в явном виде есть

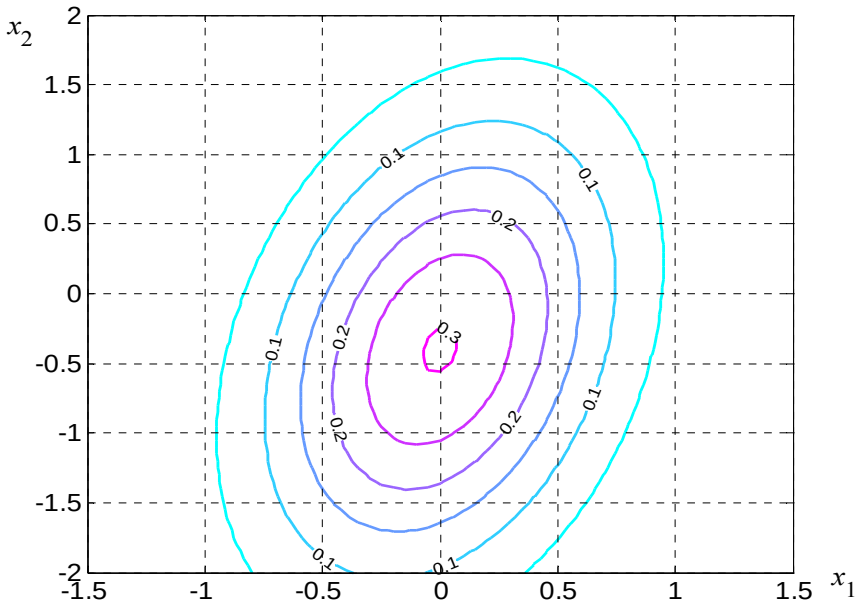
$$w_{\text{norm}}^{(2)}(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_1\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\times\right.$$

$$\left.\times\frac{(x_1-m_1)^2}{\sigma_1^2}-\frac{2\rho(x_1-m_1)(x_2-m_2)}{\sigma_1\sigma_2}+\frac{(x_2-m_2)^2}{\sigma_2^2}\right\}.$$

В качестве иллюстрации на рис. 1.2 показаны трёхмерный вид и линии уровней для двумерного гауссовского распределения с параметрами $m_1 = 0,0$; $m_2 = -0,4$; $\sigma_1^2 = 0,5$; $\sigma_2^2 = 1,1$; $\rho = 0,3$.



а)



б)

Рис. 1.2. Трёхмерный вид (*а*) и линии уровней (*б*) двумерного гауссовского распределения

Решение (1.5.22) для корреляционной матрицы (1.5.25) даёт следующие значения собственных чисел:

$$\lambda_{1,2} = \frac{1}{2}\left(\sigma_1^2 + \sigma_2^2 \pm \sqrt{\left(\sigma_1^2 + \sigma_2^2\right)^2 - 4\sigma_1^2\sigma_2^2\left(1-\rho^2\right)} \right), \qquad (1.5.26)$$

для которых системы уравнений (1.5.23), определяющие собственные векторы q_1 и q_2, имеют вид

$$\frac{1}{2}\left(\sigma_1^2 - \sigma_2^2 \mp \sqrt{\left(\sigma_1^2 + \sigma_2^2\right)^2 - 4\sigma_1^2\sigma_2^2\left(1-\rho^2\right)} \right) q_1 + \rho\sigma_1\sigma_2 q_2 = 0,$$

$$\rho\sigma_1\sigma_2 q_1 + \frac{1}{2}\left(\sigma_1^2 - \sigma_2^2 \mp \sqrt{\left(\sigma_1^2 + \sigma_2^2\right)^2 - 4\sigma_1^2\sigma_2^2\left(1-\rho^2\right)} \right) q_2 = 0,$$

откуда следует связь между q_1 и q_2:

$$q_2 = \frac{q_1}{2\rho\sigma_1\sigma_2}\left(\sigma_2^2 - \sigma_1^2 \mp \sqrt{\left(\sigma_1^2 + \sigma_2^2\right)^2 - 4\sigma_1^2\sigma_2^2\left(1-\rho^2\right)} \right).$$

Используя условие единичной нормы (1.5.24), получаем, что

$$q_1 = \frac{\sqrt{2}\rho\sigma_1\sigma_2}{\sqrt{2\rho^2\sigma_1^2\sigma_2^2 + \sigma_1^4 + \sigma_2^4 - 2\sigma_1^2\sigma_2^2\left(1-\rho^2\right) \pm \left(\sigma_2^2 - \sigma_1^2\right)\sqrt{\left(\sigma_1^2 + \sigma_2^2\right)^2 - 4\sigma_1^2\sigma_2^2\left(1-\rho^2\right)}}}, \qquad (1.5.27a)$$

$$q_2 = \frac{\left(\sigma_2^2 - \sigma_1^2 + \sqrt{\left(\sigma_1^2 + \sigma_2^2\right)^2 - 4\sigma_1^2\sigma_2^2\left(1-\rho^2\right)} \right)/\sqrt{2}}{\sqrt{2\rho^2\sigma_1^2\sigma_2^2 + \sigma_1^4 + \sigma_2^4 - 2\sigma_1^2\sigma_2^2\left(1-\rho^2\right) \pm \left(\sigma_2^2 - \sigma_1^2\right)\sqrt{\left(\sigma_1^2 + \sigma_2^2\right)^2 - 4\sigma_1^2\sigma_2^2\left(1-\rho^2\right)}}}. \qquad (1.5.27б)$$

В случае $\sigma_1 = \sigma_2 = \sigma$ формулы (1.5.26) и (1.5.27) упрощаются:

$$\lambda_{1,2} = \sigma^2(1 \pm \rho),$$

$$\mathbf{q}_1 = \left(1/\sqrt{2} \quad 1/\sqrt{2}\right),$$

$$\mathbf{q}_2 = \left(1/\sqrt{2} \quad -1/\sqrt{2}\right).$$

Итак, ортогональное преобразование $\mathbf{Q}$ переводит случайный вектор $\boldsymbol{\xi}$ в случайный вектор $\boldsymbol{\eta}$, так что $\boldsymbol{\eta} = \boldsymbol{\xi}\mathbf{Q}$. При этом математические ожидания $\mathbf{m}_\xi$ и $\mathbf{m}_\eta$ векторов $\boldsymbol{\xi}$ и $\boldsymbol{\eta}$ связаны соотношением

$$\mathbf{m}_\eta = \mathbf{m}_\xi\mathbf{Q},$$

длина (норма) вектора не меняется, ибо

$$\boldsymbol{\eta}\boldsymbol{\eta}^T = \boldsymbol{\xi}\mathbf{Q}\left(\boldsymbol{\xi}\mathbf{Q}\right)^T = \boldsymbol{\xi}\mathbf{Q}\mathbf{Q}^T\boldsymbol{\xi}^T = \boldsymbol{\xi}\boldsymbol{\xi}^T,$$

а корреляционная матрица результирующего вектора

$$\mathbf{K}_\eta = \mathbf{E}\left[\left(\boldsymbol{\eta}-\mathbf{m}_\eta\right)^T\left(\boldsymbol{\eta}-\mathbf{m}_\eta\right)\right] = \mathbf{E}\left[\left(\boldsymbol{\xi}\mathbf{Q}-\mathbf{m}_{\boldsymbol{\xi}}\mathbf{Q}\right)^T\left(\boldsymbol{\xi}\mathbf{Q}-\mathbf{m}_{\boldsymbol{\xi}}\mathbf{Q}\right)\right] =$$

$$= \mathbf{E}\left[\mathbf{Q}^T\left(\boldsymbol{\xi}-\mathbf{m}_{\boldsymbol{\xi}}\right)^T\left(\boldsymbol{\xi}-\mathbf{m}_{\boldsymbol{\xi}}\right)\mathbf{Q}\right] = \mathbf{Q}^T\mathbf{E}\left[\left(\boldsymbol{\xi}-\mathbf{m}_{\boldsymbol{\xi}}\right)^T\left(\boldsymbol{\xi}-\mathbf{m}_{\boldsymbol{\xi}}\right)\right]\mathbf{Q} =$$

$$= \mathbf{Q}^T\mathbf{K}_\xi\mathbf{Q} = \mathbf{Q}\mathbf{K}_{\boldsymbol{\xi}}\mathbf{Q}^T = \boldsymbol{\Lambda}$$

является диагональной, и элементы, стоящие на главной диагонали — это собственные числа корреляционной матрицы $\mathbf{K}_\xi$ исходного вектора. Иными словами, компоненты $\eta_1, \ldots, \eta_n$ некоррелированы, а их дисперсии $\sigma^2_{\eta k}$ равны

$$\sigma^2_{\eta k} = \mathbf{E}\left[\left(\eta_k - m_{\eta k}\right)^2\right] = \lambda_k, \qquad k = 1, \ldots, n.$$

Например, для рассмотренного случая двумерного нормального вектора $\boldsymbol{\xi} = (\xi_1, \xi_2)$ с одинаковыми дисперсиями $\sigma_1 = \sigma_2 = \sigma$ в ортогонализированном векторе $\boldsymbol{\eta} = (\eta_1, \eta_2)$ компоненты имеют математические ожидания

$$m_{\eta 1} = \frac{m_{\xi 1} + m_{\xi 2}}{\sqrt{2}}, \quad m_{\eta 2} = \frac{m_{\xi 1} - m_{\xi 2}}{\sqrt{2}}$$

и дисперсии

$$\sigma^2_{\eta 1} = \sigma^2(1+\rho), \qquad \sigma^2_{\eta 2} = \sigma^2(1-\rho).$$

Теперь, после того, как получен вектор $\boldsymbol{\eta}$ некоррелированных величин, вычисление количества информации, содержащееся в векторе $\boldsymbol{\xi}$, является несложной задачей. Поскольку ортогональное преобразование не изменяет информационные характеристики, дифференциальные энтропии вектор $\boldsymbol{\xi}$ и $\boldsymbol{\eta}$ совпадают, отсюда:

$$h(\boldsymbol{\xi}) = h(\boldsymbol{\eta}) = \frac{n}{2}\log 2\pi e + \frac{1}{2}\log\det\mathbf{K}_\eta = \frac{n}{2}\log 2\pi e + \frac{1}{2}\log\det\boldsymbol{\Lambda} =$$

$$= \frac{n}{2}\log 2\pi e + \frac{1}{2}\sum_{k=1}^{n}\log\lambda_k = \frac{1}{2}\sum_{k=1}^{n}\log 2\pi e\lambda_k. \qquad (1.5.28)$$

В частности, для двумерного нормального вектора

$$h(\boldsymbol{\xi}) = \log 2\pi e + \log\sigma^2\sqrt{1-\rho^2}. \qquad (1.5.29)$$

1.6. ВЗАИМНАЯ ИНФОРМАЦИЯ. КОЛИЧЕСТВО ВЗАИМНОЙ ИНФОРМАЦИИ ДЛЯ ДИСКРЕТНЫХ СООБЩЕНИЙ

В предыдущих разделах были исследованы характеристики, касающиеся содержания информации в случайных величинах, получаемой при непосредственном наблюдении за их значениями. Однако на практике не меньший, а подчас и больший интерес представляет собой попытка получения информации о каком-то источнике посредством наблюдения за состояниями другого источника, каким-то образом связанного с первым. В этой связи оказывается целесообразным ввести понятие *взаимной информации*, характеризующей степень информационной зависимости двух и более источников.

Рассмотрим, прежде всего, случай, когда сообщения выбираются источниками независимым образом. Пусть $X = \{x_i\}$ ($i = 1, 2, \ldots, l_x$) и $Y = \{y_j\}$ ($j = 1, 2, \ldots, l_y$) — два дискретных источника с независимыми элементами, для которых заданы вероятности $p(x_i)$ и $p(y_j)$ выбора элементов, а также условные вероятности $p(x_i / y_j)$ или $p(y_j / x_i)$. Тогда вероятность совместного события (x_i, y_j) равна

$$p\left(x_i, y_j\right) = p\left(x_i / y_j\right) p\left(y_j\right) = p\left(y_j / x_i\right) p\left(x_i\right),$$

и могут быть введены следующие характеристики, описывающие информационные соотношения между двумя источниками.

Взаимная информация $I(x_i; y_j)$ двух символов относительно друг друга, т. е. количество информации о символе x_i, доставляемое принятым символом y_j:

$$I\left(x_i; y_j\right) = \log \frac{p\left(x_i / y_j\right)}{p\left(x_i\right)} = \log \frac{p\left(x_i, y_j\right)}{p\left(x_i\right) p\left(y_j\right)}. \qquad (1.6.1)$$

Заметим, что это определение симметрично: $I(x_i; y_j) = I(y_j; x_i)$. Взаимная информация является обобщением собственной информации, содержащейся в событии x_i, поскольку

$$I\left(x_i; x_i\right) = \log \frac{1}{p\left(x_i\right)} = -\log p\left(x_i\right) = I\left(x_i\right). \qquad (1.6.2)$$

Также следует заметить, что в отличие от $I(x_i)$ взаимная информация может принимать и отрицательные значения. Так, если $p(x_i / y_j) < p(x_i)$, то

$I(x_i; y_j) < 0$; это означает, что наблюдение за сообщением y_j делает появление сообщения x_i менее вероятным, чем оно было до наблюдения.

Величина $I(x_i; y_j)$, определяемая согласно (1.6.1), обладает следующими свойствами:

- если сообщения x_i и y_j независимы, т. е. $p(x_i, y_j) = p(x_i)p(y_j)$, то $I(x_i; y_j) = 0$, и сообщение y_j не несёт никакой информации о x_i;
- если сообщение y_j с полной определённостью указывает на сообщение x_i, т. е. $p(x_i / y_j) = 1$, то y_j несёт в себе полную информацию о x_i, т. е. $I(x_i; y_j) = I(x_i)$.
- для произвольных источников X и Y справедливы соотношения

$$I\left(x_i; y_j\right) \le I\left(x_i\right); \quad I\left(x_i; y_j\right) \le I\left(y_j\right).$$

Из выражения (1.6.1) следует, что

$$I\left(x_i; y_j\right) = \log p\left(x_i / y_j\right) - \log p\left(x_i\right) = I\left(x_i\right) - I\left(x_i / y_j\right), \qquad (1.6.3)$$

где $I(x_i / y_j)$ имеет смысл условной собственной информации символа x_i при известном значении y_j. Иными словами, $I(x_i; y_j)$ определяет разность между количеством информации, заключённой в сообщении x_i, и тем количеством информации, которое осталось не переданным после приёма y_j.

На основании понятия взаимной информации можно ввести ряд величин, характеризующих информационные соотношения между двумя источниками.

Количество информации, доставляемое принятым символом y_j, относительно источника X:

$$I\left(X; y_j\right) = \sum_{i=1}^{l_x} I\left(x_i; y_j\right) p\left(x_i / y_j\right) = \sum_{i=1}^{l_x} \log \frac{p\left(x_i / y_j\right)}{p\left(x_i\right)} p\left(x_i / y_j\right). \qquad (1.6.4a)$$

Количество информации, доставляемое принятым символом x_i, относительно источника Y:

$$I\left(Y; x_i\right) = \sum_{j=1}^{l_y} I\left(y_j; x_i\right) p\left(y_j / x_i\right) = \sum_{j=1}^{l_y} \log \frac{p\left(y_j / x_i\right)}{p\left(y_j\right)} p\left(y_j / x_i\right). \qquad (1.6.4б)$$

Средняя взаимная информация между источниками X и Y:

$$I(X;Y) = \mathbf{E}\left[I\left(x_i, y_j\right)\right] = \sum_{i=1}^{l_x} \sum_{j=1}^{l_y} p\left(x_i, y_j\right) \log \frac{p\left(x_i, y_j\right)}{p\left(x_i\right) p\left(y_j\right)}. \qquad (1.6.5)$$

Используя неравенство $\ln x \le x - 1$, из соотношения (1.6.5) нетрудно показать, что взаимная информация $I(X;Y)$ является неотрицательной величиной, т. е. $I(X;Y) \ge 0$.

По аналогии с тем, как это делалось при рассмотрении характеристик источников с памятью, введём характеристики *взаимной энтропии* двух источников. Для совместного события (x_i, y_j) энтропия $H(XY)$ множества совместных событий равна

$$H\left(XY\right) = -\sum_{i=1}^{l_x}\sum_{j=1}^{l_y} p\left(x_i, y_j\right)\log p\left(x_i, y_j\right). \qquad (1.6.6)$$

Условной энтропией источника X при заданном источнике Y назовём величину

$$H\left(X/Y\right) = -\sum_{j=1}^{l_y} p(y_j)\sum_{i=1}^{l_x} p(x_i/y_j)\log p(x_i/y_j) =$$

$$= -\sum_{j=1}^{l_y}\sum_{i=1}^{l_x} p(x_i, y_j)\log p(x_i/y_j). \qquad (1.6.7a)$$

Аналогично, условной энтропией источника Y при заданном источнике X назовём величину

$$H\left(Y/X\right) = -\sum_{i=1}^{l_x} p(x_i)\sum_{j=1}^{l_y} p(y_j/x_i)\log p(y_j/x_i) =$$

$$= -\sum_{j=1}^{l_y}\sum_{i=1}^{l_x} p(x_i, y_j)\log p(y_j/x_i). \qquad (1.6.7б)$$

Легко видеть, что

$$H\left(XY\right) = H\left(X\right) + H\left(Y/X\right) = H\left(Y\right) + H\left(X/Y\right). \qquad (1.6.8)$$

При этом, когда источники X и Y независимы, $H(Y/X) = H(Y)$, $H(X/Y) = H(X)$ и, следовательно,

$$H(XY) = H(X) + H(Y).$$

Выразим среднюю взаимную информацию $I(X; Y)$ через характеристики взаимной энтропии. Поскольку

$$\sum_{j=1}^{l_y} p\left(x_i, y_j\right) = p\left(x_i\right), \qquad \sum_{i=1}^{l_x} p\left(x_i, y_j\right) = p\left(y_j\right),$$

очевидно,

$$I(X;Y) = H(X) - H(X/Y) = H(Y) - H(Y/X) =$$

$$= H(X) + H(Y) - H(XY). \qquad (1.6.9)$$

Кроме того, из неотрицательности взаимной информации $I(X;Y)$ справедливо неравенство

$$H\left(X/Y\right) \le H\left(X\right). \qquad (1.6.10)$$

Все рассмотренное может быть естественным образом обобщено на случай взаимной информации для числа источников, большего двух. В этом случае соотношения (1.6.1)–(1.6.10) остаются в силе с той лишь разницей, что появляется дополнительная условная вероятностная зависимость. Например, в случае трёх источников X, Y, и Z соотношение (1.6.10) трансформируется следующим образом:

$$H\left(X/YZ\right) \le H\left(X/Z\right). \qquad (1.6.10a)$$

Рассмотрим следующий пример, являющийся моделью передачи сигналов по двоичному каналу связи с помехами (рис. 1.2). Подробнее этот вопрос будет рассмотрен далее.

Источник выдаёт независимые сообщения x_1 и x_2 с априорными вероятностями $p(x_1) = 3/4$ и $p(x_2) = 1/4$. При этом из-за наличия помех сообщение x_1 с вероятностью $p(y_1 / x_1) = 5/8$ переходит в сообщение y_1 (“прямой”, т. е. правильный переход) и с вероятностью $p(y_2 / x_1) = 3/8$ — в сообщение y_2 (“непрямой”, т. е. ошибочный переход), а сообщение x_2 с вероятностью $p(y_2 / x_2) = 7/8$ переходит в сообщение y_2 (правильный переход) и с вероятностью $p(y_1 / x_2) = 1/8$ — в сообщение y_1 (ошибочный переход).

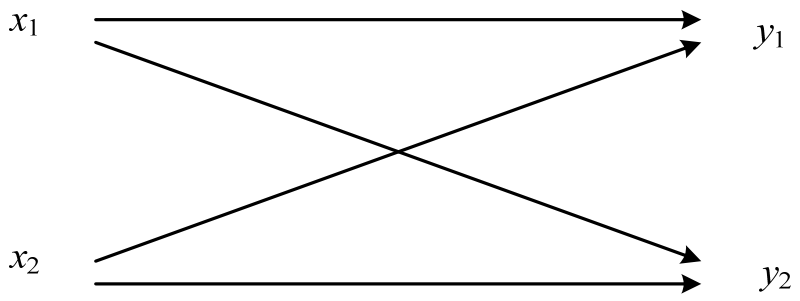


Рис. 1.3. Модель двоичного канала связи

Найдём, прежде всего, неизвестные вероятностные характеристики взаимного ансамбля сигналов. Легко видеть, что безусловные вероятности $p(y_1)$ и $p(y_2)$ сигналов y_1 и y_2 одинаковы:

$$p(y_1) = p(y_1/x_1)\,p(x_1) + p(y_1/x_2)\,p(x_2) = \frac{1}{2};$$

$$p(y_2) = p(y_2/x_1)\,p(x_1) + p(y_2/x_2)\,p(x_2) = \frac{1}{2}.$$

Также можно найти совместные вероятности $p(x_i, y_j)$, которые оказываются равными

$$p(y_1, x_1) = p(y_1/x_1)\,p(x_1) = \frac{15}{32}; \qquad p(y_1, x_2) = p(y_1/x_2)\,p(x_2) = \frac{1}{32};$$

$$p(y_2, x_1) = p(y_2/x_1)\,p(x_1) = \frac{9}{32}; \qquad p(y_2, x_2) = p(y_2/x_2)\,p(x_2) = \frac{7}{32}.$$

Наконец, условные вероятности $p(x_i / y_j)$ имеют значения

$$p(x_1/y_1) = \frac{p(y_1/x_1)\,p(x_1)}{p(y_1)} = \frac{15}{16}; \qquad p(x_1/y_2) = \frac{p(y_2/x_1)\,p(x_1)}{p(y_2)} = \frac{9}{16};$$

$$p(x_2/y_1) = \frac{p(y_1/x_2)\,p(x_2)}{p(y_1)} = \frac{1}{16}; \qquad p(x_2/y_2) = \frac{p(y_2/x_2)\,p(x_2)}{p(y_2)} = \frac{7}{16}.$$

На основании найденных вероятностей вычислим информационные характеристики. Собственная информация о сообщениях x_1, x_2, y_1 и y_2 равна соответственно 0,42, 2,00, 1,00 и 1,00 бит. Тогда энтропия источника составляет $H(X) = 0{,}815$.

Взаимная информация имеет следующие значения:

$$I(x_1, y_1) = 0{,}32 \text{ бит}; \quad I(x_1, y_2) = -0{,}42 \text{ бит};$$

$$I(x_2, y_1) = -2{,}00 \text{ бит}; \quad I(x_2, y_2) = 0{,}81 \text{ бит},$$

откуда можно найти величину средней взаимной информации, которая оказывается равной $I(X,Y) = 0{,}146$.

Одно из важнейших свойств взаимной информации состоит в том, что она не может увеличиться при функциональном преобразовании. Докажем это свойство.

Пусть некоторое преобразование φ отображает элементы исходного источника X на элементы другого множества Z, так что каждый элемент $z = \varphi(z) \in Z$ является образом некоторого (возможно, даже, не одного — в случае неоднозначного преобразования) элемента $x \in X$. Полагая также, что задан ансамбль $XY = \{p(x, y),\ x \in X,\ y \in Y\}$ посредством совместных вероятностей (из которых могут быть вычислены все одномерные и условные вероятности), можно считать, что изначально определена средняя взаимная информация $I(X; Y)$ между алфавитами X и Y.

Преобразование $Z = \varphi(X)$ определяет ансамбль $ZY = \{p(z, y),\ z \in Z,\ y \in Y\}$, для которого

$$p(z,y) = \sum_{x:\varphi(x)=z} p(x,y), \qquad (1.6.11)$$

так что определена средняя взаимная информация $I(Z;Y)$ между алфавитами Z и Y. Докажем справедливость следующего утверждения.

Для любого отображения $Z = \varphi(X)$ *ансамбля* X *в ансамбль* Z *справедливо соотношение*

$$I(X;Y) \geq I(Z;Y), \qquad (1.6.12)$$

Причём равенство имеет место для взаимно-однозначного отображения, когда каждому элементу $z \in Z$ *соответствует единственный элемент* $x \in X$.

Для доказательства рассмотрим совместный ансамбль XYZ, состоящий из всевозможных упорядоченных троек (x, y, z) и совместных вероятностей $p(x, y, z)$ появления таких троек. Исходя из заданных совместных "тройных" вероятностей, можно вычислить и совместные "двойные"

$$p(x,y) = \sum_{Z} p(x,y,z);$$

$$p(y,z) = \sum_{X} p(x,y,z);$$

$$p(x,z) = \sum_{Y} p(x,y,z),$$

и "одиночные"

$$p(x) = \sum_{Y} p(x,y) = \sum_{Z} p(x,z);$$

$$p(y) = \sum_{X} p(x,y) = \sum_{Z} p(y,z);$$

$$p(z) = \sum_{X} p(x,z) = \sum_{Y} p(y,z).$$

вероятности, а также все необходимые условные вероятности, например,

$$p(y / x) = \frac{p(x,y)}{p(x)}.$$

На основании (1.6.3) вычисляются условная взаимная информация $I(y; z / x)$ между элементами y и z при фиксированном x:

$$I(y;z/x)=I(y/x)-I(y/x,z)=\log\frac{p(y/x,z)}{p(y/x)},$$

а также взаимная информация[1] $I((x,y);z)$ между парой (x,y) и одиночным элементом z:

$$I((x,y);z)=I(x,y)-I((x,y)/z)=\log\frac{p((x,y)/z)}{p(x,y)}.$$

Представим $I((x,y);z)$ в виде суммы, в которой фигурирует взаимная информация между x и z, а также взаимная информация между y и z:

$$I((x,y);z)\equiv I((y,x);z)=I(y)+I(x/y)-I(y/z)-I(x/y,z)=$$
$$=I(y;z)+I(x;z/y).$$

Аналогично\ы образом можно получить симметричное выражение

$$I((x,y);z)=I(x;z)+I(y;z/x).$$

Теперь, усредняя полученные выражения по всем трём алфавитам, получим соотношения

$$I(XY;Z)=I(Y;Z)+I(X;Z/Y)=I(X;Z)+I(Y;Z/X), \qquad (1.6.13)$$

а также симметричные им, например,

$$I(XZ;Y)=I(Z;X)+I(X;Z/Y)=I(X;Y)+I(Z;Y/X), \qquad (1.6.13a)$$

отражающие *свойство аддитивности взаимной информации*.

Пусть теперь алфавит Z представляет собой множество значений при отображении $z=\varphi(x)$. Так как при выбранном элементе $x\in X$ однозначно определён элемент $z\in Z$, и он не зависит от любого $y\in Y$, т. е.

$$p\left(z/x,y\right)=p\left(z/x\right)=\begin{cases}1, & z=\varphi(x);\\ 0, & z\neq\varphi(x);\end{cases}$$

распределение вероятностей на тройках удовлетворяет условию

$$p\left(x,y,z\right)=p\left(z/x,y\right)p\left(x,y\right)=p\left(z/x\right)p\left(x,y\right),$$

из которого видно, что условная взаимная информация между любыми элементами ансамблей Z и Y равна нулю:

$$I(z;y/x)=\log\frac{p(z/x,y)}{p(z/x)}=\log\frac{p(z/x)}{p(z/x)}=0,$$

а, следовательно, равна нулю и средняя условная взаимная информация

[1] Необходимо различать обозначения $I(x,y)$ — собственная информация, связанная с появлением пары (x,y) и $I(x;y)$ — взаимная информация между элементами x и y.

$I(Z; Y/X)$. Отсюда, используя свойство аддитивности взаимной информации (1.6.13), можно записать

$$I(XZ;Y) = I(X;Y) + I(Y;Z/X) = I(X;Y). \qquad (1.6.14)$$

С другой стороны, в силу неотрицательности средней взаимной информации справедливо

$$I(XZ;Y) = I(Z;Y) + I(X;Y/Z) \geq I(Z;Y). \qquad (1.6.15)$$

Сопоставляя (1.6.14) и (1.6.15), получаем требуемое соотношение (1.6.12). При этом равенство имеет место в том случае, когда условная взаимная информация $I(X;Y/Z) = 0$, а это, как нетрудно видеть, выполняется, если для всех (x, y, z), принадлежащих множеству XYZ, справедливо соотношение

$$p(x/y,z) = p(x/z),$$

означающее, что при выбранном $z \in Z$ сообщение $x \in X$ вероятностно не зависит от $y \in Y$. Данное условие всегда выполняется, если x и z однозначно определяют друг друга, т. е. при взаимно-однозначном преобразовании.

В доказанном утверждении неявно предполагалось, что преобразование $z = \varphi(x)$ является детерминированным. На самом деле оно также справедливо и для случайных преобразований, если для них выполняется условие (1.6.13).

Свойству невозрастания информации при преобразованиях можно дать следующую физическую интерпретацию. Пусть наблюдаются сообщения источника X, по которым желательно получить информацию о некотором объекте, состояния которого образуют множество Y. Например, X — множество возможных сигналов на выходе канала связи, а Y — множество посылаемых в канал сообщений. Тогда можно утверждать, что никакая обработка наблюдений не может увеличить средней информации об интересующем объекте. В лучшем случае информация сохранится, но как правило — потеряется.

В дальнейшем для краткости количество средней взаимной информации часто будем называть просто взаимной информацией.

1.7. КОЛИЧЕСТВО ВЗАИМНОЙ ИНФОРМАЦИИ ДЛЯ НЕПРЕРЫВНЫХ СООБЩЕНИЙ

Пусть ξ и η — непрерывные, вообще говоря, зависимые случайные величины, задаваемые одномерными плотностями вероятностей $w_\xi(x)$ и

$w_\eta(x)$ соответственно, при этом статистическая связь между ξ и η характеризуется двумерной плотностью $w^{(2)}_{\xi\eta}(x,y)$. По аналогии с тем, как это делалось в одномерном случае, разобьём соответствующие множества значений X и Y на дискретное число интервалов шириной Δx и Δy. Тогда для произвольных интервалов $\left[k\Delta x;(k+1)\Delta x\right]$ и $\left[i\Delta y;(i+1)\Delta y\right]$ могут быть введены следующие вероятности:

$$p^{(k)}_{\xi} = P_{\xi}\left(x^*_k\right) = P\left\{\xi \in \left[k\Delta x;(k+1)\Delta x\right]\right\} = \int\limits_{k\Delta x}^{(k+1)\Delta x} w_{\xi}(x)dx \approx w(x^*_k)\Delta x;$$

$$p^{(i)}_{\eta} = P_{\eta}\left(y^*_i\right) = P\left\{\eta \in \left[i\Delta y;(i+1)\Delta x\right]\right\} = \int\limits_{i\Delta y}^{(i+1)\Delta y} w_{\eta}(y)dy \approx w(y^*_i)\Delta y;$$

$$p^{(k,i)}_{\xi\eta} = P_{\xi\eta}\left(x^*_k, y^*_i\right) = P\left\{\xi \in \left[k\Delta x;(k+1)\Delta x\right], \eta \in \left[i\Delta y;(i+1)\Delta y\right]\right\} =$$

$$= \int\limits_{k\Delta x}^{(k+1)\Delta x}\int\limits_{i\Delta y}^{(i+1)\Delta y} w^{(2)}_{\xi\eta}(x,y)dxdy \approx w^{(2)}_{\xi\eta}\left(x^*_k, y^*_i\right)\Delta x\Delta y.$$

Рассмотрим для полученных дискретных сообщений характеристики взаимной информации. Условная энтропия $\tilde{H}\left[\xi/\eta\right]$, т. е. среднее количество информации, получаемое о дискретных выборках ξ по наблюдению за дискретными выборками η равна

$$\tilde{H}\left[\xi/\eta\right] = -\sum_{(k)}\sum_{(i)} P_{\xi\eta}\left(x^*_i, y^*_j\right)\log\frac{P_{\xi\eta}\left(x^*_i, y^*_j\right)}{P_{\eta}\left(y^*_j\right)} =$$

$$= -\sum_{(k)}\sum_{(i)} w^{(2)}_{\xi\eta}\left(x^*_k, y^*_i\right)\Delta x\Delta y\log\frac{w^{(2)}_{\xi\eta}\left(x^*_k, y^*_i\right)\Delta x\Delta y}{w_{\eta}\left(y^*_i\right)\Delta y} =$$

$$= -\sum_{(k)}\sum_{(i)} w^{(2)}_{\xi\eta}\left(x^*_k, y^*_i\right)\log\frac{w^{(2)}_{\xi\eta}\left(x^*_k, y^*_i\right)}{w_{\eta}\left(y^*_i\right)}\Delta x\Delta y - \sum_{(k)}\sum_{(i)} w^{(2)}_{\xi\eta}\left(x^*_k, y^*_i\right)\log(\Delta x)\Delta x\Delta y.$$

Аналогичным образом условная энтропия $\tilde{H}\left[\eta/\xi\right]$, т. е. среднее количество информации, получаемое о дискретных выборках η по наблюдению за дискретными выборками ξ равна

$$\tilde{H}\left[\eta/\xi\right]=-\sum_{(k)}\sum_{(i)}w^{(2)}_{\xi\eta}\left(x^*_k,y^*_i\right)\log\frac{w^{(2)}_{\xi\eta}\left(x^*_k,y^*_i\right)}{w_{\xi}\left(x^*_{k,}\right)}\Delta x\Delta y-$$

$$-\sum_{(k)}\sum_{(i)}w^{(2)}_{\xi\eta}\left(x^*_k,y^*_i\right)\log(\Delta y)\Delta x\Delta y\,.$$

Устремим Δx и Δy к нулю. Тогда "истинная" условная энтропия $H\left[\xi/\eta\right]$ непрерывного сообщения имеет вид:

$$H\left[\xi/\eta\right]=\lim_{\substack{\Delta x\to 0\\ \Delta y\to 0}}\tilde{H}\left[\xi/\eta\right]=-\lim_{\substack{\Delta x\to 0\\ \Delta y\to 0}}\sum_{(k)}\sum_{(i)}w^{(2)}_{\xi\eta}\left(x^*_{k,}\,y^*_i\right)\log\frac{w^{(2)}_{\xi\eta}\left(x^*_{k,}\,y^*_i\right)}{w_{\eta}\left(y^*_i\right)}\Delta x\Delta y-$$

$$-\lim_{\substack{\Delta x\to 0\\ \Delta y\to 0}}\sum_{(k)}\sum_{(i)}w^{(2)}_{\xi\eta}\left(x^*_{k,}\,y^*_i\right)\log(\Delta x)\Delta x\Delta y=h\left[\xi(t)/\eta(t)\right]-\lim_{\Delta x\to 0}\log\Delta x\,,$$

где, по аналогии с (1.4.2), величина

$$h(\xi/\eta)=-\iint_{(X,Y)}w^{(2)}_{\xi\eta}\left(x,y\right)\log\frac{w^{(2)}_{\xi\eta}\left(x,y\right)}{w_{\eta}\left(y\right)}dxdy \qquad (1.7.1a)$$

называется *дифференциальной условной энтропией* случайной величины ξ по наблюдению случайной величины η. Таким же образом, вводя дифференциальную условную энтропию случайной величины η по наблюдению случайной величины ξ

$$h(\eta/\xi)=-\iint_{(X,Y)}w^{(2)}_{\xi\eta}\left(x,y\right)\log\frac{w^{(2)}_{\xi\eta}\left(x,y\right)}{w_{\xi}\left(x\right)}dxdy\,, \qquad (1.7.1б)$$

можно записать условную энтропию $H(\eta/\xi)$:

$$H(\eta/\xi)=h\left[\eta(t)/\xi(t)\right]-\lim_{\Delta y\to 0}\log\Delta y\,.$$

Тогда взаимная информация $I(\xi;\eta)$ между величинами ξ и η равна

$$I(\xi;\eta)=\lim_{\substack{\Delta x\to 0\\ \Delta y\to 0}}\left[\tilde{H}\left(\xi\right)-\tilde{H}\left(\xi/\eta\right)\right]=$$

$$=h\left(\xi\right)-\lim_{\Delta x\to 0}\log\Delta x-h\left(\xi/\eta\right)+\lim_{\Delta x\to 0}\log\Delta x=h\left(\xi\right)-h\left(\xi/\eta\right),$$

т. е. разности между собственной и условной дифференциальными энтропиями какого-либо одного процесса. Учитывая симметричность $I(\xi;\eta)$, окончательно имеем

$$I(\xi;\eta)=h(\xi)-h(\xi/\eta)=h(\eta)-h(\eta/\xi) \qquad (1.7.2)$$

Взаимная информация по-прежнему является относительной величиной, но изучение одного непрерывного источника посредством наблюдения за реализациями другого, как это ни парадоксально, снимает необходимость введения дополнительного эталонного источника (сравните с разд. 1.4), относительного которого сравнивается неопределённость получения реализации непрерывного сообщения. Иными словами, невозможность идеального, абсолютно точного наблюдения отдельно за каждым непрерывным источником трансформируется в практически состоятельную характеристику измерения количества взаимной информации между такими источниками.

Запишем соотношение (1.7.2) в виде

$$I(\xi,\eta) = -\int_{-\infty}^{\infty} w_\xi(x)\log w_\xi(x)\,dx + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w_{\xi\eta}^{(2)}(x,y)\log\frac{w_{\xi\eta}^{(2)}(x,y)}{w_\eta(y)}\,dxdy =$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w_{\xi\eta}^{(2)}(x,y)\log\frac{w_{\xi\eta}^{(2)}(x,y)}{w_\eta(y)w_\xi(x)}\,dxdy\,, \qquad (1.7.3)$$

из которого видно, что если $w_{\xi\eta}^{(2)}(x,y) = w_\eta(y)w_\xi(x)$, т. е. величины ξ и η независимы, то взаимная информация между ними равна нулю.

Обобщим введённые понятия взаимной информации на многомерные непрерывные величины. В разд. 1.5 было введена дифференциальная энтропия $h(\boldsymbol{\xi})$ для многомерного распределения $w_{\boldsymbol{\xi}}(\mathbf{x})$, описывающего последовательность (вектор) случайных величин $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$. По аналогии с этим, введём условные дифференциальные энтропии $h(\boldsymbol{\xi}/\boldsymbol{\eta})$ и $h(\boldsymbol{\eta}/\boldsymbol{\xi})$ для совокупности из $2n$ случайных величин $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ и $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$, задаваемых $2n$-мерным совместным распределением $w_{\boldsymbol{\xi\eta}}(\mathbf{x}, \mathbf{y})$:

$$h\left[\boldsymbol{\xi}/\boldsymbol{\eta}\right] = -\iint_{X^n Y^n} w_{\boldsymbol{\xi\eta}}(\mathbf{x},\mathbf{y})\log\frac{w_{\boldsymbol{\xi\eta}}(\mathbf{x},\mathbf{y})}{w_{\boldsymbol{\eta}}(\mathbf{y})}\,d\mathbf{x}d\mathbf{y}\,. \qquad (1.7.4a)$$

$$h\left[\boldsymbol{\eta}/\boldsymbol{\xi}\right] = -\iint_{X^n Y^n} w_{\boldsymbol{\xi\eta}}(\mathbf{x},\mathbf{y})\log\frac{w_{\boldsymbol{\xi\eta}}(\mathbf{x},\mathbf{y})}{w_{\boldsymbol{\xi}}(\mathbf{x})}\,d\mathbf{x}d\mathbf{y}\,. \qquad (1.7.4б)$$

Тогда среднюю взаимную информацию $I(\boldsymbol{\xi};\, \boldsymbol{\eta})$ между последовательностями $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ и $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ определим как

$$I(\boldsymbol{\xi},\boldsymbol{\eta}) = h(\boldsymbol{\xi})] - h(\boldsymbol{\xi}/\boldsymbol{\eta}) = h(\boldsymbol{\eta}) - h(\boldsymbol{\eta}/\boldsymbol{\xi})\,. \qquad (1.7.5)$$

Данная характеристика является непосредственным обобщением (1.7.3) и обладает всеми свойствами, присущими взаимной информации.

Рассмотрим некоторые примеры вычисления $I(\xi; \eta)$ и $I(\boldsymbol{\xi}; \boldsymbol{\eta})$ для непрерывных случайных величин.

Пример 1. Найдём взаимную информацию $I(\xi, \eta)$ между случайными величинами ξ и $\eta = \xi + \zeta$, где независимые величины ξ и ζ имеют нормальное распределение с нулевыми средними и дисперсиями σ_ξ^2 и σ_ζ^2 соответственно. В такой постановке задача является моделью канала, в котором действует шум с "интенсивностью" ζ, а ξ и η являются значениями сигналов на входе и выходе канала соответственно.

Поскольку величины ξ и ζ нормальные, η — также нормальная с нулевым средним и дисперсией $\sigma_\eta^2 = \sigma_\xi^2 + \sigma_\zeta^2$. Тогда при вычислении взаимной информации по формуле

$$I(\eta, \xi) = h(\eta) - h(\eta / \xi)$$

безусловная энтропия равна

$$h(\eta) = \log\sqrt{2\pi e\sigma_\eta^2} = \log\sqrt{2\pi e(\sigma_\xi^2 + \sigma_\zeta^2)}\,,$$

а поскольку при фиксированном значении $\xi = x$ распределение случайной величины η совпадает с распределением случайной величины ζ (с точностью до математического ожидания, не влияющего на дифференциальную энтропию), условная энтропия имеет вид

$$h(\eta / \xi) = -\int_{-\infty}^{\infty} w_\xi(x)dx \int_{-\infty}^{\infty} w_\eta(y / \xi)\log w_\eta(y / \xi)dy =$$

$$= -\int_{-\infty}^{\infty} w_\xi(x)dx \int_{-\infty}^{\infty} w_\zeta(z)\log w_\zeta(z)dz = \log\sqrt{2\pi e\sigma_\zeta^2}\,.$$

Тогда

$$I(\eta; \xi) = -\log\sqrt{2\pi e\sigma_\zeta^2} + \log\sqrt{2\pi e(\sigma_\xi^2 + \sigma_\zeta^2)} = \frac{1}{2}\log\left(1 + \frac{\sigma_\xi^2}{\sigma_\zeta^2}\right). \qquad (1.7.6)$$

Рассмотрим этот же пример с точки зрения статистической обработки. Пусть получена статистическая выборка (табл. 1.4) о значениях пар (x, z) случайных величин ξ и ζ (табл. 1.4).

Таблица 1.4

Статистическая выборка значений пар (x, z)

(x, z)		(x, z)		(x, z)		(x, z)	
0,830	0,018	–0,199	–0,327	0,725	0,749	1,219	2,394
0,361	–0,772	0,459	1,544	–0,080	–0,037	0,326	1,872
0,764	3,344	–0,129	2,083	–0,181	–2,193	0,607	0,739
–0,192	–0,188	0,745	–1,261	–0,105	3,969	0,194	2,228
–0,107	0,611	0,200	–3,087	–0,236	0,678	0,525	3,873
0,153	–2,716	0,695	–1,137	0,183	–0,892	0,301	–0,456
–0,803	0,006	0,947	–0,686	–0,572	–2,287	–0,281	–0,191
0,528	2,330	–0,354	0,779	–0,935	–0,466	–0,347	–2,551
–0,013	0,003	–0,916	–1,552	0,306	0,423	0,483	–2,005
–1,312	–2,726	–0,641	1,677	0,254	–0,630	–0,697	0,276
0,900	–3,674	–0,736	1,515	0,181	0,514	0,101	–0,097
0,305	3,685	0,003	0,044	0,403	1,150	0,854	–1,285
0,322	2,396	0,075	2,281	–1,010	–5,750	–0,130	–0,686
0,234	–1,690	–0,170	1,766	0,415	–0,332	0,215	3,219
1,344	–0,203	0,263	1,882	0,217	1,240	–0,029	–1,391
0,168	–0,924	0,151	–2,524	–0,350	–0,604	–0,201	0,669
0,207	1,827	–0,506	-0,544	–0,028	1,038	–0,813	0,273
0,044	3,043	0,015	–2,368	0,080	0,692	0,755	1,672
0,066	1,468	–0,385	0,798	0,191	1,357	1,216	1,012
0,046	–0,285	0,183	–1,568	0,273	–1,455	–0,595	–1,058
–0,220	–1,620	–0,267	2,193	0,374	–0,528	0,073	3,955
–0,796	1,703	–0,202	2,288	0,497	0,602	–1,006	0,797
–0,118	0,496	–0,262	–3,128	0,349	–0,649	–0,391	–1,955
–0,003	4,418	–0,088	–1,356	–0,114	0,889	–0,918	–0,374
–0,778	–0,289	–0,314	–0,711	–1,437	3,261	0,389	0,685

Анализ представленных в табл. 1.4 данных показывает, что каждую из величин ξ или ς можно с высокой степенью доверия считать нормально распределённой. Например, при использовании критерия согласия χ^2 это решение можно принять с уровнем значимости 0,08. В качестве иллюстра-

ции на рис. 1.4 представлены эмпирические $F_1^*(x)$, $F_2^*(x)$ и теоретические $F_1(x)$, $F_2(x)$ функции распределения для ξ и ς соответственно.

Точечные оценки корреляционных параметров дают следующую корреляционную матрицу:

$$\mathbf{K} = \begin{pmatrix} 0,242 & 0,057 \\ 0,057 & 3,977 \end{pmatrix},$$

на главной диагонали которой располагаются точечные оценки дисперсии, и в рассматриваемом примере они равны $s_1^2 = 0,242$ и $s_2^2 = 3,977$. На основе матрицы $\mathbf{K}$ можно построить нормированную корреляционную матрицу (матрицу корреляционных коэффициентов),

$$\mathbf{K}' = \begin{pmatrix} 1,000 & 0,059 \\ 0,059 & 1,000 \end{pmatrix},$$

где на главной диагонали всегда стоят единицы, а на побочной — значения коэффициента корреляции ρ, лежащие в диапазоне [–1; 1].

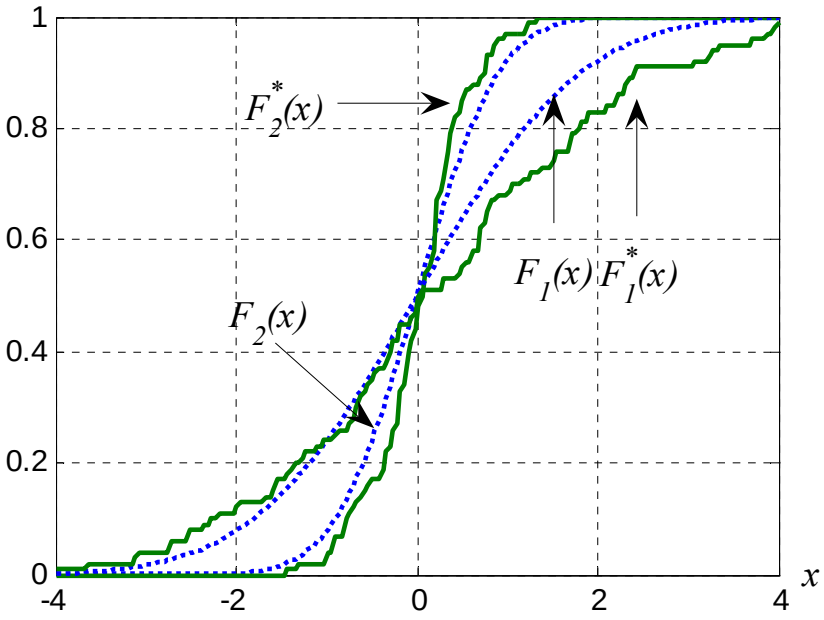


Рис. 1.4. Эмпирические $F^*(x)$ и теоретические $F(x)$ функции распределения

Близость |ρ| к единице показывает сильную коррелированность величин ξ и ς; напротив, близкое к нулю значение ρ (что и имеет место в данном случае) означает их практическую независимость. Это видно и на рис. 1.5, на котором нанесено облако точек (x, z).

Теперь, отображая условия задачи на представленные в табл. 1.4 данные и подставляя соответствующие значения в (1.7.6), получаем значение взаимной информации $I(\xi;\ \varsigma) = 0{,}043$, что является достаточно малой величиной. Это и не удивительно, поскольку при трактовке этой задачи как модели передачи полезной информации по каналу с шумом отношение дисперсий $\sigma^2_\xi / \sigma^2_\zeta$ имеет смысл отношения сигнал-шум, и в данном примере "интенсивность" шума существенно превосходит "интенсивность" полезного сигнала.

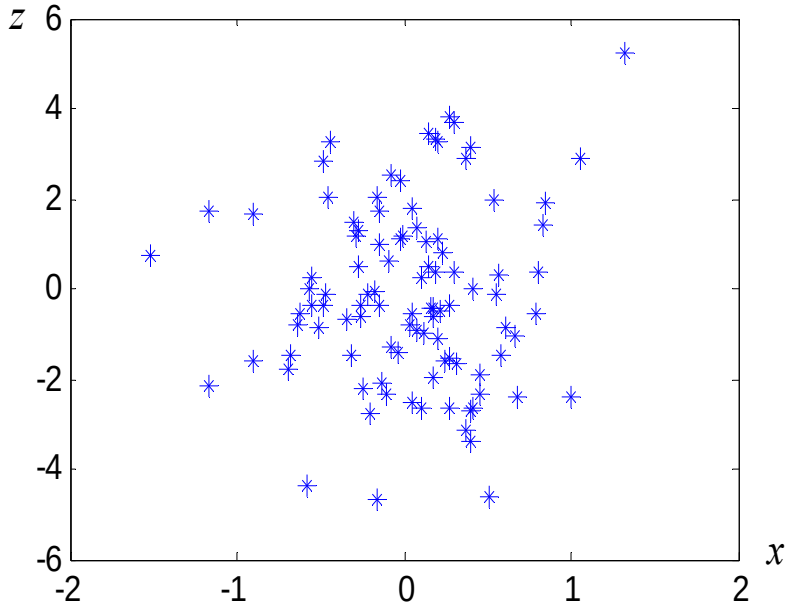


Рис. 1.5. Облако точек (x, z)

Пример 2. Пусть ξ и η — зависимые нормальные величины, характеризуемые математическими ожиданиями $\mathbf{E}[\xi] = a_\xi$, $\mathbf{E}[\eta] = a_\eta$, дисперсиями $\mathbf{D}[\xi] = \sigma_\xi^2$, $\mathbf{D}[\eta] = \sigma_\eta^2$ и коэффициентом корреляции ρ:

$$w^{(2)}_{\xi\eta}(x,y) = \frac{1}{2\pi\sigma_\xi\sigma_\eta\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\times\right.$$

$$\left.\times\left[\frac{\left(x-a_\xi\right)^2}{\sigma_\xi^2} - \frac{2\rho\left(x-a_\xi\right)\left(y-a_\eta\right)}{\sigma_\xi\sigma_\eta} + \frac{\left(y-a_\eta\right)^2}{\sigma_\eta^2}\right]\right\}. \qquad (1.7.7)$$

Тогда, переходя к натуральным логарифмам, имеем:

$$I(\xi;\eta)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w_{\xi\eta}^{(2)}(x,y)\log\frac{1}{\sqrt{1-\rho^2}}dxdy+$$

$$+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w_{\xi\eta}^{(2)}(x,y)\left\{-\frac{1}{2}\left[\frac{\left(x-a_\xi\right)^2}{\left(1-\rho^2\right)\sigma_\xi^2}-\frac{2\rho\left(x-a_\xi\right)\left(y-a_\eta\right)}{\left(1-\rho^2\right)\sigma_\xi\sigma_\eta}+\right.\right.$$

$$\left.\left.+\frac{\left(y-a_\eta\right)^2}{\left(1-\rho^2\right)\sigma_\eta^2}-\frac{\left(x-a_\xi\right)^2}{\left(1-\rho^2\right)\sigma_\xi^2}-\frac{\left(y-a_\eta\right)^2}{\sigma_\eta^2}\right]\right\}=-\frac{1}{2}\ln\left(1-\rho^2\right)-$$

$$-\frac{1}{2}\left[\frac{1}{1-\rho^2}-\frac{2\rho^2}{1-\rho^2}+\frac{1}{1-\rho^2}-1-1\right]=-\frac{1}{2}\ln\left(1-\rho^2\right).$$

Итак, взаимная информация между двумя нормальными зависимыми источниками ξ и η, характеризующимися коэффициентом корреляции ρ, определяется выражением

$$I(\xi;\eta)=-\frac{1}{2}\ln\left(1-\rho^2\right)\text{ натуральных единиц.}\qquad(1.7.8)$$

Обратимся теперь к многомерным случайным величинам и рассмотрим пример, являющийся многомерным аналогом уже рассмотренного выше примера.

Пример 3. Пусть заданы два независимых вектора нормальных величин $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ и $\boldsymbol{\varsigma} = (\varsigma_1, \ldots, \varsigma_n)$ с нулевыми математическими ожиданиями и корреляционными матрицами $\mathbf{K}_\xi$ и $\mathbf{K}_\varsigma$ соответственно. Требуется найти взаимную информацию между векторами $\boldsymbol{\xi}$ и $\boldsymbol{\eta} = \boldsymbol{\xi} + \boldsymbol{\varsigma}$.

При вычислении $I(\boldsymbol{\xi}; \boldsymbol{\eta})$ по формулам (1.7.5) безусловная энтропия $h[\boldsymbol{\eta}]$ согласно (1.5.19) составляет

$$h(\boldsymbol{\eta})=\frac{n}{2}\log 2\pi e+\frac{1}{2}\log\det\mathbf{K}_\eta,$$

где элементы $\left(b_\eta\right)_{kj}$ корреляционной матрицы $\mathbf{K}_\eta$ в силу независимости векторов $\boldsymbol{\xi}$ и $\boldsymbol{\varsigma}$ равны

$$\left(b_\eta\right)_{kj}=\int_{X^nZ^n}\left(\left(x^{(k)}+z^{(k)}\right)-\mathbf{E}\left[\xi_k+\varsigma_k\right]\right)\left(\left(x^{(j)}+z^{(j)}\right)-\mathbf{E}\left[\xi_j+\varsigma_j\right]\right)\times$$

$$\times w_{\boldsymbol{\xi}\boldsymbol{\varsigma}}(\mathbf{x},\mathbf{z})d\mathbf{x}d\mathbf{z}=\int_{X^n}\left(\left(x^{(k)}\right)-\mathbf{E}\left[\xi_k\right]\right)\left(\left(x^{(j)}\right)-\mathbf{E}\left[\xi_j\right]\right)w_\xi(\mathbf{x})d\mathbf{x}+$$

$$+\int_{Z^n}\Big(\big(z^{(k)}\big)-\mathbf{E}\big[\varsigma_k\big]\Big)\Big(\big(z^{(j)}\big)-\mathbf{E}\big[\varsigma_j\big]\Big)w_{\varsigma}(\mathbf{z})d\mathbf{z}=\big(b_{\xi}\big)_{kj}+\big(b_{\varsigma}\big)_{kj},$$

так что саму корреляционную матрицу $\mathbf{K}_{\eta}$ можно записать как

$$\mathbf{K}_{\eta}=\mathbf{K}_{\xi}+\mathbf{K}_{\varsigma}.$$

Условная энтропия $h(\boldsymbol{\eta}\,/\,\boldsymbol{\xi})$ аналогично одномерному случаю определяется плотностью вероятности $w_{\boldsymbol{\eta}}(\mathbf{y}\,/\,\boldsymbol{\xi})=w_{\boldsymbol{\zeta}}(\mathbf{y}-\mathbf{x})$, поэтому средняя взаимная информация принимает вид

$$I(\boldsymbol{\xi};\boldsymbol{\eta})=\frac{1}{2}\log\det\big(\mathbf{K}_{\xi}+\mathbf{K}_{\zeta}\big)-\frac{1}{2}\log\det\big(\mathbf{K}_{\zeta}\big)=\frac{1}{2}\log\frac{\det\big(\mathbf{K}_{\xi}+\mathbf{K}_{\zeta}\big)}{\det\big(\mathbf{K}_{\zeta}\big)}. \quad (1.7.9)$$

При этом по умолчанию предполагается, что $\det\mathbf{K}_{\zeta}\neq 0$.

В частности, когда корреляционные матрицы имеют диагональный вид:

$$\mathbf{K}_{\xi}=\begin{pmatrix}\sigma_{\xi 1}^2 & 0 & \cdots & 0\\ 0 & \sigma_{\xi 2}^2 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ 0 & 0 & \cdots & \sigma_{\xi n}^2\end{pmatrix},\quad \mathbf{K}_{\zeta}=\begin{pmatrix}\sigma_{\zeta 1}^2 & 0 & \cdots & 0\\ 0 & \sigma_{\zeta 2}^2 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ 0 & 0 & \cdots & \sigma_{\zeta n}^2\end{pmatrix},$$

взаимная информация равна

$$I(\boldsymbol{\xi};\boldsymbol{\eta})=\frac{1}{2}\log\frac{\prod_{k=1}^{n}\big(\sigma_{\xi k}^2+\sigma_{\zeta k}^2\big)}{\prod_{k=1}^{n}\sigma_{\zeta k}^2}=\frac{1}{2}\sum_{k=1}^{n}\log\left(1+\frac{\sigma_{\xi k}^2}{\sigma_{\zeta k}^2}\right), \quad (1.7.10)$$

что является обобщением соотношения (1.7.6) для взаимной информации между одномерными случайными величинами.

В заключение данного раздела получим выражения, определяющие количество взаимной информации между дискретным и непрерывным сообщениями.

Пример 4. Пусть дискретная случайная величина ξ принимает набор значений $\{x_i\}(i=1,\ \ldots,\ l)$ с вероятностями $\{p_i\}$, а непрерывная случайная величина ζ описывается распределениями $w_{\zeta}(z)$ и $w_{\zeta}(z\,/\,x)$, где z — значения из континуального множества Z. Разобьём это множество значений на дискретное число интервалов шириной Δz:

$$Z = \bigcup_{(k)} \left[k\Delta z; (k+1)\Delta z \right].$$

Количество взаимной информации $\tilde{I}(\xi;\zeta)$ между ξ и квантованными, с точностью до Δz, значениями ς согласно (1.6.5), равно

$$\tilde{I}(\xi;\zeta) = \sum_{i=1}^{l} \sum_{(k)} p\left(x_i, z_k^*\right) \log \frac{p\left(x_i, z_k^*\right)}{p\left(x_i\right) p\left(z_k^*\right)},$$

где $z_k^* \in \left[k\Delta z; (k+1)\Delta z \right]$, $k = \ldots, -1, 0, 1, \ldots$.

Заменим приближённо $p(z_k^*) \approx w_\zeta(z_k^*)\Delta z$ и

$$p(x_i, z_k^*) = p(z_k^* / x_i) p\left(x_i\right) \approx p(x_i) w_\zeta(z_k^* / x_i)\Delta z,$$

тогда

$$\tilde{I}(\xi;\zeta) = \sum_{i=1}^{l} p(x_i) \sum_{(k)} w_\zeta(z_k^* / x_i) \log \frac{w_\zeta(z_k^* / x_i)}{w_\zeta(z_k^*)} \Delta z,$$

и при стремлении $\Delta z \to 0$ получаем выражение для взаимной информации между дискретной и непрерывной величинами:

$$I(\xi;\zeta) = \sum_{i=1}^{l} p(x_i) \int_{(Z)} w_\zeta\left(z / x_i\right) \log \frac{w_\zeta\left(z / x_i\right)}{w_\zeta(z)} dz. \qquad (1.7.10)$$

Пусть дискретная случайная величина ξ равновероятно принимает значения в интервале $[-A_0; A_0]$, т. е.

$$x_i = A_0(m - 2i + 1) / (1 - m),\ I = 1, \ldots, m.$$

Например:

при $m = 2$ значения $x_1 = -A_0$; $x_2 = A_0$;

при $m = 4$ значения $x_1 = -A_0$; $x_2 = -A_0/3$; $x_1 = A_0/3$; $x_2 = A_0$; и т. д.

Далее, пусть случайная величина ζ представляет собой сумму дискретной случайной величины ξ и не зависящей от неё гауссовской случайной величины ν, имеющей нулевое среднее и дисперсию σ^2:

$$\zeta = \xi + \nu.$$

Условное распределение $w_\zeta(z / x_i)$ — это гауссовское распределение с дисперсией σ^2 и средним x_i, а безусловное распределение $w_\zeta(z)$ получается путём усреднения $w_\zeta(z / x_i)$ по всем дискретным значениям:

$$w_\zeta(z) = \sum_{i=1}^{l} w_\zeta\left(z / x_i\right) p(x_i) = \frac{1}{l\sigma^2\sqrt{2\pi}} \sum_{i=1}^{l} \exp\left(-\frac{(z - x_i)^2}{2\sigma^2} \right).$$

Тогда для взаимной информации между ξ и ζ можно записать следующее выражение:

$$I(\xi,\zeta)=\frac{b}{\sqrt{2\pi}}\sum_{i=1}^{l}p(x_i)\int_{-\infty}^{\infty}\exp\left(-b^2(x-d_i)^2/2\right)\times$$

$$\times\log\frac{\exp\left(-b^2(x-d_i)^2/2\right)}{\sum_{k=1}^{l}p(x_k)\exp\left(-b^2(x-d_k)^2/2\right)}dx, \qquad (1.7.11)$$

где введён параметр $b = A_0/\sigma^2$.

С инженерной точки зрения рассматриваемый пример можно трактовать как передачу каналу полезной информации (значений случайной величины ξ) по дискретно-непрерывному каналу, в котором действует гауссовская помеха (случайная величина ν). В этой случае параметр b имеет смысл “отношения сигнал-шум” (отношение амплитудного значения полезного сигнала к среднеквадратическому значению помехи).

На рис. 1.6 показаны зависимости количества взаимной информации от параметра b при различных количествах m входных значений.

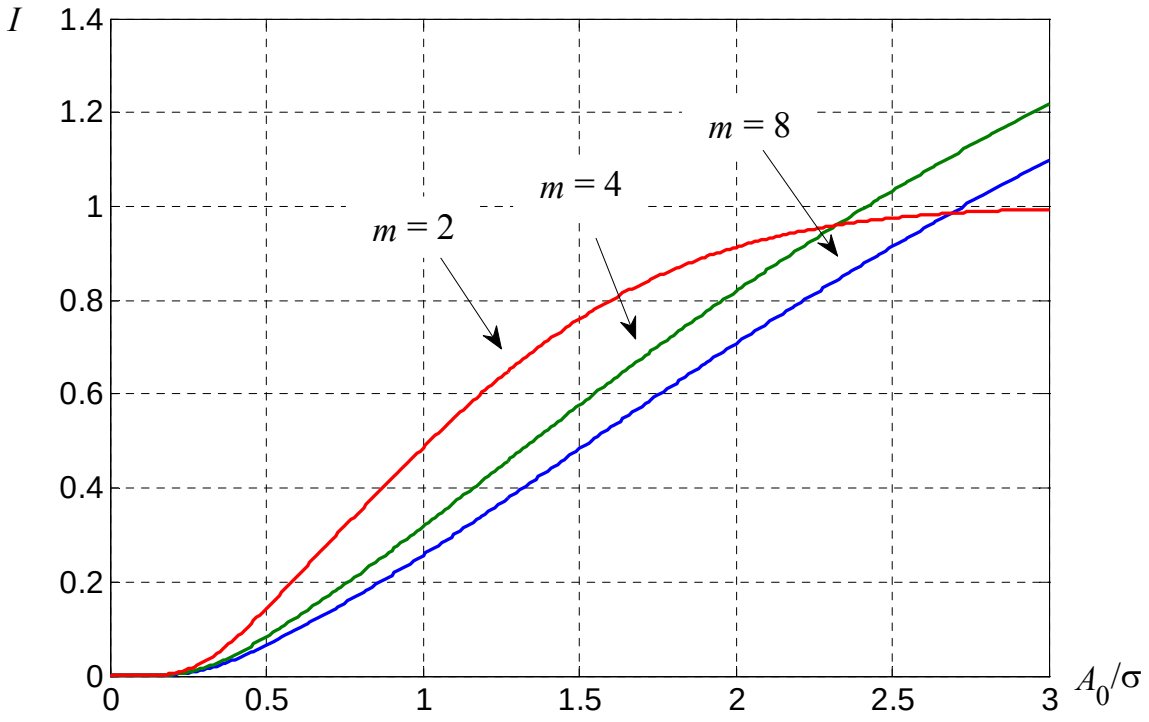


Рис. 1.6. Количество взаимной информации между равновероятной дискретной и гауссовской случайными величинами

При малых значениях b (значения помехи превалируют над значениями полезной информации) взаимная информация близка к нулю, и по мере роста b значения взаимной информации возрастают, выходя на насыщение: при $m = 2$ значения I стремятся к 1; при $m = 4$ значения I стремятся к 2.

Отметим, что при больших значениях m (реально при $m \geq 10$) рассматриваемый пример может служить приближённым решением задачи вычисления взаимной информации между двумя непрерывными случайными величинами: равномерной на интервале $[-A_0; A_0]$ и гауссовской. Точное решение такое задачи затруднительно невозможности аналитического вычисления интегралов в (1.7.3) для рассматриваемых распределений.

1.8. ВЫПУКЛОСТЬ ВЗАИМНОЙ ИНФОРМАЦИИ

В данном разделе будет исследовано одно из важнейших свойств взаимной информации — свойство выпуклости по отношению к вероятностным ансамблям. Наличие такого свойства у функции позволяет относительно легко ее максимизировать в соответствующих областях, и такого рода задачи составляют целое теоретико-информационное направление.

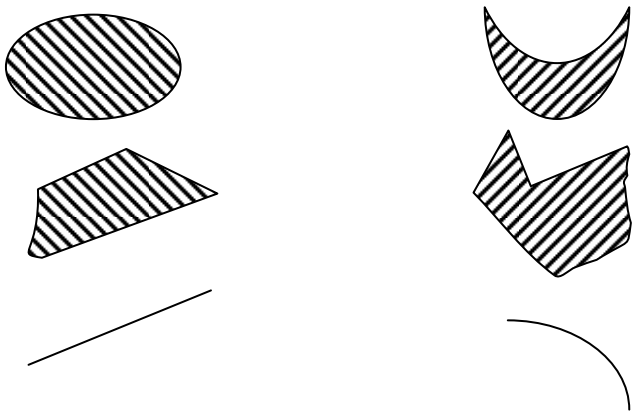
Прежде всего, напомним определения выпуклой области и выпуклой функции [5].

Пусть $\mathbf{v} = [v_1\ v_2\ \ldots\ v_n]$ — n-мерный вектор с вещественными компонентами, определёнными в некоторой области B векторного пространства. Тогда *область B является выпуклой, если для любых принадлежащих ей векторов $\mathbf{v}_1$ и $\mathbf{v}_2$ и для любого числа $\alpha \in [0;1]$ результирующий вектор $\mathbf{v} = \alpha\mathbf{v}_1 + (1 - \alpha)\mathbf{v}_2$ также принадлежит области B*.

Геометрически это означает, что когда α изменяется от 0 до 1, вектор $\mathbf{v}$ перемещается по отрезку прямой линии между $\mathbf{v}_1$ и $\mathbf{v}_2$. Иными словами, область является выпуклой, если для каждой принадлежащей ей пары точек отрезок прямой линии между этими точками также принадлежит области. На рис. 1.7 приведены примеры плоских (двумерных) выпуклых и невыпуклых областей.

Поскольку основным приложением к выпуклым областям в рассматриваемых вопросах являются вероятностные векторы, т. е. векторы, компоненты которых неотрицательны и в сумме равны единице, конкретизи-

руем понятие выпуклой области по отношению именно к вероятностным векторам.



Выпуклые области Невыпуклые области

Рис. 1.7. Примеры выпуклых и невыпуклых областей

Пусть B — конечномерное пространство вероятностных векторов. Тогда, если v_{1k} и v_{2k} ($k = 1, \ldots, n$) — компоненты вероятностных векторов $\mathbf{v}_1$ и $\mathbf{v}_2$, то, очевидно, для всех компонентов $v_k = \alpha v_{1k} + (1-\alpha)v_{2k}$ результирующего вектора выполняются свойства $v_k \geq 0$ и

$$\sum_{k=1}^{n} v_k = \alpha \sum_{k=1}^{n} v_{1k} + (1-\alpha)\sum_{k=1}^{n} v_{2k} = \alpha + 1 - \alpha = 1.$$

Таким образом, пространство вероятностных векторов всегда является выпуклой областью.

В другом случае — при рассмотрении непрерывных ансамблей под областью B понимается множество функций, являющихся плотностями вероятностей для соответствующих случайных величин. Нетрудно показать, что в этом случае область B также всегда является выпуклой областью линейного пространства.

Далее определим понятие выпуклости функции.

Функция $f(\mathbf{v})$ называется выпуклой в выпуклой области B, если для любых неотрицательных чисел $b_1, \ldots, b_s$ ($s = 1, 2, \ldots$) сумма которых равна единице, и любых векторов $\mathbf{v}_1, \ldots, \mathbf{v}_s$ из B выполняется неравенство

$$\sum_{i=1}^{s} b_i f\left(\mathbf{v}_i\right) \le f\left(\sum_{i=1}^{s} b_i \mathbf{v}_i\right). \qquad (1.8.1)$$

В противном случае функция y = f(x) называется вогнутой[1].

Неравенству (1.8.1) можно дать простую геометрическую трактовку. Пусть $s = 2$, тогда $b_1 = b$, $b_2 = 1 - b_1 = 1 - b$, и соотношение

$$bf\left(\mathbf{v}_1\right) + (1-b) f\left(\mathbf{v}_2\right) \le f\left[b\mathbf{v}_1 + (1-b)\mathbf{v}_2\right]$$

применительно к зависимости $f(b)$ означает (рис. 1.8), что хорда, соединяющая две точки плоской кривой, лежит под этой кривой или совпадает с ней, если сама кривая является линией.

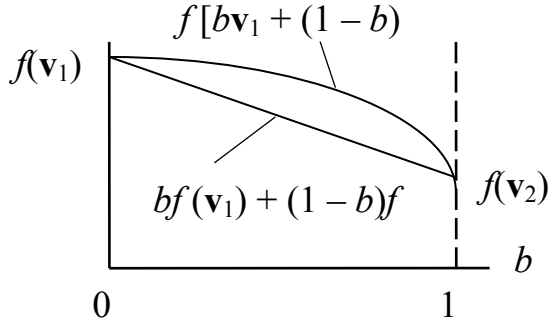


Рис. 1.8. Геометрическая иллюстрация выпуклости

Аналогичное геометрическое свойство справедливо и в многомерном случае: часть плоскости (гиперплоскости при $s \ge 4$) на которой лежат точки $f(\mathbf{v}_1)$, …, $f(\mathbf{v}_s)$, находится ниже поверхности $f\left(\sum_{i=1}^{s} b_i \mathbf{v}_i\right)$ (или совпадает с ней).

Дифференцируемая выпуклая функция обладает следующими важными свойствами:

- для любых векторов $\mathbf{v}_1$ и $\mathbf{v}_2$ справедливо соотношение

$$f(\mathbf{v}_2) - f(\mathbf{v}_1) \ge \nabla^T f(\mathbf{v}_1)(\mathbf{v}_2 - \mathbf{v}_1);$$

- матрица Гессе (матрица вторых смешанных производных) функции $f(\mathbf{v})$ является знакоопределённой;
- в области B функция $f(\mathbf{v})$ имеет только один экстремум.

---

[1] Иногда применяется терминология “функция, выпуклая вверх” и “функция, выпуклая вниз”.

Пусть $XY = \{(x, y), p(x, y)\}$ — дискретный ансамбль с совместным распределением вероятностей $p(x, y) = p(x / y)p(y) = p(y / x)p(x)$. Тогда средняя взаимная информация между ансамблями X и Y определяется выражением

$$I(X;Y) = \sum_X \sum_Y p\left(y / x\right) p(x) \log \frac{p\left(y / x\right)}{p(y)} =$$

$$= \sum_X \sum_Y p\left(x / y\right) p(y) \log \frac{p\left(x / y\right)}{p(x)}. \qquad (1.8.2)$$

Аналогичное выражение имеет место для непрерывного ансамбля $XY = \{(x, y), w^{(2)}(x, y)\}$, задаваемого совместной плотностью вероятностей $w^{(2)}(x, y) = w(x / y)w(y) = w(y / x)w(x)$:

$$I(X;Y) = \int_X \int_Y w^{(2)}\left(y / x\right) w(x) \log \frac{w^{(2)}\left(y / x\right)}{w(y)} dxdy =$$

$$= \int_X \int_Y w^{(2)}\left(x / y\right) w(y) \log \frac{w^{(2)}\left(x / y\right)}{w(x)} dxdy. \qquad (1.8.3)$$

При рассмотрении свойств выпуклости функции $I(X; Y)$ оказывается необходимым различать способы образования совместных распределений.

В первом случае считаются заданными условные вероятности (или распределения $w(x / y)$). Тогда взаимная информация является функцией (или функционалом) безусловного распределения вероятностей $\mathbf{p} = [p_1 \ldots p_l]$ (или безусловной плотности вероятности $w(x)$), и $I(X; Y)$ можно рассматривать как функцию (функционал), определённую на элементах выпуклой области B, образованной соответствующими вероятностными объектами.

Во втором случае считаются заданными безусловные вероятности (плотности вероятности) на одном их ансамбле, например, на X. Образуем выпуклую область B для дискретного случая совокупностью всех матриц, каждая строка которой есть вероятностный вектор, а для непрерывного — совокупностью функций двух переменных $w(y/x)$, обладающих тем свойством (свойством условной плотности вероятности), что при каждом $x \in X$ она неотрицательна, и её интеграл по множеству Y равен 1. Легко видеть, что множество условных плотностей также является выпуклой областью.

Действительно, если $w_1(y / x)$ и $w_2(y / x)$ — две условные плотности вероятности, то для любого $\alpha \in [0;1]$ результирующая функция

$$w(y/x) = \alpha w_1(y/x) + (1-\alpha) w_2(y/x)$$

неотрицательна при каждом значении $x \in X$ и

$$\int_Y w(y/x)dy = \alpha \int_Y w_1(y/x)dy + (1-\alpha)\int_Y w_2(y/x)dy = 1.$$

Таким образом, во втором случае взаимная информация является функцией условных распределений вероятностей на ансамбле *Y*, т. е. функцией от элементов матрицы с вероятностными векторами $p(y/x)$, $x \in X$, $y \in Y$ в дискретном случае или функционалом от условной плотности вероятности $w(y/x)$, $x \in X$, $y \in Y$ в непрерывном случае.

Поставим вопрос, обладает ли взаимная информация свойством выпуклости? Оказываются справедливыми следующие утверждения.

1. *Взаимная информация I(X; Y) является выпуклой функцией в области B, образованной всеми безусловными распределениями на множестве X (вероятностными векторами или плотностями вероятностей) при заданных на множестве Y условных распределениях.*

2. *Взаимная информация I(X; Y) является вогнутой функцией в области B, образованной всеми условными распределениями на множестве Y (матрицами с вероятностными векторами) при заданных на множестве X безусловных распределениях.*

Докажем подробно справедливость первого из этих утверждений для дискретного случая; метод доказательства для второго случая, а также для непрерывных ансамблей аналогичен (см. задачу 1.11).

Введём в рассмотрение дополнительно к ансамблю *XY*, где заданы условные вероятности $p(y/x)$, а вероятностный вектор $\mathbf{p} = [p_1 \ldots p_l]$ определяет функциональную зависимость, множество *Z* элементов $\{z_1, \ldots, z_s\}$, вероятностные соотношения для которых есть

$$p(x, y, z_j) = p(x/z_j)\, p(y/x)\, p(z_j), \quad j = 1, \ldots, s. \qquad (1.8.4)$$

Тогда при каждом фиксированном $z_j \in Z$ на множестве *XY* определены условные распределения

$$p(x, y/z_j) = p(y/x)\, p(x/z_j), \qquad (1.8.5)$$

которым соответствуют значения условной взаимной информации

$$I(X;Y/z_j)=\sum_X\sum_Y p\left(y,x/z_j\right)p(x)\log\frac{p\left(y/x,z_j\right)}{p(y/z_j)}. \qquad (1.8.6)$$

С другой стороны, на множестве XY определено безусловное распределение вероятностей, получаемое из (1.8.5) усреднением по всему множеству Z:

$$p(x,y)=\sum_Z p\left(x,y/z_j\right)p\left(z_j\right)=\sum_Z p\left(y/x\right)p\left(x/z_j\right)p\left(z_j\right)=p\left(y/x\right)p\left(x\right), \qquad (1.8.7)$$

где

$$p\left(x\right)=\sum_{j=1}^{s}p\left(x/z_j\right)p\left(z_j\right), \qquad (1.8.8)$$

и, таким образом, определена безусловная взаимная информация $I(X; Y)$.

Каждое из условных распределений $p(x / z_j)$ ($j = 1, \ldots, s$) представляет собой вероятностный вектор, а величина $I(X; Y/z_j)$ — значение условной взаимной информации на таком векторе. Тогда

$$\sum_{j=1}^{s}I(X;Y/z_j)p(z_j)=I(X;Y/Z)$$

представляет собой взаимную информацию ансамбля XY относительно множества Z, и она является левой частью неравенства (1.8.1). Таким образом, для доказательства утверждения необходимо доказать неравенство

$$I(X;Y/Z)\le I(X;Y)$$

в предположении, что справедливо (1.8.4).

Для доказательства неравенства используем свойство аддитивности взаимной информации (1.6.14):

$$I(XY;Z)=I(X;Y)+I(Y;Z/X)=I(Y;Z)+I(X;Z/Y).$$

Из (1.8.4) получаем, что для всех $(x,y,z)\in XYZ$ выполняется

$$\frac{p\left(y,z/x\right)}{p\left(y/x\right)p\left(z/x\right)}=\frac{p\left(x\right)}{p\left(x\right)}\frac{p\left(y,z/x\right)}{p\left(y/x\right)p\left(z/x\right)}=\frac{p\left(x,y,z\right)}{p\left(x\right)p\left(y/x\right)p\left(z/x\right)}=$$

$$=\frac{p\left(x/z\right)p\left(y/x\right)p\left(z\right)}{p\left(x\right)p\left(y/x\right)p\left(z/x\right)}=1,$$

т. е. условная взаимная информация $I(y; z / x) = 0$, а следовательно, равна нулю и средняя информация $I(Y; Z / X)$, так что

$$I(X;Y) = I(Y;Z) + I(X;Z/Y).$$

Теперь, учитывая неотрицательность средней взаимной информации $I(Y;Z)$, можно записать требуемое соотношение

$$I(X;Y) = I(Y;Z) + I(X;Z/Y) \geq I(X;Z/Y).$$

Итак, средняя взаимная информация является выпуклой функцией в области B всех безусловных распределений вероятностных векторов на множестве X при заданных условных распределениях на множестве Y.

Непрерывный случай отличается от дискретного только тем, что вероятностные распределения на тройках (x, y, z) задаются посредством соотношений смешанного типа

$$w(x/z)w(y/z)p(z),$$

когда ансамбли X и Y непрерывны, а Z — дискретный. При этом структура доказательства полностью сохраняется.

Для иллюстрации свойства выпуклости взаимной информации рассмотрим следующий простой пример.

Пусть, как и в разд.1.6, $X = \{0, 1\}$ — двоичный источник на входе канала, а $Y = \{0, 1\}$ — множество сообщений на выходе канала. Вероятности двоичных символов на входе и выходе канала равны p_1 и $1 - p_1$, p_2 и $1 - p_2$ соответственно (рис. 1.9). При этом канальные вероятности, соответствующие прямым и непрямым переходам символов 0 и 1, одинаковы — такой канал называется *двоичным симметричным каналом* (ДСК).

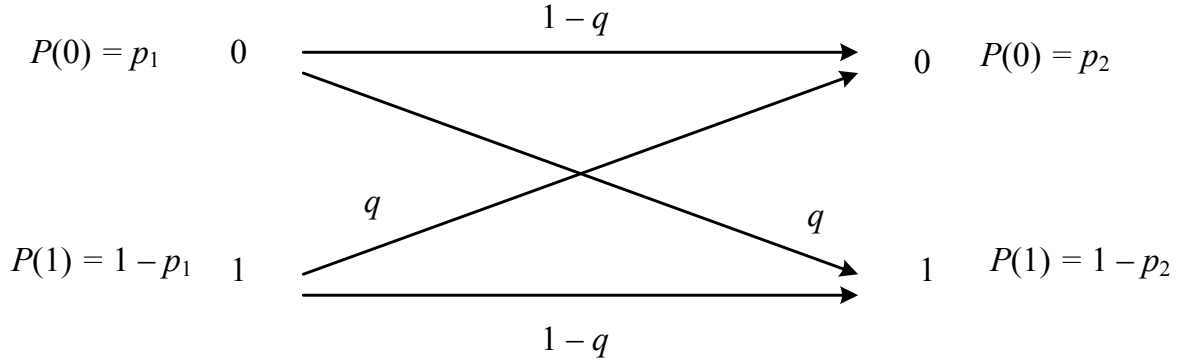


Рис. 1.9. Переходные вероятности ДСК

Нетрудно выразить вероятности выходных символов через вероятности символов на входе:

$$p_2 = (1 - p_1)q + p_1(1 - q);$$

$$1 - p_2 = p_1 q + (1 - p_1)(1 - q).$$

Таким образом, взаимную информацию $I(X; Y)$ можно рассматривать как функцию двух переменных p_1 и q. Имеем:

$$I\big(X(p_1, q); Y(p_1, q)\big) = H\big(Y\big) - H\big(Y / X\big),$$

где

$$H\big(Y\big) = -p_2 \log p_2 - (1 - p_2) \log p_2 =$$

$$= -\big[(1 - p_1)q + p_1(1 - q)\big] \log\big[(1 - p_1)q + p_1(1 - q)\big] -$$

$$-\big[p_1 q + (1 - p_1)(1 - q)\big] \log\big[p_1 q + (1 - p_1)(1 - q)\big].$$

$$H\big(Y / X\big) = -p_1\big[(1 - q)\log(1 - q) + q \log q\big] - \big(1 - p_1\big)\big[q \log q + (1 - q)\log(1 - q)\big] =$$

$$= -q \log q - (1 - q)\log(1 - q).$$

Итак, взаимная информация равна

$$I\big[p_1, q\big] = -\big[(1 - p_1)q + p_1(1 - q)\big] \log\big[(1 - p_1)q + p_1(1 - q)\big] -$$

$$-\big[p_1 q + (1 - p_1)(1 - q)\big] \log\big[p_1 q + (1 - p_1)(1 - q)\big] + q \log q + (1 - q)\log(1 - q), \quad (1.8.9)$$

В соответствие с высказанными выше утверждениями рассмотрим два случая:

- задаются канальные вероятности (условные распределения), и взаимная информации трактуется только как функция p_1;
- задаются входные вероятности (фактически — только p_1), и взаимная информации рассматривается как функция канальной (условной) вероятности q.

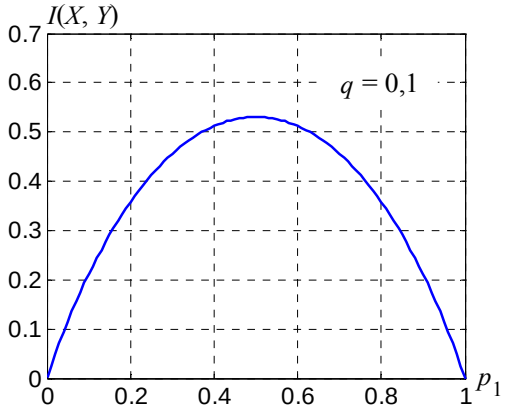
На рис. 1.10, *а* показан вид взаимной информации как функции от p_1 для заданного значения канальной вероятности $q = 0{,}1$. Видно, что $I(p_1)$ — выпуклая функция, принимающая нулевые значения при $p_1 = 0$ и $p_1 = 1$, а максимальное значение, равное в общем случае

$$I_{\max}(X; Y) = 1 + q \log q I(1 - q)\log(1 - q),$$

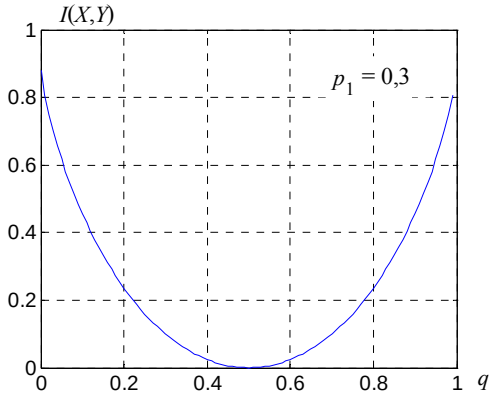
достигается при $p_1 = 0{,}5$, т. е. при равновероятных входных сообщениях.

На рис. 1.9, *б* показан вид взаимной информации как функции от q для заданного значения входной вероятности $p_1 = 0{,}3$, из которого видно, что $I(q)$ — вогнутая функция, равная нулю при $q = 0{,}5$ и принимающая при $q = 0$ и $q = 1$ максимальное значение

$$-p_1 \log p_1 - (1 - p_1)\log(1 - p_1).$$



а)



б)

Рис. 1.10. Взаимная информация для ДСК: *а*) — при заданном $q = 0{,}1$; *б*) — при заданном $p_1 = 0{,}3$

Обратимся теперь к вопросу о методики нахождения максимума взаимной информации для более широкого класса источников. Одним из возможных путей является использование метода неопределённых множителей Лагранжа, уже применённого в разд. 1.2 для максимизации энтропии, когда максимизируется безусловным образом функция $f(\mathbf{v}) - \lambda \sum_{(k)} v_k$, а значение множителя λ находится из условия $\sum_{(k)} v_k = 1$.

Однако часто встречаются задачи поиска экстремума при ограничениях, представляющих собой неравенства, и желательно иметь необходимые (а по возможности — и достаточные) условия экстремума, не зависящие от типа ограничений и от свойств функций, задающих ограничения. В общем виде такие задачи, представляющие собой предмет нелинейного программирования [6. 7], формулируются следующим образом:

найти максимум функции $f(\mathbf{v})$ *при заданных ограничениях-равенствах*

$$u_i(\mathbf{v}) = 0,\ i = 1, \ldots, t \qquad (1.8.10а)$$

и ограничениях-неравенствах

$$g_j(\mathbf{v}) \geq 0,\ j = 1, \ldots, s. \qquad (1.8.10б)$$

Если в ограничении фигурирует обратное неравенство $g_j(\mathbf{v}) \leq 0$, то его можно трактовать как $-g_j(\mathbf{v}) \geq 0$, сводя условие к (1.8.10*б*). Также задачу минимизации функции $f(\mathbf{v})$ можно трактовать как задачу максимиза-

ции функции – $f(\mathbf{v})$, так что представленная формулировка является общей для целого класса задач.

Фигурирующие в (1.8.10*б*) ограничения-неравенства для разных точек многомерного пространства можно разделить на два множества: *активные*, для точек в которых $g_j(\mathbf{v}) = 0$, и *неактивные*, для точек в которых $g_j(\mathbf{v}) > 0$. Если функции $g_j(\mathbf{v})$ непрерывны, то существует некоторая окрестность точки $\mathbf{v}$, в пределах которой сохраняют свои действия неактивные ограничения.

Заметим, что если ограничения-равенства $u_i(\mathbf{v}) = 0$ линейные, то они могут быть записаны в виде двух ограничений-неравенств: $u_i(\mathbf{v}) \geq 0$ и $u_i(\mathbf{v}) \leq 0$, поэтому иногда в формулировке задачи оптимизации соотношения (1.8.10*а*) отсутствуют.

При нахождении условного экстремума функции многих переменных широко используется результат, известный в теории нелинейного программирования как *теорема Куна–Таккера*.

Пусть $\mathbf{v}_0$ *— вектор-решение, дающий максимум функции* $f(\mathbf{v})$ *при ограничениях* (1.8.10). *Тогда, при достаточно общих предположениях о свойствах функций* $f(\mathbf{v})$, $u_i(\mathbf{v})$ *и* $g_j(\mathbf{v})$ *существует набор чисел* $\lambda_j \geq 0$ ($j = 1, \ldots, s$) и γ_i ($i = 1, \ldots, t$), *таких что* $\lambda_j g_j(\mathbf{v}_0) = 0$, *и имеет место равенство*

$$\nabla f(\mathbf{v}_0) + \sum_{j=1}^{s} \lambda_j \nabla g_j(\mathbf{v}_0) + \sum_{i=1}^{t} \gamma_i \nabla u_i(\mathbf{v}_0) = 0 . \qquad (1.8.11)$$

Равенство $\lambda_j g_j(\mathbf{v}_0) = 0$ означает, что в каждой допустимой точке, т. е. в точке, удовлетворяющей соотношениям (1.8.10), либо $g_j(\mathbf{v}) = 0$ (действие активного ограничения), либо значение $\lambda_j = 0$, и соответствующее слагаемое выпадает из суммы.

Условия Куна–Таккера сами по себе не дают метода отыскания экстремальных точек, и в общем случае нахождение таких точек аналитическими методами затруднительно. Однако применение (1.8.11) к вероятностным векторам позволяет существенно упростить задачу. В частности, оказывается (см. задачу 1.13), что для выпуклых функций условия Куна–Таккера являются не только необходимыми, но и достаточными. Докажем это.

Пусть множество векторов, на котором ищется экстремум, является выпуклым множеством вероятностных векторов $\mathbf{p} = (p_1, \ldots p_l)$, определяемым совокупностью $l + 1$ ограничений

$$u(\mathbf{p}) = \sum_{k=1}^{l} p_k - 1 = 0 \qquad (1.8.12a)$$

и

$$g_j(\mathbf{p}) = p_j \geq 0,\ j = 1, \ldots, l. \qquad (1.8.12б)$$

Тогда, поскольку $\partial u(\mathbf{p}) / \partial p_k = 1$ и $\partial g_j(\mathbf{p}) / \partial p_k = \delta_{jk}$ ($j, k = 1, \ldots, l$), использование (1.8.11) приводит к соотношениям

$$\frac{\partial f(\mathbf{p})}{\partial p_j}\Big|_{\mathbf{p}=\mathbf{p}_0} = -\lambda_j - \gamma .$$

Так как λ_j — неотрицательные числа ($\lambda_j > 0$ при активном ограничении $p_j = 0$ и $\lambda_j = 0$ при неактивном ограничении $p_j > 0$), переобозначая $\mu = -\gamma$, можно сказать, что $\partial f(\mathbf{p}) / \partial p_j \,|_{\mathbf{p}=\mathbf{p}_0} = \mu$, если для j-й координаты $p_j > 0$, и $\partial f(\mathbf{p}) / \partial p_j \,|_{\mathbf{p}=\mathbf{p}_0} \leq \mu$, если для j-й координаты $p_j = 0$.

Итак, условия Куна–Таккера для максимизации функции на множестве вероятностных векторов можно сформулировать в следующем виде:

Для того, чтобы вероятностный вектор $\mathbf{p}$ *максимизировал выпуклую функцию* $f(\mathbf{v})$ *при наличии ограничений* (1.8.12) *необходимо и достаточно существование числа* μ*, такого, что*

$$\frac{\partial f(\mathbf{p})}{\partial p_j}\Big|_{\mathbf{p}=\mathbf{p}_0} \begin{cases} = \mu, & p_{0j} > 0 \\ \leq \mu, & p_{0j} = 0 \end{cases},\ j = 1, \ldots, l. \qquad (1.8.13)$$

Условия (1.8.13) в принципе не отличаются от необходимых условий максимизации функции

$$f(\mathbf{p}) - \mu\left(\sum_{k=1}^{l} p_k - 1\right)$$

традиционным методом неопределённых множителей Лагранжа, и если максимум находится внутри области B существования вероятностных векторов, когда $p_j > 0$ для всех j, то (1.8.13) выполняются со знаком равенства. Однако в общем случае нет гарантии того, что решение удовлетворяет вероятностным ограничениям. В частности, может оказаться, что *безусловный* максимум функции $f(\mathbf{p})$ лежит вне области B в точке, где хотя бы одна

из компонент p_j меньше нуля. Отсюда следует, что *условный максимум функции* $f(\mathbf{p})$ *лежит на границе области B*, где одна или более компонент p_j равны нулю, как это показано на рис. 1.11 для случая $l = 2$.

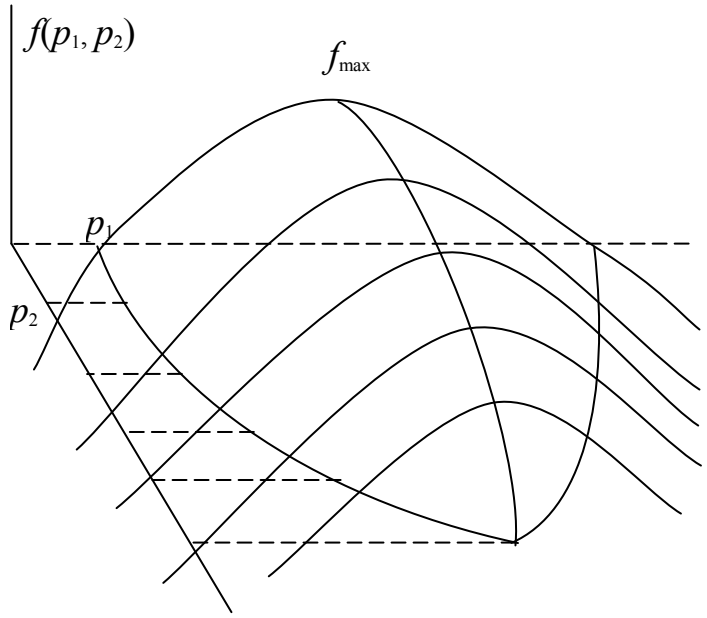


Рис. 1.11. Вид выпуклой функции с максимумом при $p_2 = 0$

Таким образом, если будет найден вероятностный вектор $\mathbf{p}_0$ и число μ, удовлетворяющие соотношениям (1.8.13), то задача максимизации функции $f(\mathbf{p})$ будет решена. К сожалению, в большинстве случаев аналитическое решение этой задачи крайне затруднительно, и требуется привлечение методов нелинейного программирования. Тем не менее, в ряде случаев удаётся решить задачу, не прибегая к таким методам, что и будет продемонстрировано в последующих разделах.

В заключение данного раздела заметим, что минимизация выпуклой функции $f(\mathbf{p})$ на множестве вероятностных векторов производится, исходя из точно таких же соображений. Формально данную задачу можно трактовать как задачу максимизации на том же множестве функции $-f(\mathbf{p})$. При этом необходимыми и достаточными условиями является существование вероятностного вектора $\mathbf{p}_0$ и числа μ, удовлетворяющих соотношениям

$$\left.\frac{\partial f(\mathbf{p})}{\partial p_j}\right|_{\mathbf{p}=\mathbf{p}_0} \begin{cases} = \mu, & p_{0j} > 0 \\ \geq \mu, & p_{0j} = 0 \end{cases}, \; j = 1, \ldots, l. \qquad (1.8.13a)$$

1.9. ЭПСИЛОН-ЭНТРОПИЯ

Как было показано в разделе 1.4, энтропия непрерывной величины стремится к бесконечности при уменьшении интервала квантования, поскольку неопределённость реализации одного из бесконечного множества состояний может быть сколь угодно велика. Однако требование бесконечной точности воспроизведения непрерывной реализации вследствие ограниченной разрешающей способности информационно-измерительных систем не имеет смысла. С точки зрения наблюдателя события, отличающиеся (в определённом смысле) на величину, не большую некоторой заданной, определяемой либо конструктивными особенностями устройств наблюдения, либо природными особенностями самого наблюдателя (зрением, слухом), должны восприниматься для него как несущие идентичную информацию.

Более того, подобные соображения можно отнести и к дискретным источникам, несмотря на то, что взаимная информация для них всегда является конечной величиной: анализ дискретного источника посредством наблюдения его реализаций, отличающихся от реализаций исходного источника на некоторую заданную величину, должен давать *практически* ту же информацию. Таким образом, уместно дать такое определение количества информации, которое опиралось бы на точность воспроизведения исходного сообщения, независимо от того, является он дискретным или непрерывным.

Пусть источник X в каждый момент времени выбирает дискретное или непрерывное сообщение x. Анализ источника X посредством наблюдений за вторичным источником Y будем трактовать как замену (аппроксимацию) сообщения $x \in X$ значением $y \in Y$.

Качество, или ошибку аппроксимации будем характеризовать мерой различия d между элементами x и y. Ввести меру различия можно, разумеется, многими способами: как детерминированными, посредством отображения X в Y, так и вероятностными с помощью задания набора условных вероятностей $p(y / x)$ или условных плотностей вероятностей $w(y / x)$[1]. При

[1] Заметим, что детерминированный способ аппроксимации также может быть описан с помощью условных вероятностей в виде вырожденного распределения

этом нужно помнить, что отображение x в y, как правило, неоднозначно: имеется несколько (в непрерывном случае — бесконечно много) значений $x \in X$, которые аппроксимируются одним и тем же значением $y \in Y$.

Итак, пусть источник X задан набором вероятностей $p(x)$ или плотностью вероятности $w(x)$. Тогда для каждого набора условных распределений $p(y/x)$ или условной плотности вероятности $w(y/x)$ определена функция $d(x, y)$, представляющая собой случайную величину на ансамбле XY. Ее математическое ожидание, представляющее собой среднюю ошибку аппроксимации, равно

$$\Delta = \mathbf{E}\big[d(x,y)\big] = \sum_X \sum_Y d(x,y)\, p(y/x)\, p(x), \qquad (1.9.1a)$$

когда источники X и Y дискретны,

$$\Delta = \mathbf{E}\big[d(x,y)\big] = \int_X \int_Y d(x,y) w(y/x) w(x) dx dy \qquad (1.9.1б)$$

когда источники X и Y непрерывны.

Например, если в качестве меры различия выбрать квадрат разности между значениями x и y, т. е.

$$d(x,y) = \big(x - y\big)^2, \qquad (1.9.2)$$

то средняя ошибка является среднеквадратической. В частности, при нулевом математическом ожидании $\mathbf{E}[(x - y)] = 0$ выражение (1.9.1) представляет собой дисперсию ошибки. Выбор квадратичного критерия качества целесообразен в том случае, когда особенно нежелательны большие ошибки.

Поскольку описание дискретного источника может быть осуществлено также плотностями вероятностей с помощью соответствующего набора дельта-функций, далее в этом разделе вводимые понятия будут, главным образом, касаться непрерывных источников с возможной переформулировкой по отношению к дискретному случаю. Если какие-то примеры будут относиться исключительно к дискретным источникам, то это будет особо оговариваться.

$$p(x/y) = \begin{cases} 1, & y = y(x); \\ 0, & y \neq y(x); \end{cases}$$

для дискретного или в виде дельта функции $w(y/x) = \delta\big(y - y(x)\big)$ — для непрерывного случая, так что вероятностная аппроксимация включает в себя детерминированную.

Наряду со средней ошибкой аппроксимации на ансамбле XY определим взаимную информацию $I(X; Y)$ между источниками X и Y:

$$I(X;Y) = \int_X \int_Y w(y/x) w(x) \log \frac{w(y/x)}{w(y)} dxdy, \qquad (1.9.3)$$

где $w(y)$ находится по заданным плотностям $w(x)$ и $w(x/y)$:

$$w(y) = \int_X w(y/x) w(x) dx. \qquad (1.9.4)$$

Введём далее класс функций $\Omega(\varepsilon)$, куда поместим все условные плотности вероятностей $w(y/x)$, для которых средняя ошибка аппроксимации Δ не превосходит некоторого заданного значения ε:

$$\Omega(\varepsilon) = \{ w(y/x) : \Delta \le \varepsilon \}. \qquad (1.9.5)$$

Тогда функция от ε вида

$$H(\varepsilon) = \min_{w(y/x)} I(X;Y), \qquad (1.9.6)$$

где минимум разыскивается по всем условным плотностям $w(y/x)$, принадлежащим классу $\Omega(\varepsilon)$, называется[1] *эпсилон-энтропией источника X относительно меры аппроксимации d*.

Боле строго эпсилон-энтропия определяется не через минимум, а через инфинум, т. е. точную нижнюю грань множества.

Инфинумом $\inf_{a \in A} A$ множества A называется число a_0, такое, что для всех элементов a из A справедливо неравенство $a_0 \le a$, и каково бы ни было $\gamma > 0$, существует такой элемент $a' \in A$, что $a' < a_0 + \gamma$. Если инфинум достигается на самом множестве A, т. е. $a_0 \in A$, то он называется минимумом множества $\min_{a \in A} A$. В частности, минимум всегда существует для конечного множества. Если A — бесконечное множество, то инфинум может не достигаться ни на одном элементе из A. Пусть, например, A — множество значений функции $f(n) = 1 + 1/n$, заданной на натуральном ряде чисел 1, 2, … . Нетрудно видеть, что $\inf A = 1$, но $f(n) \ne 1$ ни для одного элемента из A.

В большинстве практических задач минимум и инфинум на рассматриваемых множествах совпадают, поэтому за определение эпсилон-энтропии примем (1.9.6).

Итак, эпсилон-энтропия — это минимальное по всем условным распределениям $w(x/y)$ (а поскольку задано $w(x)$, то и по совместным распре-

---

[1] Термин “эпсилон-энтропия” предложен А. Н. Колмогоровым. Иногда в литературе, особенно, в переводной англоязычной, употребляются термины “скорость создания сообщений” (предложен К. Э. Шенноном), “скорость как функция искажения”, “функция скорость–искажение” и др.

делениям $w(y, x)$) количество взаимной информации, необходимое для того, чтобы значениями наблюдаемого источника Y воспроизвести с заданной точностью значения исходного источника X.

Рассмотрим некоторые свойства введённого понятия.

Прежде всего, заметим, что эпсилон-энтропия как средняя взаимная информация является неотрицательной величиной. Далее, определённая для неотрицательных значений параметра точности ε, она является невозрастающей функцией от ε. Действительно, если $\varepsilon_1 > \varepsilon_2$, то, как нетрудно видеть из (1.9.5), класс $\Omega(\varepsilon_1)$ содержит в себе класс $\Omega(\varepsilon_2)$. А поскольку минимум по некоторому множеству не может быть больше, чем минимум по части этого множества, следует заключить, что $H(\varepsilon_1) \le H(\varepsilon_2)$. Таким образом, минимальное значение эпсилон-энтропии есть нуль.

Как было показано в разд. 1.8, средняя взаимная информация является либо выпуклой функцией на множестве входных распределений $w(x)$, либо вогнутой функцией на множестве условных распределений $w(y / x)$. Для эпсилон-энтропии это свойство однозначно конкретизируется: *эпсилон-энтропия* $H(\varepsilon)$ *является вогнутой функцией по отношению к параметру* ε.

Покажем, что это действительно так. Рассмотрим два класса $\Omega(\varepsilon_1)$ и $\Omega(\varepsilon_2)$ в которых минимум (1.9.6) достигается на условных плотностях $w_1(y / x)$ и $w_2(y / x)$ соответственно. Выберем произвольное число $\alpha \in [0;1]$ и с его помощью построим функцию

$$w(y / x) = \alpha w_1(y / x) + (1 - \alpha) w_2(y / x),$$

для которой, как нетрудно видеть,

$$\int_X \int_Y d(x, y) w(y / x) w(x) dx = \int_X \int_Y d(x, y) \left[\alpha w_1(y / x) + (1 - \alpha) w_2(y / x)\right] w(x) dx \le$$

$$\le \int_X \int_Y \left[\alpha w_1(y / x)\varepsilon_1 + (1 - \alpha) w_2(y / x)\varepsilon_2\right] w(x) dx = \alpha\varepsilon_1 + (1 - \alpha)\varepsilon_2 .$$

Отсюда можно заключить, что условная плотность вероятности $w(y / x)$, принадлежит классу

$$\Omega\left(\alpha\varepsilon_1 + (1 - \alpha)\varepsilon_2\right) = \left\{w(y / x) : \Delta \le \alpha\varepsilon_1 + (1 - \alpha)\varepsilon_2\right\},$$

а соответствующая эпсилон-энтропия $H\left(\alpha\varepsilon_1 + (1 - \alpha)\varepsilon_2\right)$ как минимальная взаимная информация удовлетворяет неравенству

$$H\left(\alpha\varepsilon_1 + (1 - \alpha)\varepsilon_2\right) \le I(X; Y),$$

где $I(X;\ Y)$ — средняя взаимная информация между источниками X и Y, вычисленная по функции $w(y / x)$.

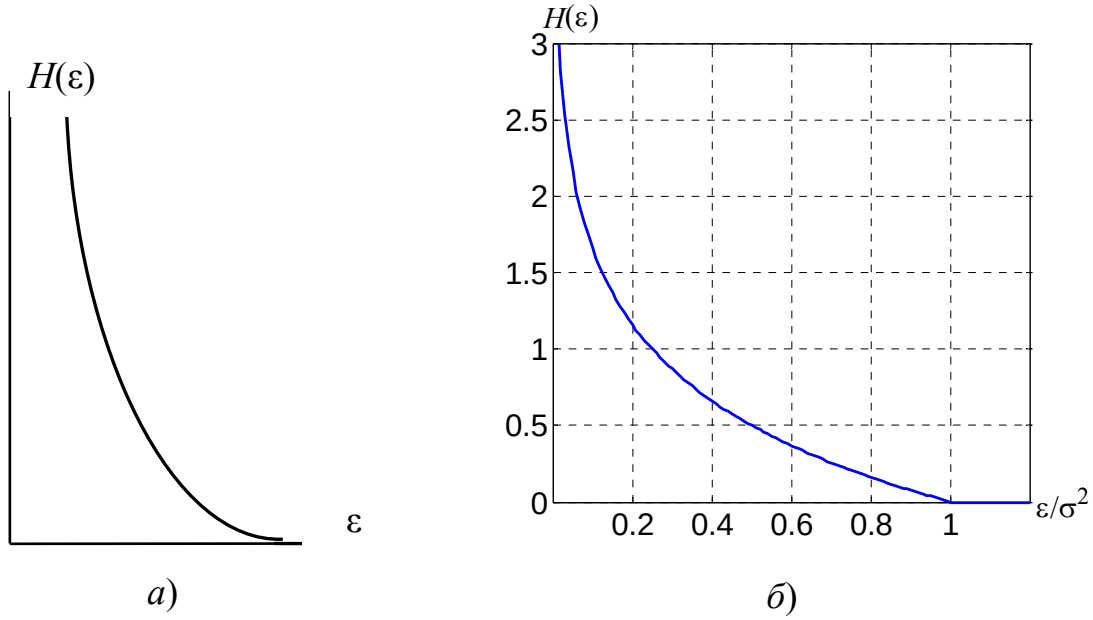
Теперь, используя свойство вогнутости взаимной информации относительно условных распределений, можно записать соотношение

$$H\left(\alpha\varepsilon_1 + (1-\alpha)\varepsilon_2\right) \le I(X;Y) \le$$

$$\le \alpha I_1(X;Y) + (1-\alpha) I_2(X;Y) = \alpha H(\varepsilon_2) + (1-\alpha) H(\varepsilon_2),$$

означающее вогнутость эпсилон-энтропии относительно параметра ε.

Для иллюстрации доказанных свойств на рис. 1.12, *а* показан типичный вид эпсилон-энтропии как функции ε.



а) *б*)

Рис. 1.12. Эпсилон-энтропия: *а*) — типичный график;
б) — для гауссовского источника

Нахождение эпсилон-энтропии в общем случае представляет собой достаточно сложную задачу. Рассмотрим пример её вычисления для простого, но важного случая — гауссовского источника X с нулевым средним, описываемого плотностью вероятности

$$w(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{x^2}{2\sigma^2}\right],$$

относительно непрерывного источника Y при использовании квадратичной меры (1.9.2). При этом для всех условных вероятностей $w(y / x)$ из класса $\Omega(\varepsilon)$ средняя ошибка аппроксимации есть

$$\Delta = \int_X \int_Y (x-y)^2 w(y/x) w(x) dx dy \le \varepsilon . \qquad (1.9.7)$$

Другими словами, значения $y \in Y$ с некоторой погрешностью z воспроизводят значения $x \in X$, т. е. $x = y + z$.

Задача вычисления эпсилон-энтропии на самом деле содержит себе две подзадачи: во-первых, найти минимальное значение (1.9.6), а во-вторых, указать то условное распределение $w(y/x)$ (или, возможно, семейство таких распределений), на которых этот минимум достигается.

Запишем (1.9.6) через разность дифференциальных энтропий:

$$H(\varepsilon) = \min_{\Omega(\varepsilon)} \left[h(X) - h(X/Y) \right].$$

Первое слагаемое $h(X)$ не зависит от условных плотностей $w(y/x)$ и для гауссовского источника, согласно (1.5.13), равно $(1/2)\log 2\pi e\sigma^2$. Поэтому можно записать, что

$$H(\varepsilon) = \frac{1}{2}\log 2\pi e\sigma^2 - \max_{\Omega(\varepsilon)} h(X/Y) .$$

Поскольку $z = x - y$, при фиксированном значении y величины z и x отличаются друг от друга только математическим ожиданием и, следовательно, имеют одинаковые дифференциальные энтропии. Тогда, вводя в рассмотрение источник Z всех возможных значений z (этот источник физически не существует; он является “внутренней переменной” задачи) и учитывая свойства дифференциальной энтропии, можно записать

$$h(X/Y) = h(Z/Y) \le h(Z) , \qquad (1.9.8)$$

откуда

$$H(\varepsilon) \ge \frac{1}{2}\log 2\pi e\sigma^2 - \max_{\Omega(\varepsilon)} h(Z) .$$

Так как источник Z — гауссовский, для которого дисперсия не превосходит величины ε, учитывая экстремальность гауссовского распределения по отношению к дифференциальной энтропии, когда

$$h(Z) \le \frac{1}{2}\log 2\pi e\varepsilon , \qquad (1.9.9)$$

справедлива оценка

$$H(\varepsilon) \ge \frac{1}{2}\log 2\pi e\sigma^2 - \frac{1}{2}\log 2\pi e\varepsilon = -\frac{1}{2}\log\frac{\varepsilon}{\sigma^2} = \frac{1}{2}\log\frac{\sigma^2}{\varepsilon} . \qquad (1.9.10)$$

Тогда, если найдётся такое распределение $w(y/x)$, для которого неравенство (1.9.10) превращается в равенство, то величина $(1/2)\log(\sigma^2/\varepsilon)$ и есть искомая эпсилон-энтропия. Искомое равенство будет достигнуто, если, в свою очередь, будет достигнуто равенство в соотношениях (1.9.8) и (1.9.9). Попробуем выбрать условную плотность $w(y/x) \in \Omega(\varepsilon)$, удовлетворяющую указанным требованиям.

Равенство в (1.9.8) возможно в том и только в том случае, когда источники Y и Z независимы. Равенство в (1.9.9) достигается, когда Z — гауссовский случайный источник с нулевым средним и дисперсией ε. Таким образом, с неизбежностью оказывается что Y также гауссовский, т. е. источник X является суммой независимых гауссовских источников Y и Z. При этом, когда источник X имеет дисперсию σ^2, дисперсии источников Y и Z равны соответственно $(\sigma^2 - \varepsilon)$ и ε.

Выпишем искомую плотность $w(y/x)$ в явном виде. Если $w_X(x)$, $w_Y(y)$ и $w_Z(z) = w_Z(y - x)$ — безусловные плотности источников X, Y и Z соответственно, то функция

$$w(y/x) = \frac{w(x/y)w_Y(y)}{w_X(x)} = \frac{w_Z(z)w_Y(y)}{w_X(x)} = \frac{w_Z(x-y)w_Y(y)}{w_X(x)}, \qquad (1.9.11)$$

и является тем условным распределением, на котором реализуется эпсилон-энтропия.

Соотношения (1.9.10) и (1.9.11), казалось бы, полностью решают поставленную задачу. Однако в приведённых рассуждениях имеется одна тонкость: они справедливы лишь при условии $\sigma^2 \geq \varepsilon$, определяющем возможность существования источника Z с неотрицательной дисперсией. В противном случае такого источника существовать не может, и эпсилон-энтропия должна вычисляться каким-то другим способом.

Оказывается, что для случая $\sigma^2 < \varepsilon$ задача нахождения эпсилон-энтропии оказывается ещё более простой. Дело в том, что при таком соотношении существует возможность универсальной аппроксимации нулевым значением источника Y, поскольку при $y = y_0 = 0$ среднеквадратическая ошибка равна

$$\Delta = \mathbf{E}\left[(X - y_0)^2\right] = \mathbf{E}\left[X^2\right] = \sigma^2 < \varepsilon.$$

Это означает, что $w(y/x) = \delta(y - y_0) = \delta(y)$ и, как следствие, $H(\varepsilon) = 0$.

Полученные результаты можно объединить одной формулой:

$$H(\varepsilon)=\frac{1}{2}\log\max\left\{1,\sigma^2/\varepsilon\right\}, \qquad (1.9.12)$$

которая и является окончательным решением задачи. На рис. 1.12, *б* представлена зависимость эпсилон-энтропии как функции нормированного аргумента ε / σ^2.

ВОПРОСЫ И ЗАДАНИЯ К ГЛАВЕ 1

1.1. Показать, не прибегая к методике отыскания условного экстремума функций многих переменных, что энтропия двоичного источника как функция от вероятности p одного из сообщений имеет вид, показанный на рис. 1.1, *а*. Для этого с помощью производной найти области возрастания, убывания, экстремума и показать, что экстремальное значение есть максимум.

1.2. Пусть $X=\{x_1,\ x_2,\ \ldots,\ x_l\}$ — источник без памяти, и требуется найти такое распределение вероятностей появления элементов $p(x_k)$, при котором энтропия $H(X)$ максимальна, а $p(x_k)=\varepsilon$, где ε — заданное число. Показать, используя, например, метод Лагранжа, что

$$H(X)\le -\varepsilon\log\varepsilon-(1-\varepsilon)\log\frac{1-\varepsilon}{l-1},$$

а равенство достигается только в том случае, когда все сообщения кроме x_1 имеют одинаковые вероятности.

1.3. Докажите неравенство $\ln x\le x-1$ двумя способами: рассматривая экстремальные свойства функции $f(x)=\ln x-x+1$, а также используя свойства знакопеременных (лейбницевских) рядов. Каков геометрический смысл этого неравенства?

1.4. Пусть $X=\{x_1,\ x_2,\ \ldots\}$ — множество, состоящее из бесконечного числа элементов.

а) Привести пример такого распределения вероятностей, для которого энтропия

$$H(X)=-\sum_{k=1}^{\infty}p(x_k)\log p(x_k)$$

является конечной. В частности, каким должно быть распределение, чтобы $H(X)=2$?

б) Полагая, что $p(x_1)=1-p$, а $p(x_i)=q^{i-1}$, $i>1$, где p — заданное число, а $q<1$, найти q как функцию от p и показать, что энтропия источника равна

$$H(X)=-(1-p)\log(1-p)-\frac{q\log q}{1-q^2}.$$

1.5. Найти количество информации, содержащееся в одном кадре квантованного телевизионного сигнала, считая, что в кадре содержится 625 строк, сигнал, соответствующий каждой строке представляет собой последовательность из 600 случайных по амплитуде и некоррелированных импульсов, каждый из которых может равновероятно принимать значения от 0 до 8 В, а каждый импульс квантуется с шагом квантования 1 В.

1.6. Найти дифференциальную энтропию следующих распределений и выразить ее через дисперсию:

$$a)\ w(x)=\begin{cases}\dfrac{1}{b-a}, & x\in[a;b];\\ 0, & x\notin[a;b];\end{cases}; \qquad \underline{б})\ w(x)=\begin{cases}\dfrac{4(x-a)}{(b-a)^2}, & a<x<\dfrac{b+a}{2}\\ \dfrac{4(b-x)}{(b-a)^2}, & \dfrac{a+b}{2}<x<b;\\ 0, & x<a, x>b;\end{cases}$$

$$в)\ w(x)=\frac{1}{\pi}\frac{h}{h^2+(x-a)^2}; \quad г)\ w(x)=\frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right),\ x>0;$$

$$д)\ w(x)=\frac{\lambda}{2}\exp\left(-\lambda|x-\mu|\right); \qquad е)\ w(x)=\lambda\exp(-\lambda x),\ x>0.$$

1.7. Найти дифференциальную энтропию случайной величины $\xi=A\sin\omega t$, где A и ω — положительные постоянные величины, а t равномерно распределено на интервале $[-\pi/\omega;\pi/\omega]$.

1.8. Сигнал со средним значением a может принимать только положительные значения. Найти распределение значений такого сигнала, обладающее максимальной энтропией.

1.9. Показать, что при заданной энтропии нормальное распределение имеет наименьшую из всех одномерных распределений дисперсию.

1.10. Для рассмотренного в разд. 1.3 примера

КОЛОКОЛ_ОКОЛО_КОЛОКОЛА

найти энтропию и избыточность, полагая, что источник обладает памятью на два элемента. сравнить полученные результаты с аналогичными значениями для источника без памяти, а также источника с памятью на один элемент.

1.11. Докажите справедливость утверждения 2 в разд. 1.9 для взаимной информации в случае дискретных и непрерывных ансамблей.

1.12. Пусть непрерывный источник X имеет равномерное распределение $w_X(x) = 1/(2c)$ на интервале $[-c; c]$, и для источника Y задана условная плотность $w(y/x) = 1/(2b)$, $y \in [-b + x; b + x]$, $b \le c$. Найти:

а) распределение значений источника Y;

б) взаимную информацию между источниками X и Y;

в) среднюю ошибку аппроксимации для критерия качества $d(x, y) = |x - y|$;

г) эпсилон-энтропию относительно указанного $d(x, y)$.

1.13. Докажите, что условия Куна–Таккера являются достаточными для максимизации выпуклой функции на множестве вероятностных векторов.

1.14. Покажите, что если плотности вероятности $w_1(x)$ и $w_2(x)$ отличаются только математическим ожиданием, то соответствующие им дифференциальные энтропии совпадают. Обобщите этот результат на многомерный случай.

ГЛАВА 2. ИНФОРМАЦИОННЫЕ ХАРАКТЕРИСТИКИ СЛУЧАЙНЫХ ПРОЦЕССОВ

В предыдущем разделе были рассмотрены основные свойства энтропии для случайных величин. Обобщим эти свойства таким образом, чтобы иметь возможность оценивать информационные характеристики *случайных процессов*. Некоторые шаги в данном направлении уже были предприняты, когда имело место рассмотрение количества информации между многомерными случайными величинами. Однако при этом не производилось принципиального действия — *предельного перехода* в информационных характеристиках, введённых для последовательностей с конечным числом элементов. Именно предельный переход определяет информационные характеристики случайных процессов, и в данной главе этому механизму и будет уделено основное внимание.

Как известно [8], случайный процесс $\xi(t)$ можно задать посредством совокупности случайных величин, определённых на множестве значений параметра t, который для будем считать временем. Обратимся сначала к простейшему случаю дискретного времени и конечного множества возможных значений процесса в каждый момент времени, а затем обобщим введённые понятия на континуальные множества.

2.1. ЭНТРОПИЯ ДИСКРЕТНОГО СЛУЧАЙНОГО ПРОЦЕССА

Рассмотрим конечные реализации случайного процесса, состоящие из n случайных величин. Если множество значений каждой случайной величины содержит l элементов, то общее число отличных друг от друга реализаций равно l^n, и появление на выходе источника каждой такой реализации можно рассматривать как исход некоторой случайной величины $\Psi_n = (\xi^{(n)}, \xi^{(n-1)}, \ldots, \xi^{(1)})$, имеющей l^n значений. Зная распределение исходных "элементарных" случайных величин и характер связей между отдельными величинами, задаваемых соответствующими условными вероятностями, в принципе можно вычислить вероятность $p_k(\Psi_n)$ каждой из l^n реализаций величины Ψ_n, которые равны

$$p_k\left(\Psi_n\right) = P\left(\xi^{(1)}\right) P\left(\xi^{(2)} / x^{(1)}\right) \cdots P\left(\xi^{(n)} / x^{(n-1)}, \ldots, x^{(1)}\right). \qquad (2.1.1)$$

Обозначим условную энтропию n-го события при известных предыдущих ν событиях через

$$H\left(\xi^{(n)} / X^{\nu}\right) \equiv H\left(\xi^{(n)} / x^{(n-1)}, \ldots, x^{(n-\nu)}\right), \qquad n > \nu \geq 0, \qquad (2.1.2)$$

где $X^{\text{н}}$ означает произведение ансамблей, образованное предыдущими ν событиями.

Пусть $A = \{a_1, \ldots, a_M\}$ и $B = \{b_1, \ldots, b_N\}$ — два конечных множества. Тогда множество, элементы которого представляют собой все возможные упорядоченные пары (a_i, b_j), $i = 1, \ldots, M$, $j = 1, \ldots, N$, называется *произведением множеств* A и B и обозначается AB. Согласно этому определению множества AB и BA, вообще говоря, различаются. Если A и B совпадают, то произведение множеств обозначается как A^2 (или B^2).

Аналогичным образом определяются произведения более чем двух множеств. А именно, произведение $A_1 A_2 \ldots A_\nu$ представляет собой множество всех последовательностей $(a^{(1)}, a^{(2)}, \ldots, a^{(\nu)})$ длины ν таких, что первый элемент $a^{(1)}$ принадлежит множеству A_1, второй элемент $a^{(2)}$ — множеству A_2 и т. д., ν-й элемент — множеству A_ν. Если все множества совпадает между собой $A_1 = A_2 = \ldots = A_\nu = A$, то такое произведение обозначается как A^ν.

В силу соотношений (1.6.10) и (1.6.10*a*) приведённые выше энтропии удовлетворяют неравенствам

$$H\left(\xi^{(n)} / X^{\nu}\right) \leq H\left(\xi^{(n)} / X^{\nu-1}\right) \leq \ldots \leq H\left(\xi^{(n)} / x^{(n-1)}\right) \leq H\left(\xi^{(n)}\right). \qquad (2.1.3)$$

Данные неравенства устанавливают тот факт, что условная энтропия n-го события является невозрастающей функцией от числа уже определённых предшествующих событий, однако они не указывают, каким образом величина $H(\xi^{(n)} / X^{\nu})$ изменяется при увеличении n. В частности, нельзя ничего сказать о монотонности её поведения, а это, в свою очередь, оставляет пока открытым вопрос о предельном поведении значений последовательности условных энтропий.

Более определённый ответ на такой вопрос может быть получен в классе стационарных процессов, для которых любые конечномерные распределения вероятностей не зависят от выбора начала отсчёта времени:

$$P\left(\xi^{(n)} / x^{(n-1)}, \ldots, x^{(n-\nu)}\right) = P\left(\xi^{(m)} / x^{(m-1)}, \ldots, x^{(m-\nu)}\right), \; n > \nu, \; m > \nu. \qquad (2.1.4)$$

В силу равенства (2.1.4) условные энтропии событий, порождаемые стационарным процессом, также зависят только от относительного положения этих событий в последовательности, например,

$$H\left(\xi^{(n)} / X^{\nu}\right) = H\left(\xi^{(m)} / X^{\nu}\right). \qquad (2.1.5)$$

Рассмотрим теперь условные энтропии последовательных событий при условии, что все предыдущие события заданы, т. е. рассмотрим последовательность энтропий $H(\xi^{(n)} / X^{n-1})$. Согласно (2.1.5), в каждой из условных энтропий, фигурирующих в (2.1.3), число n можно заменить на любое целое число, большее заданного числа предшествующих событий. В частности, в первой энтропии можно заменить n на $\nu + 1$, во второй — на ν, в третьей — на $\nu - 1$ и т. д. В результате таких замен получим следующую цепочку неравенств:

$$H\left(\xi^{(\nu+1)} / X^{\nu}\right) \le H\left(\xi^{(\nu)} / X^{\nu-1}\right) \le \ldots \le H\left(\xi^{(2)} / x^{(1)}\right) \le H\left(\xi^{(1)}\right) \le \log l, \qquad (2.1.6)$$

теперь уже устанавливающую тот факт, что для стационарного процесса условные энтропии событий при условии, что заданы все предшествующие события, образуют монотонно невозрастающую последовательность. Следовательно, для рассматриваемой последовательности существует предельное значение

$$\lim_{\nu\to\infty} H\left(\xi^{(\nu)} / X^{\nu-1}\right) = H\left(\xi / X^{\infty}\right), \qquad (2.1.7)$$

принадлежащее, как это следует из общих свойств энтропии, интервалу $[0; \log l]$.

Существование предела (2.1.7) для последовательности условных энтропий наводит на мысль о возможности существования подобной характеристики и для безусловной энтропии случайного процесса. Рассмотрим энтропию множества n-элементных реализаций

$$H\left(\Psi_n\right) = -\sum_{k=1}^{l^n} p_k\left(\Psi_n\right) \log p_k\left(\Psi_n\right). \qquad (2.1.8)$$

Значение $H(\Psi_n)$, разумеется, в общем случае зависит от n. Желая образовать инвариантную по отношению к n характеристику случайного процесса, введём в рассмотрение величину

$$H = \lim_{n\to\infty} H_n = \frac{H\left(\Psi_n\right)}{n}, \qquad (2.1.9)$$

имеющую смысл энтропии стационарного источника, отнесённой к одному сообщению. Если указанный предел существует, то величина H и будет являться энтропией случайного процесса.

Покажем, что предел (2.1.9) действительно существует. Для этого на

основании (1.6.8) запишем энтропию множества n-элементных реализаций $H(\Psi_n)$ произвольного стационарного процесса в виде

$$H\left(\Psi_n\right) \equiv H\left(\xi^{(1)},\ldots,\xi^{(n)}\right) = H\left(\xi^{(1)}\right) + H\left(\xi^{(2)} / x^{(1)}\right) + H\left(\xi^{(3)} / x^{(2)}x^{(1)}\right) +$$

$$+\ldots + H\left(\xi^{(n-1)} / X^{n-2}\right) + H\left(\xi^{(n)} / X^{n-1}\right) = \sum_{k=1}^{n} H\left(\xi^{(k)} / X^{k-1}\right).$$

Данное представление наряду со свойством (2.1.6) монотонного невозрастания условной энтропии позволяет сделать следующую оценку:

$$H\left(\Psi_n\right) \geq nH\left(\xi^{(n)} / X^{n-1}\right) \geq nH\left(\xi^{(n+1)} / X^{n}\right), \qquad (2.1.10)$$

откуда

$$H\left(\Psi_{n+1}\right) = H\left(\Psi_n\right) + H\left(\xi^{(n+1)} / X^{n}\right) \leq \frac{n+1}{n} H\left(\Psi_n\right),$$

или, согласно (2.1.9),

$$H_{n+1} \leq H_n .$$

Итак, последовательность $\{H_n\}$ есть монотонно невозрастающая положительная последовательность, имеющая, следовательно, предел при $n \to \infty$, который и может считаться инвариантной характеристикой всего наблюдаемого процесса, т. е. характеристикой дискретного стационарного источника.

Введённое понятие энтропии случайного процесса обладает следующим интересным и важным свойством.

Пусть стационарный источник сообщений с объёмом алфавита l_1 имеет (с учётом всех вероятностных связей) энтропию H_1 и избыточность ρ_1. Произведём операцию, называемую *укрупнением алфавита*, при которой каждая последовательность[1] $\left(x_{k_1}^{(1)}, x_{k_2}^{(2)}, \ldots, x_{k_n}^{(n)}\right)$ из любых n элементов первичного источника считается одним элементом y_i нового, вторичного источника. Очевидно, объем вторичного источника есть $l_2 = l_1^n$, а его энтропия равна $H_2 = nH_1$.

---

[1] Здесь и в дальнейших обозначениях элементов $x_{k_\nu}^{(\nu)}$ ($k_\nu = 1, \ldots, l$) верхний индекс показывает порядковый номер элемента в какой-то последовательности, а нижний — позицию этого элемента в алфавите. Часто, чтобы не загромождать запись излишней детализацией, верхние или нижние индексы — если это не влияет на существо изложения — могут быть опущены.

Определим избыточность ρ_2 вторичного источника. Максимальная энтропия для источника с объёмом $l_2 = l_1^n$ составляет

$$H_{2\max} = \log l_2 = n \log l_1,$$

откуда, учитывая что $\rho_1 = 1 - H_1 / \log l_1$ — избыточность первичного источника, имеем

$$\rho_2 = 1 - \frac{H_2}{H_{2\max}} = 1 - \frac{nH_1}{n \log l_1} = 1 - \frac{H_1}{\log l_1} = \rho_1 .$$

Из последнего выражения следует, что при укрупнении алфавита избыточность не изменяется. Однако, по крайней мере, интуитивно понятно (более строго этот вопрос рассмотрен далее), что укрупнение алфавита ослабляет взаимные вероятностные связи между элементами сообщения: если выбрать величину n достаточно превосходящей протяжённость действия вероятностных связей между элементами первичного источника, то вероятностными связями между укрупнёнными элементами можно пренебречь. Поскольку избыточность в процессе укрупнения не изменяется, то она должна практически полностью определяться неравномерностью распределения элементов вторичного источника. Таким образом, операция укрупнения алфавита может служить для “декорреляции” элементов сообщения, т. е. для устранения взаимных вероятностных связей между ними. Данное свойство будет впоследствии использовано при рассмотрении алгоритмов сжатия информации.

2.2. МАРКОВСКИЕ ИСТОЧНИКИ

Многие общие идеи предыдущего раздела могут быть более понятны при рассмотрении специального класса источников, называемых *марковскими*, о которых уже упоминалось в разд. 1.3. Такие источники определяются алфавитом сообщений $X = \{x_1, \ldots, x_l\}$, множеством *состояний* $\{S_1, \ldots, S_L\}$, а также вероятностями перехода между состояниями. В каждый момент времени источник порождает определённое сообщение (“букву”) и переходит в какое-то состояние, так что возникает (в общем случае — бесконечная) последовательность сообщений $\mathbf{x} = \{x^{(1)}, x^{(2)}, \ldots\}$ и соответствующая им последовательность состояний $\mathbf{S} = \{S^{(1)}, S^{(2)}, \ldots\}$. Будем считать, что когда генерируется n-элементная последовательность $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$, источник находится в некотором состоянии S_q ($q = 1, \ldots, L = l^n$).

Изучая вероятностные характеристики источника, можно рассматривать вероятности последовательных сообщений $P\left\{x_k^{(b)} / x^{(b-1)}, x^{(b-2)}, \ldots\right\}$, которые, вообще говоря, будут меняться при учёте все большего числа предыдущих символов. Однако, если такие вероятности зависят только от одного предыдущего символа, т. е.

$$P\left\{x_k^{(b)} / x^{(b-1)}, x^{(b-2)}, \ldots\right\} = P\left\{x_k^{(b)} / x_m^{(b-1)}\right\}, \qquad (2.2.1)$$

то такой источник называется *марковским источником первого порядка*[1] или, просто *марковским источником*. В дальнейшем о таких источниках и пойдёт речь.

Для марковского источника генерирование на некотором b-м шаге очередного символа x_i означает переход в состояние $S_q^{(b)} = i$, определяемое парой $\left(x_i^{(b)}, x_j^{(b-1)}\right)$, так что последовательности символов $\left(x_{j_1}^{(1)}, x_{j_2}^{(2)}, \ldots, x_{j_n}^{(n)}\right)$ соответствует цепочка состояний $j_1 \to j_2 \to \ldots \to j_n$.

Обозначим через $p_{ji}(1)$ условную вероятность *одношагового* перехода $j \to i$ на некотором b-м шаге в определённое состояние i при условии, что известно предыдущее состояние j:

$$p_{ji}(1) \equiv p_{ji} = P\left\{x_i^{(b)} / x_j^{(b-1)}\right\} = P\left\{S^{(b)} = i / S^{(b-1)} = j\right\}. \qquad (2.2.2)$$

Тогда полное поведение марковского источника можно задать посредством квадратной матрицы $\boldsymbol{\pi}$ размерности $q \times q$:

$$\boldsymbol{\pi} = \begin{pmatrix} p_{11} & p_{21} & \cdots & p_{q1} \\ p_{12} & p_{22} & \cdots & p_{q2} \\ \cdots & \cdots & \cdots & \cdots \\ p_{1q} & p_{2q} & \cdots & p_{qq} \end{pmatrix}, \qquad (2.2.3)$$

называемой *матрицей переходных вероятностей*. Такая матрица — стохастическая, т. е. сумма её элементов в каждой строке есть 1. Если элементы матрицы переходных вероятностей не зависят от номера шага, на котором происходит переход из одного состояния в другое, то такой источник называется *однородным*, в противном случае — *неоднородным*.

[1] Заметим, что при такой терминологии источник без памяти, генерирующий элементы независимым образом, может называться марковским источником нулевого порядка.

Иногда бывает удобно нумеровать состояния, начиная с нуля, а не с единицы. В этом случае индексы матрицы $\boldsymbol{\pi}$ следует уменьшить на единицу.

Вероятность $p_{ji}(2)$ *двухшагового* перехода $j \to i$ через произвольное промежуточное состояние $S = k$, т. е. вероятность последовательности одношаговых переходов $j \to k \to i$ на основании формулы полной вероятности равна

$$p_{ji}(2) = P\left\{S^{(b)} = i / S^{(b-1)} = k, S^{(b-2)} = j\right\} = \sum_{k=1}^{q} p_{jk} p_{ki} ,$$

и эти вероятности представляют собой матричные элементы для матрицы, которая является произведением матрицы $\boldsymbol{\pi}$ на саму себя. Таким образом, если обозначить через $\boldsymbol{\pi}_2$ матрицу двухшаговых переходов, то

$$\boldsymbol{\pi}_2 = \boldsymbol{\pi} \cdot \boldsymbol{\pi} = \boldsymbol{\pi}^2 .$$

Рассуждая аналогичным образом, можно прийти к выводу, что вероятность $p_{ji}(n)$ для *n-шагового* перехода $j \to i$ равна

$$p_{ji}(n) = \sum_{k=1}^{q} p_{jk}(n_1) p_{ki}(n_2), \; n_1 + n_2 = n, \qquad (2.2.4a)$$

т. е. матрица n-шаговых переходов $\boldsymbol{\pi}_n$ представляет собой произведение степеней матрицы $\boldsymbol{\pi}$ так, чтобы сумма этих степеней составляла n:

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}^{n_1} \cdot \boldsymbol{\pi}^{n_2} = \boldsymbol{\pi}^n , \; n_1 + n_2 = n . \qquad (2.2.4б)$$

Задание матрицы переходных вероятностей (2.2.3) позволяет полностью описывать изменения состояния источника за один шаг и, кроме того, определять вероятности заданных *траекторий*, т. е. последовательностей состояний, которые проходит источник. Для этого, прежде всего, необходимо задать вероятность начального состояния $p_{0i} = P\left\{S^{(0)} = i\right\} = P\left(x_i^{(0)}\right)$. Тогда вероятность траектории

$$i \to j_1 \to j_2 \to \cdots \to j_n$$

с учётом марковского свойства равна

$$P\left\{i \to j_1 \to j_2 \to \cdots \to j_n\right\} \equiv P\left\{\left(x_i^{(0)}, x_{j_1}^{(1)}, x_{j_2}^{(2)}, \ldots, x_{j_n}^{(n)}\right)\right\} =$$

$$= P\left\{x_{j_n}^{(n)} \Big/ x_{j_{n-1}}^{(n-1)}, \ldots, x_{j_1}^{(1)}, x_i^{(0)}\right\} P\left\{x_{j_{n-1}}^{(n-1)}, \ldots, x_{j_1}^{(1)}, x_i^{(0)}\right\} =$$

$$= P\left\{x_{j_n}^{(n)} \Big/ x_{j_{n-1}}^{(n-1)}\right\} P\left\{x_{j_{n-1}}^{(n-1)}, \ldots, x_{j_1}^{(1)}, x_i^{(0)}\right\} = \ldots =$$

$$= P\left\{x_{j_n}^{(n)} \Big/ x_{j_{n-1}}^{(n-1)}\right\} P\left\{x_{j_{n-1}}^{(n-1)} \Big/ x_{j_{n-2}}^{(n-2)}\right\} \cdots P\left\{x_{j_1}^{(1)} \Big/ x_i^{(0)}\right\} p_{0i} =$$

$$= p_{j_{n-1}j_n} p_{j_{n-2}j_{n-1}} \cdots p_{j_1 i} p_{0i}. \qquad (2.2.5)$$

Обобщением простой цепи Маркова является *марковская цепь n-го порядка* — процесс, вероятностное описание которого предполагает учёт $n > 1$ предыдущих состояний (см. разд. 1.3).

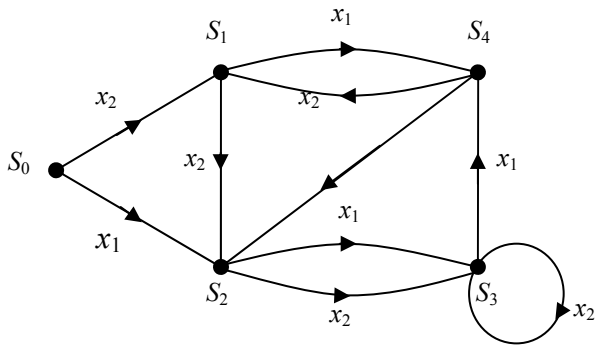


Рис. 2.1. Марковский источник с непериодическими состояниями

Функционирование марковского источника можно проиллюстрировать схемами, подобными той, что показана на рис. 2.1. В данном примере каждый узел схемы представляет собой одно из четырёх состояний S_1, …, S_4, а каждая ветвь указывает (согласно стрелке) изменение состояния, произошедшее в результате генерирования очередного символа: либо x_1, либо x_2. При этом необходимо задать начальное состояние S_0, в котором находится источник, перед тем, как сгенерировать самый первый символ.

Графическое представление марковского источника удобно тем, что, задавая для каждого состояния условные вероятности $P(x_k / S_q)$ появления символов, можно вычислить переходные вероятности из одного состояния в другое $P(S_q / S_r)$ путём простого суммирования вероятностей появления символов, которые приводят к такому переходу. Так, для рассматриваемого примера

$$P(S_4 / S_1) = P(x_1 / S_1);$$
$$P(S_3 / S_2) = P(x_1 / S_2) + P(x_2 / S_2) = 1.$$

Матрица условных вероятностей полностью описывает источник, а следовательно, и весь ансамбль сообщений, образованный генерированием последовательностей символов.

Свойства марковских последовательностей давно и широко исследованы [9], и далее приведён краткий обзор некоторых основных понятий

и результатов, необходимых для понимания общих теоретико-информационных свойств.

Состояние S_q называется *несущественным*, если существует такое состояние S_r и натуральное число n_0 (шаг генерирования), для которых отлична от нуля вероятность $p_{qr}(n_0)$, в то время как для любых натуральных n "обратная" вероятность $p_{rq}(n) = 0$. В противном случае такое состояние S_q называется *существенным*.

Существенные состояния S_q и S_r называются *сообщающимися*, если существуют такие натуральные значения t и s, что $p_{qr}(t) > 0$ и одновременно $p_{rq}(s) > 0$.

Например, для источника, который может находиться в одном из четырёх состояний $S_1 \ldots S_4$ и описывается матрицей переходных вероятностей

$$\boldsymbol{\pi} = \begin{pmatrix} 0 & p & q & 0 \\ p & 0 & 0 & q \\ 0 & 0 & p & q \\ 0 & 0 & p & q \end{pmatrix} = \begin{pmatrix} P(S_1/S_1) & P(S_1/S_2) & P(S_1/S_3) & P(S_1/S_4) \\ P(S_2/S_1) & P(S_2/S_2) & P(S_2/S_3) & P(S_2/S_4) \\ P(S_3/S_1) & P(S_3/S_2) & P(S_3/S_3) & P(S_3/S_4) \\ P(S_4/S_1) & P(S_4/S_2) & P(S_4/S_3) & P(S_4/S_4) \end{pmatrix},$$

где $p > 0$, $q > 0$ и $p + q = 1$, состояния S_1 и S_2 являются несущественными, в то время как состояния S_2 и S_3 — существенные и сообщающиеся.

Состояние S_q называется *возвратным*, если вероятность возвращения в него равна единице; в противном случае такое состояние называется *невозвратным*. В примере на рис. 2.1 все состояния кроме начального S_0 являются возвратными.

Состояние S_q называется *нулевым*, если вероятность $p_{qq}(n)$ стремится к нулю при $n \to \infty$ и *ненулевым* в противном случае.

Состояние S_q называется *периодическим* с (целочисленным) периодом $T_q > 1$, если возвращение в него возможно только после $m = kT_q$ шагов, где k — произвольное натуральное число. Иными словами, для периодического состояния из условия $p_{qq}(n) > 0$ следует, что n кратно T_q, или, что эквивалентно, T_q является наибольшим общим делителем множества тех n, для которых отлична от нуля вероятность

$$f_q(n) = P\left(S^{(n)} = S_q, S^{(n-1)} \neq S_q, \ \ldots \ , S^{(1)} \neq S_q \,\middle/\, S^{(0)} = S_q\right)$$

того, что источник, выйдя из состояния S_q, впервые вернётся в него ровно через n шагов.

Суммарная вероятность

$$F_q = \sum_{n=1}^{\infty} f_q(n)$$

определяет событие, состоящее в том, что источник, выйдя из состояния S_q, когда-либо вернётся в это состояние.

На рис. 2.1 все состояния непериодические, в то время как для другого источника, представленного на рис. 2.2, наоборот, все состояния являются периодическими с периодом, равным 2.

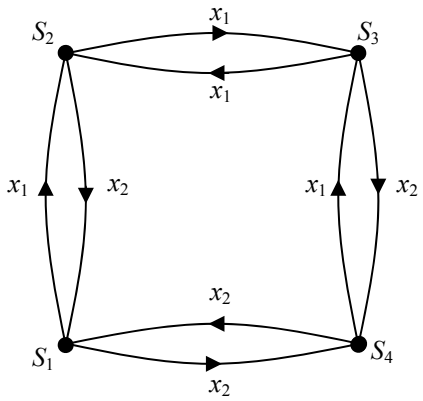


Рис. 2.2. Марковский источник с периодическими состояниями

Множество состояний называется *замкнутым*, если одношаговые переходы из любого состояния не выводят за пределы этого множества. При этом марковская цепь называется *неразложимой*, если любое подмножество её состояний, отличное от множества всех состояний, незамкнуто; в противном случае цепь называется *разложимой*.

Цепь, показанная на рис. 2.1, разложима, поскольку подмножество, образованное всеми состояниями, кроме начального S_0, является замкнутым. Однако если исключить начальное состояние, то оставшиеся состояния образуют неразложимую цепь.

Очевидно, как только источник попадает в одно из замкнутых подмножеств, все другие подмножества становятся ему недоступны. Исходя из этого, изучение свойств цепи можно проводить, рассматривая независимо замкнутые подмножества. Более строго указанное свойство можно сформулировать следующим образом.

Не существует марковской цепи с конечным числом состояний, *состоящей только из невозвратных состояний*. *Возвратные состояния*,

имеющиеся в цепи, можно разбить на замкнутые множества, соответствующие неразложимым цепям, в которые можно попасть из невозвратных состояний, но не наоборот.

Неразложимая цепь (*подцепь*) *с конечным числом состояний может содержать только возвратные состояния. Если одно из её состояний периодическое, то все состояния также являются периодическими с тем же периодом.*

Одной из основных задач исследования марковского источника является поиск безусловных вероятностей того, что в произвольный (дискретный) момент времени источник окажется в некотором заданном состоянии.

Рассмотрим непериодический неразложимый марковский источник с конечным числом состояний, в котором $P(S_q / S_r)$ — вероятность попадания из состояния S_r в состояние S_q ровно за m шагов. Тогда справедливо следующее свойство.

Для всех пар состояний (S_q, S_r) *существуют предельные* (*финальные*) *вероятности*

$$\lim_{m\to\infty} P_m\left(S_q / S_r\right) = P(S_q), \tag{2.2.6}$$

удовлетворяющие системе уравнений

$$P(S_q) = \sum_{(i)} P\left(S_q / S_i\right) P(S_i), \tag{2.2.7}$$

где суммирование ведётся по всем состояниям источника.

Заметим, что определитель системы линейных уравнений (2.2.7) обращается в нуль, поскольку

$$\sum_{(i)} P\left(S_q / S_i\right) = 1, \tag{2.2.8}$$

и, следовательно, финальные вероятности $P(S_q)$ определяются с точностью до постоянного множителя. Разрешить эту неопределённость можно, воспользовавшись условием нормировки для состояний

$$\sum_{(q)} P\left(S_q\right) = 1. \tag{2.2.9}$$

Сформулированное свойство означает то, что вероятность нахождения источника в каком-либо состоянии S_q стремится к предельному значению, когда число последовательных переходов между состояниями неограниченно возрастает. Более того, этот предел — единственный, и он не

зависит от начального состояния источника. Таким образом, порождаемые источником последовательности символов попадают в пределе в условия статистического равновесия, и если вероятность нахождения в каком-либо состоянии S_q равна значению $P(S_q)$, удовлетворяющему системе (2.2.7), то эти условия статистического равновесия существуют с самого начала.

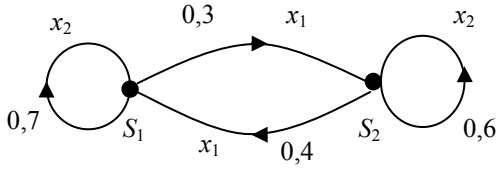
Иными словами, при некоторых условиях в марковском источнике с течением (дискретного) времени устанавливается стационарный режим, при котором источник продолжает “блуждать” по своим состояниям, но вероятности этих состояний от времени уже не зависят. Такие состояния и называются стационарными (финальными, предельными).

Установить справедливость (2.2.7) для конечного непериодического неразложимого марковского источника достаточно просто. Пусть в произвольный m-й момент времени источник, стартуя из начального состояния S_r ($r = 1, \ldots, L$), оказался в состоянии S_i ($i = 1, \ldots, L$), причём вероятность такого события равна $P_m(S_i / S_r) \equiv p_{ri}(m)$. Тогда на следующем шаге источник с вероятностью p_{iq} может перейти в одно из состояний S_q ($q = 1, \ldots, L$), и вероятность этого события — того, что источник, стартуя из начального состояния S_r за $m + 1$ шаг окажется в состоянии S_q — согласно формуле полной вероятности равна

$$p_{rq}(m+1) \equiv P_{m+1}\left(S_q / S_r\right) = \sum_{i=1}^{L} P_m\left(S_i / S_r\right) p_{iq} \, .$$

Теперь, предполагая существование у источника предельных вероятностей $\lim\limits_{m\to\infty} P_m\left(S_q / S_r\right) = P(S_q)$ и переходя в последнем выражении к предельному переходу $m \to \infty$, получаем искомое соотношение (2.2.7).

Рассмотрим в качестве примера простейший непериодический неразложимый двоичный источник, графическое представление которого показано на рис. 2.3. Источник может находиться в двух состояниях. В состоянии S_1 с вероятностью 0,7 генерируется символ x_2, после чего источник остаётся в прежнем состоянии, и с вероятностью 0,3 генерируется символ x_1, после чего источник переходит в состояние S_2. В состоянии S_2 свойства источника аналогичны: генерирование с вероятностью 0,6 символа x_2, оставаясь в прежнем состоянии, и генерирование с вероятностью 0,4 символа x_1 с последующим переводом в состояние S_1.



$$\pi = \begin{pmatrix} 0,6 & 0,4 \\ 0,3 & 0,7 \end{pmatrix}$$

Рис. 2.3. Пример марковского источника

Решая для такого источника систему (2.2.7), легко получить финальные вероятности состояний, которые оказываются равными

$$P(S_1) = \frac{0,4}{0,7} = 0,571 \text{ и } P(S_2) = \frac{0,3}{0,7} = 0,429.$$

Проследим теперь механизм перехода в стационарный режим, считая, что начальным состоянием является состояние S_1, т. е. $p_{01} = 1$, $p_{02} = 0$. Последовательно имеем

на первом шаге:

$$P_1(S_1) = P(S_1 / S_1)p_{01} + P(S_1 / S_2)p_{02} = 0,7 \cdot 1 + 0,4 \cdot 0 = 0,700$$

$$P_1(S_2) = 1 - P_1(S_1) = 1 - 0,700 = 0,300;$$

на втором шаге:

$$P_2(S_1) = P(S_1 / S_1)P_1(S_1) + P(S_1 / S_2)P_1(S_2) = 0,7 \cdot 0,7 + 0,4 \cdot 0,3 = 0,610$$

$$P_2(S_2) = 1 - P_2(S_1) = 1 - 0,610 = 0,390;$$

на третьем шаге:

$$P_3(S_1) = P(S_1 / S_1)P_2(S_1) + P(S_1 / S_2)P_2(S_2) = 0,7 \cdot 0,61 + 0,4 \cdot 0,39 = 0,583;$$

$$P_3(S_2) = 1 - P_3(S_1) = 1 - 0,583 = 0,417;$$

на четвёртом шаге:

$$P_4(S_1) = P(S_1 / S_1)P_3(S_1) + P(S_1 / S_2)P_3(S_2) = 0,7 \cdot 0,583 + 0,4 \cdot 0,417 = 0,575;$$

$$P_4(S_2) = 1 - P_4(S_1) = 1 - 0,575 = 0,425.$$

Таким образом, в рассматриваемом примере уже на четвёртом шаге безусловные вероятности отличаются от финальных лишь в третьем знаке, т. е. практически можно считать, что стационарный режим наступает после первых четырёх шагов.

Кроме того, проведя аналогичные выкладки, можно убедиться в том, что в данном случае финальные вероятности не зависят от начальных условий (начальных состояний).

Рассмотрим разложимый марковский источник, состоящий из двух неразложимых подцепей C_1 и C_2. Соотношение (2.2.6) применимо к любой паре (S_q, S_r) состояний, принадлежащей любой из подцепей, но неприменимо, если эти состояния находятся в разных подцепях, поскольку в этом случае, очевидно, $P(S_q / S_r) = 0$. Тогда выражение для финальных вероятностей несколько видоизменяется, принимая следующий вид

$$\lim_{m\to\infty} P_m\left(S_q / S_r\right) = \begin{cases} P\left(S_q / C_1\right), & S_r \in C_1 \\ P\left(S_q / C_2\right), & S_r \in C_2 \end{cases}, \qquad (2.2.10)$$

При этом вероятность $P(S_q / C_1)$ равна нулю, если S_q принадлежит подцепи C_2 и, соответственно, вероятность $P(S_q / C_2)$ равна нулю, если S_q принадлежит подцепи C_1. В этом случае суммирование в (2.2.7) может вестись только по состояниям той подцепи, к которой принадлежит S_q. Тогда, если $P(C_1)$ и $P(C_2)$ — вероятности того, что начальное состояние принадлежит к подцепям C_1 и C_2 соответственно, то финальные вероятности равны

$$P\left(S_q\right) = \begin{cases} P\left(S_q / C_1\right)P(C_1), & S_r \in C_1 \\ P\left(S_q / C_2\right)P(C_2), & S_r \in C_2 \end{cases}. \qquad (2.2.11)$$

Кроме того, если финальные вероятности $P(S_q)$, задаваемые (2.2.11), равны вероятностям начальных состояний, то вероятность нахождения источника в каком-либо состоянии не зависит от времени, также не зависят от времени все статистические характеристики генерируемых последовательностей, и источник с самого начала находится в стационарном режиме.

Пусть теперь разложимый марковский источник с конечным числом состояний обладает как невозвратными состояниями, так и несколькими неразложимыми подцепями. Тогда для такого источника справедливы следующие свойства.

Источник, находящийся в невозвратном состоянии, в конечном итоге с единичной вероятностью окажется в возвратном состоянии. Если он содержит более одной неразложимой подцепи, то условная вероятность $P(C_j / S_q)$ того, что из невозвратного состояния S_q будет достигнуто какое-то состояние, принадлежащее неразложимой подцепи C_j, является решением системы уравнений

$$P\left(C_j / S_q\right) - \sum_{(r)} P\left(C_j / S_r\right)P\left(S_r / S_q\right) = P_1\left(C_j / S_q\right). \qquad (2.2.12).$$

В выражении (2.2.12) *суммирование ведётся по всем невозвратным состояниям, а* $P_1(C_j / S_q)$ *— вероятность того, что переход из состояния* S_q *в подцепь* C_j *осуществится за один шаг.*

Попав однажды в какое-либо возвратное состояние, источник больше не может оставить ту подцепь, к которой принадлежит это возвратное состояние.

Понятие финальных вероятностей позволяет говорить об асимптотическом поведении энтропии непериодического марковского источника с конечным числом состояний. Для неразложимого источника, находящегося в состоянии S_q, соответствующая условная энтропия равна

$$H(X / S_q) = -\sum_{(k)} P(x_k / S_q) \log P(x_k / S_q),$$

где, как обычно, суммирование ведётся по всем возможным символам источника.

Тогда, в соответствии с выражением (1.3.4), можно ввести в рассмотрение асимптотическое значение энтропии $H(X / X^{\infty})$, когда длина генерируемых последовательностей символов стремится к бесконечности, путём усреднения условной энтропии $H(X / S_q)$ по всем возможным состояниям:

$$H\left(X / X^{\infty}\right) = \sum_{(q)} H(X / S_q) P(S_q) =$$

$$= -\sum_{(q)} \sum_{(k)} P(x_k / S_q) P(S_q) \log P(x_k / S_q). \qquad (2.2.13)$$

Если источник является разложимым, то асимптотическая энтропия может быть получена усреднением (2.2.13) по всем неразложимым подцепям. При этом невозвратные состояния могут не приниматься во внимание, надо лишь учитывать, что они определяют вероятность попадания источника в состояние некоторой подцепи.

Если S_0 — начальное состояние источника, то с помощью (2.2.12) можно вычислить вероятность $P(C_j / S_0)$ того, что источник в конечном итоге достигнет состояний, принадлежащих подцепи C_j. Таким образом, асимптотическое значение энтропии разложимого источника определяется следующим выражением:

$$H\left(X / X^{\infty}\right) = \sum_{j=1}^{N} P\left(C_j / S_0\right) \sum_{q=1}^{R_j} P\left(S_q / C_j\right) H\left(X / S_q\right), \quad (2.2.14)$$

где R_j — число состояний подцепи C_j, N — общее число подцепей, а условная вероятность $P(S_q / C_j)$ определяет нахождение источника в состоянии S_q в предположении, что состояния, принадлежащие той же подцепи, уже были достигнуты.

Обратимся теперь к периодическим марковским источникам. Неразложимые периодические марковские источники обладают свойствами, аналогичными (2.2.6).

Если T — период неразложимого периодического марковского источника с конечным числом состояний, то эти состояния можно разбить на T подмножеств $X_1, X_2, \ldots, X_T$, так, что одношаговый переход переводит источник из одного подмножества в другое. Пусть i — целое положительное число и $P_{iT}(S_q / S_r)$ — вероятность того, что источник достигнет состояния S_q из состояния S_r за iT шагов. Тогда

$$\lim_{i\to\infty} P_{iT}\left(S_q / S_r\right) = P(S_q / X_j),\ j = 1,\ \ldots,\ T, \qquad (2.2.15)$$

если S_q и S_r принадлежат одному подмножеству X_j и

$$\lim_{i\to\infty} P_{iT}\left(S_q / S_r\right) = 0, \qquad (2.2.16)$$

если S_q и S_r принадлежат разным подмножествам.

Вероятности $P(S_q / X_j)$ удовлетворяют системе уравнений

$$P(S_q / X_j) = \sum_{(r)} P(S_r / X_j) P_T(S_q / S_r), \qquad (2.2.17)$$

где суммирование ведётся по всем возможным состояниям в подмножестве X_j.

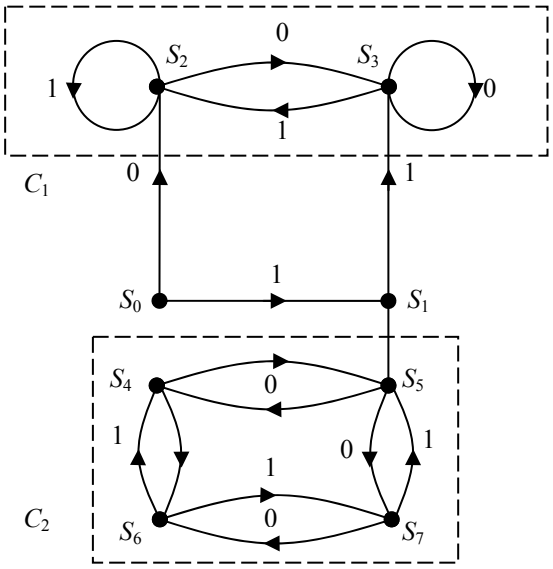
По сути, вероятности $P(S_q / X_j)$ представляют собой асимптотическое значение условной вероятности того, что источник, находясь в некотором состоянии подмножества X_j, через iT одношаговых переходов окажется в определённом состоянии S_q того же подмножества. Тогда, если $P_0(X_\gamma)$ — вероятность того, что начальное состояние принадлежит некоторому подмножеству X_γ, то финальная вероятность нахождения источника в состоянии $S_q \in X_j$ после $m = iT + \tau$ одношаговых переходов (τ — целое число, $0 \le \tau < T$) задаётся соотношением

$$\lim_{i\to\infty} P_{iT+\tau}\left(S_q\right) = P_0(X_\gamma) P(S_q / X_j), \qquad (2.2.18)$$

где

$$\gamma = \begin{cases} j - \tau, & j \geq \tau; \\ T + j - \tau, & j < \tau. \end{cases}$$

Видно, что и в данном случае последовательности состояний, принимаемых источником, и, следовательно, последовательности генерируемых символов приближаются к состоянию статистического равновесия, но имеющего дополнительный “периодический” оттенок.



Состояние	$P(x_k/S_q)$	
	$x_k = 0$	$x_k = 1$
S_0	0,2	0,8
S_1	0,5	0,5
S_2	0,7	0,3
S_3	0,6	0,4
S_4	0,3	0,7
S_5	0,6	0,4
S_6	0,9	0,1
S_7	0,2	0,8

Рис. 2.4. Пример марковского источника

Для неразложимых периодических марковских источников с конечным числом состояний асимптотическая энтропия $\lim\limits_{i\to\infty} H\left(X^{(m+1)} / X^m\right)$ суть периодическая функция от числа $m = iT + \tau$ ранее сгенерированных символов. Усредняя (2.2.15) по всем подмножествам $X_1, X_2, \ldots, X_T$, получаем следующее выражение:

$$\lim_{m\to\infty} H\left(X^{(m+1)} / X^m\right) = \sum_{\gamma=1}^{T} P_0(X_\gamma) \sum_{(r)} P(S_r / X_j) H(X / S_r). \qquad (2.2.19)$$

Для пояснения изложенных выше свойств марковских источников рассмотрим пример, представленный на рис. 2.4, где графически показаны возможные переходы и состояния двоичного источника, а в расположенной справа таблице приведены значения условных вероятностей $P(x_k / S_q)$.

Пытаясь дать физическую трактовку поведения рассматриваемого источника, представим, что S_0 является начальным состоянием. Далее, из

начального состояния источник переходит в промежуточное невозвратное состояние S_1 (некий переходный процесс), из которого затем — в один из двух статистически устойчивых процессов, каждый их которых связан с двумя подмножествами $X_1 = \{S_2, S_3\}$ и $X_2 = \{S_4, S_5, S_6, S_7\}$. Функционированию источника в рамках указанных подмножеств уже можно придать смысл генерирования символов с определёнными статистическими связями.

Таким образом, исходную цепь можно разложить на две неразложимых подцепи: непериодическую C_1 и периодическую C_2 — ту, что содержит периодические с периодом $T = 2$ состояния $S_4, \ldots, S_7$.

Рассматривая сначала непериодическую подцепь C_1, запишем систему уравнений для финальных вероятностей:

$$P(S_2) = P(S_2 / S_2)P(S_2) + P(S_2 / S_3)P(S_3);$$
$$P(S_3) = P(S_3 / S_2)P(S_3) + P(S_3 / S_3)P(S_3),$$

или, подставляя численные значения,

$$P(S_2) = 0,300P(S_2) + 0,400P(S_3);$$
$$P(S_3) = 0,700P(S_2) + 0,600P(S_3),$$

из которой следует соотношение

$$0,700P(S_2) = 0,400P(S_3).$$

Используя условие нормировки

$$P(S_2) + P(S_3) = 1,$$

получим вероятности двух состояний, образующих неразложимую непериодическую подцепь C_1:

$$P(S_2) = 0,364,\ P(S_3) = 0,636.$$

Далее, обратимся к периодической подцепи C_2, состояния которой можно разбить на два подмножества:

$$X_2 = X_{21} \cup X_{22} = \{S_4, S_7\} \cup \{S_5, S_6\},$$

исходя из того, что одношаговые переходы приводят из состояний одного подмножества в состояния другого подмножества. Используя таблицу на рис. 2.4, для всех условных вероятностей $P(S_q / S_r)$ данной подцепи имеем (нижний индекс T, равный в данном случае 2, для упрощения записи опустим):

$$P(S_4 / S_4) = 0,3 \cdot 0,6 + 0,7 \cdot 0,9 = 0,810;$$
$$P(S_7 / S_4) = 0,3 \cdot 0,4 + 0,7 \cdot 0,1 = 0,190;$$
$$P(S_7 / S_7) = 0,2 \cdot 0,4 + 0,8 \cdot 0,1 = 0,160;$$

$$P(S_4 / S_7) = 0,2 \cdot 0,6 + 0,8 \cdot 0,9 = 0,840;$$
$$P(S_5 / S_5) = 0,6 \cdot 0,3 + 0,4 \cdot 0,2 = 0,260;$$
$$P(S_6 / S_5) = 0,6 \cdot 0,7 + 0,4 \cdot 0,8 = 0,740;$$
$$P(S_6 / S_6) = 0,9 \cdot 0,7 + 0,1 \cdot 0,8 = 0,710;$$
$$P(S_5 / S_6) = 0,9 \cdot 0,3 + 0,1 \cdot 0,2 = 0,290.$$

Подставляя полученные значения вероятностей в (2.2.17), получим две системы уравнений:

$$P(S_4 / X_{21}) = 0,810P(S_4 / X_{21}) + 0,840P(S_7 / X_{21});$$
$$P(S_7 / X_{21}) = 0,190P(S_4 / X_{21}) + 0,160P(S_7 / X_{21});$$

и

$$P(S_5 / X_{22}) = 0,260P(S_5 / X_{22}) + 0,290P(S_6 / X_{22});$$
$$P(S_6 / X_{22}) = 0,740P(S_5 / X_{22}) + 0,710P(S_6 / X_{22}).$$

При этом имеют место условия

$$P(S_4 / X_{21}) + P(S_7 / X_{21}) = 1;$$
$$P(S_5 / X_{21}) + P(S_6 / X_{21}) = 1.$$

Решая выписанные уравнения, получим следующие значения финальных вероятностей состояний для периодической подцепи:

$$P(S_4 / X_{21}) = 0,816,\ P(S_7 / X_{21}) = 0,184,$$
$$P(S_5 / X_{22}) = 0,282,\ P(S_6 / X_{22}) = 0,718.$$

Чтобы найти асимптотическую энтропию источника необходимо подсчитать вероятности $P(C_1 / S_0)$ и $P(C_2 / S_0)$ того, что источник, выйдя из начального состояния S_0, в конечном итоге достигнет состояний подцепи C_1 или C_2. Поскольку переход в подцепь C_2 происходит только через состояние S_1, то согласно таблице на рис. 2.4,

$$P\left(C_2 / S_0\right) = P\left(S_1 / S_0\right) P\left(S_5 / S_1\right) = 0,8 \cdot 0,5 = 0,400$$

и, в силу наличия только двух подцепей,

$$P\left(C_1 / S_0\right) = 1 - P\left(C_2 / S_0\right) = 0,600.$$

Кроме того, заметим, что начальным состоянием подцепи C_2 может быть только S_5, так что вероятности того, что начальное состояние принадлежит подмножествам X_{21} и X_{22} равны соответственно

$$P_0\left(X_{21}\right) = P_0\left(\left\{S_4, S_7\right\}\right) = 0;$$

$$P_0\left(X_{22}\right)=P_0\left(\{S_5,S_6\}\right)=1.$$

Теперь можно вычислить асимптотическое значение энтропии. Для непериодической подцепи C_1 имеем:

$$H_1\left(X/X^{\infty}\right)=-P\left(S_2\right)\left[H\left(X/S_2\right)\right]-P\left(S_3\right)\left[H\left(X/S_3\right)\right]=$$
$$=-0,364\left(0,700\cdot\log_2 0,700+0,300\cdot\log_2 0,300\right)-$$
$$-0,636\left(0,600\cdot\log_2 0,600+0,400\cdot\log_2 0,400\right)=0,935 \text{ бит.}$$

Для непериодической подцепи C_2 подсчитаем сначала отдельно значения энтропии для подмножеств X_{21} и X_{22}:

$$H_2\left(X/X^{\infty},X_{21}\right)=-P\left(S_4/X_{21}\right)\left[H\left(X/S_4\right)\right]-P\left(S_7/X_{21}\right)\left[H\left(X/S_7\right)\right]=$$
$$=-0,806\left(0,300\cdot\log_2 0,300+0,700\cdot\log_2 0,700\right)-$$
$$-0,184\left(0,200\cdot\log_2 0,200+0,800\cdot\log_2 0,800\right)=0,843 \text{ бит;}$$
$$H_2\left(X/X^{\infty},X_{22}\right)=-P\left(S_5/X_{22}\right)\left[H\left(X/S_5\right)\right]-P\left(S_6/X_{22}\right)\left[H\left(X/S_6\right)\right]=$$
$$=-0,282\left(0,600\cdot\log_2 0,600+0,400\cdot\log_2 0,400\right)-$$
$$-0,718\left(0,900\cdot\log_2 0,900+0,100\cdot\log_2 0,100\right)=0,611 \text{ бит,}$$

и эти асимптотические значения подцепь C_2 принимает попеременно, поскольку $P_0\left(X_{21}\right)=0$ и $P_0\left(X_{22}\right)=1$.

Итак, можно сказать, что асимптотическое значение энтропии на одну букву сообщения состоит из двух компонент: непериодической компоненты, равной 0,935 бит, и периодической компоненты с чередующимися значениями 0,843 и 0,611 бит.

2.3. ЭРГОДИЧЕСКИЕ ИСТОЧНИКИ

Невосприимчивость к влиянию начального состояния для неразложимого источника является проявлением свойства *эргодичности*. Точная формулировка условий эргодичности достаточно громоздка и приведение их в рамках излагаемого материала вряд ли целесообразно. Эвристически эргодичность означает, что свойства почти каждой порождаемой источником последовательности совпадают со свойствами всего ансамбля последовательностей. Для более строгого определения свойства эргодичности необходимо различать два понятия среднего значения: *среднее по ансамблю* и *среднее по последовательности*.

Представим, что в наличии имеются M стационарных источников, описываемых одним и тем же распределением вероятностей, а сообщения, возникающие на их выходах, независимо разбиваются на N-элементные последовательности символов

$$\mathbf{x}_N = \left(x^{(1)}, x^{(2)}, \ \ldots, \ x^{(N)} \right).$$

Обозначим через $I^{(m)}[\mathbf{x}_N]$ ($m = 1, \ \ldots, \ M$) количество информации, возникающей при генерировании m-м источником последовательности $\mathbf{x}_N$. Это количество информации является случайной величиной, определённой на ансамбле всех возможных сообщений. Её математическое ожидание, т. е. *среднее по ансамблю* по определению есть энтропия $H_M(X^N)$ ансамбля N-элементных сообщений, и с учётом идентичности и независимости источников она равна

$$H_M\left(X^N\right) = \frac{1}{M} \sum_{m=1}^{M} I^{(m)}\left[\mathbf{x}_N\right]. \qquad (2.3.1)$$

Рассмотрим теперь один стационарный источник, генерирующий на выходе сообщения, представляющие собой N-элементные подпоследовательности символов

$$\mathbf{x}_N^{(j)} = \left(x^{(j+1)}, x^{(j+2)}, \ldots, x^{(j+N)} \right), \ j = 0, \ 1, \ \ldots$$

которым соответствуют значения количества информации $I\left[\mathbf{x}_N^{(j)}\right]$. Если в ν-элементной последовательности содержится $\mu = \nu / N$ образцов N-элементных подпоследовательностей $\mathbf{x}_N^{(j)}$, то среднее по наблюдаемой последовательности значение количества информации равно

$$H_\mu\left(X^N\right) = \frac{1}{\mu} \sum_{j=0}^{\mu-1} I\left[\mathbf{x}_N^{(j)}\right]. \qquad (2.3.2)$$

Теперь можно определить эргодический источник как источник, для которого усреднение по последовательности значений любой случайной величины, связанной с состоянием эргодического источника, с единичной вероятностью равно среднему значению той же величины усреднённой по ансамблю реализаций:

$$P\left\{ \lim_{M\to\infty} \frac{1}{M} \sum_{m=1}^{M} I^{(m)}\left[\mathbf{x}_N\right] = \lim_{\mu\to\infty} \frac{1}{\mu} \sum_{j=0}^{\mu-1} I\left[\mathbf{x}_N^{(j)}\right] \right\} = 1. \qquad (2.3.3)$$

Разумеется, практический смысл соотношение (2.3.3) имеет для конечных, но достаточно больших значений M и μ.

Лучше всего отсутствие свойства эргодичности проиллюстрировать на следующем примере. Пусть простой стационарный марковский источник X равновероятно генерирует один из четырёх символов x_1, x_2, x_3, x_4, т. е. $p(x_k) = 1/4$ ($k = 1, \ldots, 4$), а условное распределение вероятностей $p(x_k / x_p)$ ($k, p = 1, \ldots, 4$) задано матрицей

$$\pi = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/4 & 3/4 \\ 0 & 0 & 3/4 & 1/4 \end{pmatrix}.$$

Будем рассматривать последовательность символов на выходе такого источника как последовательность укрупнённых символов, каждый из которых состоит из двух подряд идущих символов x_p и x_k. Вероятности совместных событий $p(x_k, x_p)$ равны

$$p(x_k, x_p) = p(x_k / x_p)\, p(x_p),$$

поэтому имеют место следующие ненулевые значения количества информации, соответствующие сочетаниям различных пар символов:

$$I(x_1, x_1) = I(x_1, x_2) = I(x_2, x_1) = I(x_2, x_2) = -\log(1/8) = 3{,}000 \text{ бит};$$

$$I(x_3, x_3) = I(x_4, x_4) = -\log(1/16) = 4{,}000 \text{ бит};$$

$$I(x_3, x_4) = I(x_4, x_3) = -\log(3/16) = 2{,}415 \text{ бит}.$$

Тогда энтропия ансамбля сообщений $\{p(x_k / x_p)$ ($k, p = 1, \ldots, 4)\}$ равна

$$H\left[(x_k, x_p)\right] = \frac{4}{8} \cdot 3 + \frac{2}{16} \cdot 4 + \frac{2 \cdot 3}{16} \cdot 2{,}415 = 2{,}906 \text{ бит}.$$

Анализ поведения источника показывает, что возможны два типа подпоследовательностей: подпоследовательность, состоящая из символов x_1 и x_2, и подпоследовательность, состоящая из символов x_3 и x_4. Эти подпоследовательности имеют одинаковые вероятности, и можно сказать, что рассматриваемый источник фактически состоит из двух различных источников $X_1 = \{x_1, x_2\}$ и $X_2 = \{x_3, x_4\}$, равновероятно подключаемых к выходу.

Если включён источник X_1, то возможно только одно значение количества информации, равное 3,000. Если включён источник X_2, то возможны

два значения количества информации: 4,000 и 2,415 с вероятностями 1/4 и 3/4 соответственно, а среднее значение равно 2,812. Это означает, что величина среднего значения количества информации оказывается разной для различных подпоследовательностей в пределах общей последовательности. И только среднее арифметическое для двух средних значений, полученных по отдельным подпоследовательностям, совпадает с энтропией, т. е. средней по ансамблю. Таким образом, рассматриваемый источник не является эргодическим.

Ещё раз обратим внимание на то, что характерным признаком этого источника является наличие у него двух различных режимов работы, соответствующих как бы двум отдельным источникам. Такие источники, как было сказано выше, называются разложимыми, и можно показать, что любой стационарный источник может рассматриваться как комбинация эргодических источников (возможно, бесконечного числа), каждый из которых задает некоторый отличный от других режим работы всего источника.

Большинство практически интересных стационарных источников может быть аппроксимировано марковскими источниками различных порядков. С другой стороны, существует возможность моделирования реальных источников (осмысленный текст, музыка и др.) на основе задания наборов вероятностей, определяющих статистические соотношения между отдельными соотношениями. Некоторые из примеров такого моделирования будут рассмотрены ниже.

2.4. АСИМПТОТИЧЕСКАЯ РАВНОРАСПРЕДЕЛЕННОСТЬ РЕАЛИЗАЦИЙ СЛУЧАЙНОГО ПРОЦЕССА

Как было указано в разд. 2.1, осуществление конкретной реализации Ψ_n длиною в n элементов можно рассматривать как осуществление одного из l^n возможных событий с соответствующей вероятностью $p_k(\Psi_n)$. На множестве таких событий можно задать любую числовую функцию $f(\Psi_n)$, которая, очевидно, также будет случайной величиной. В частности, такой случайной величиной может быть функция

$$f\left(\Psi_n\right) = -\frac{1}{n}\log p_k\left(\Psi_n\right), \qquad (2.4.1)$$

математическое ожидание которой равно

$$\mathbf{E}\left[f\left(\Psi_n\right)\right] = -\frac{1}{n}\sum_{k=1}^{l^n} p_k\left(\Psi_n\right)\log p_k\left(\Psi_n\right) = \frac{H_n}{n}.$$

Поскольку предел $\lim_{n\to\infty}\left(H_n / n\right)$ существует и равен H — энтропии случайного процесса,

$$\lim_{n\to\infty}\mathbf{E}\left[f\left(\Psi_n\right)\right] = \lim_{n\to\infty}\mathbf{E}\left[-\frac{1}{n}\log p_k\left(\Psi_n\right)\right] = H,$$

т. е. при $n \to \infty$ математическое ожидание случайной величины $f(\Psi_n)$ имеет пределом H.

Такое соотношение между $p_k(\Psi_n)$ и H, весьма интересное уже само по себе, является, однако, лишь одним из проявлений гораздо более общего свойства дискретных эргодических процессов. Оказывается, что не только математическое ожидание величины $f(\Psi_n)$ при $n \to \infty$ имеет пределом H, но и сама величина $f(\Psi_n)$ стремится по вероятности к H при $n \to \infty$. Другими словами, при достаточно большом значении n близость $f(\Psi_n)$ к H является почти достоверным событием.

Это фундаментальное свойство эргодических процессов впервые было обнаружено ещё Шенноном [1] на примерах процесса с независимыми символами и простой марковской цепи. Тогда же он высказал предположение, что этим свойством обладают любые эргодические процессы, однако строгое доказательство этого факта было найдено Б. Макмилланом лишь в 1953 г. Для большей наглядности отмеченное выше свойство, называемое в литературе *свойством асимптотической равнораспределённости* реализаций эргодического процесса, формулируют следующим образом.

Для любых заданных $\varepsilon > 0$ и $\delta > 0$ *можно найти такое натуральное* n_ε, *что реализации любой длины* $n \geq n_\varepsilon$ *распадаются на две группы:*

- *группа реализаций, общая вероятность которых не превышает значения* δ*;*
- *группа реализаций, вероятности которых удовлетворяют неравенству*

$$\left|(1/n)\log p_k\left(\Psi_n\right) + H\right| < \varepsilon. \qquad (2.4.2)$$

Первую группу реализаций называют “маловероятной”, вторую — “высоковероятной”.

Докажем свойство асимптотической равнораспределённости для случая простой марковской эргодической последовательности[1], характеризуемой зависимостью лишь от одного предыдущего шага.

Рассмотрим произвольную реализацию Ψ_n, вероятность которой (2.1.1) в данном случае равна

$$p_k(\Psi_n) = P\left(\xi^{(1)}\right) P\left(\xi^{(2)} / x^{(1)}\right) \cdots P\left(\xi^{(n)} / x^{(n-1)}\right). \qquad (2.4.3)$$

Вероятности переходов

$$p_{ij} = P\left(\xi^{(m)} / x^{(m-1)}\right),\ 2 \le m \le n,$$

оценим с помощью закона больших чисел для эргодической марковской цепи. Согласно этому закону, число переходов μ_{ij} цепи из состояния i в j в последовательности достаточно большой длины n по вероятности стремится к величине $p_i p_{ij} n$, где p_i — вероятность i-го состояния, т. е. для любых положительных чисел δ и ε можно подобрать такое n, что

$$P\left(\left|\mu_{ij} / n - p_i p_{ij}\right| > \delta\right) < \varepsilon.$$

Иными словами, в реализации Ψ_n будет приблизительно $p_1 p_{12} n$ раз встречаться последовательность символов $x_2 x_1$, приблизительно $p_1 p_{13} n$ раз — последовательность $x_3 x_1$ и т. д., приблизительно $p_i p_{ij} n$ раз — последовательность $x_j x_i$. Тогда с точностью до сколь угодно малой составляющей вероятность реализации Ψ_n равна

$$p_k(\Psi_n) = P\left(\xi^{(1)}\right) p_{11}^{p_1 p_{11} n} \cdots p_{1n}^{p_1 p_{1n} n} p_{21}^{p_2 p_{21} n} \cdots p_{2n}^{p_2 p_{2n} n} \times \cdots \times$$

$$\times p_{n1}^{p_n p_{n1} n} \cdots p_{nn}^{p_n p_{nn} n} = P\left(\xi^{(1)}\right) \prod_{(i,j)} p_{ij}^{p_i p_{ij} n},$$

где произведение вычисляется по всем возможным парам (i, j). Отсюда

$$\frac{\log p_k(\Psi_n)}{n} = \frac{P\left(\xi^{(1)}\right)}{n} + \sum_{(i,j)} p_i p_{ij} \log p_{ij} \underset{n\to\infty}{\to} \sum_{(i,j)} p_i p_{ij} \log p_{ij}.$$

Внутренняя сумма по j:

$$\sum_{(j)} p_{ij} \log p_{ij}$$

представляет собой взятую с обратным знаком условную энтропию H_i марковского источника, находящегося в i-м состоянии. Следовательно,

[1] В частности, это будет справедливо и для последовательностей независимых величин, являющихся марковскими последовательностями нулевого порядка.

усреднение по всем состояниям (внешняя сумма по i) в итоге даёт безусловную энтропию H:

$$\sum_{(i,j)} p_i p_{ij} \log p_{ij} = -H ,$$

что и доказывает (2.4.2).

Свойство асимптотической равнораспределённости приводит к ряду важных следствий, два из которых заслуживают особого внимания.

В качестве первого следствия укажем, что соотношение (2.4.2) означает приблизительную равновероятность всех реализаций высоковероятной группы, независимо от того, как распределены вероятности отдельных элементов, и как элементы связаны между собой, т. е. $p_k(\Psi_n) = p(\Psi_n)$. В частности, по известной вероятности $p_k(\Psi_n)$ одной из реализаций высоковероятной группы может быть оценено само число N_2 реализаций в группе: $N_2 \approx 1 / p(\Psi_n)$.

Вторым следствием является утверждение о том, что при больших n высоковероятная группа охватывает лишь малую долю всех возможных реализаций. Действительно, из (2.4.2) и первого следствия имеем:

$$-\log p\left(\Psi_n\right) = \log N_2 \approx nH ,$$

откуда $N_2 = a^{nH}$, где a — основание логарифма.

С другой стороны, общее число N всех возможных реализаций равно $N = l^n = a^{n\log l}$. Таким образом, доля высоковероятных реализаций характеризуется отношением

$$\frac{N_2}{N} = a^{n\left(H-\log l\right)}, \qquad (2.4.4)$$

являющимся экспоненциальным с ростом n убыванием, поскольку $H < \log l$.

Пусть, например, $n = 100$, $l = 8$ и $H = 2{,}75$. Тогда, переходя к двоичным логарифмам, имеем $N = 8^{100} = 2^{300}$, $N_2 = 2^{100\cdot 2{,}75} = 2^{275}$. Отсюда $N_2 / N = 2^{-25} \approx 3{,}0 \cdot 10^{-8}$.

2.5. ЭНТРОПИЯ НЕПРЕРЫВНОГО СЛУЧАЙНОГО ПРОЦЕССА

Рассмотрим теперь возможность определения энтропии для случайных процессов с дискретным временем, но непрерывными значениями. Задание плотностей распределений все более высоких порядков

$$w(x,\ t),\ w^{(2)}(x_1,\ x_2;\ t_1,\ t_2),\ \ldots,\ w^{(n)}(x_1,\ \ldots,\ x_n;\ t_1,\ \ldots,\ t_n) \qquad (2.5.1)$$

означает все более и более полную характеризацию свойств такого процесса. Поскольку значения процесса непрерывны, можно вычислить многомерную дифференциальную энтропию $h_n(\xi)$, равную

$$h_n(\xi) = -\iint\ldots\int w^{(n)}(\mathbf{x},\mathbf{t})\log\left[w^{(n)}(\mathbf{x},\mathbf{t})\right]d\mathbf{x}. \qquad (2.5.2)$$

Проследим, как изменяется значение $h_n(\xi)$ с ростом n, а также при изменении статистической связи между отдельными значениями процесса. При этом для простоты и наглядности ограничимся стационарным гауссовским процессом с нулевым средним и дисперсией σ^2.

Совместное n-мерное распределение гауссовского процесса задаётся выражением (1.5.16), в котором, с учётом стационарности и нулевого среднего, перейдём от ковариаций $b_{ki} = \mathrm{Cov}[\xi_k, \xi_i]$ к коэффициентам корреляций $R_{ki} = b_{ki} / \sigma^2$:

$$w^{(n)}(\mathbf{x},\mathbf{t}) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}\sqrt{D_n}}\exp\left(-\frac{1}{2D_n\sigma^2}\sum_{k=1}^{n}\sum_{i=1}^{n}D_{ki}x_kx_i\right), \qquad (2.5.3)$$

где σ^2 — дисперсия, D_n — определитель корреляционной матрицы $\|R_{ki}\|$, а ρ_{ki} — алгебраические дополнения коэффициентов корреляции R_{ki}. Вычисление интеграла (2.5.2) приводит к выражению, аналогичному (1.5.19):

$$h_n(\xi) = \log\left[\left(2\pi\sigma^2e\right)^{n/2}D_n\right]. \qquad (2.5.4)$$

В случае статистической независимости отсчётов $x_1, \ldots, x_n$ многомерное распределение факторизуется:

$$w^{(n)}(x_1,\ \ldots,\ x_n;\ t_1,\ \ldots,\ t_n) = \prod_{k=1}^{n}w(x_k,\ t_k).$$

При этом $h_n(\xi) = nh_1(\xi)$, что является верхней оценкой для дифференциальной энтропии n-го порядка, ибо $0 \le D_n \le 1$, и наличие корреляционных связей приводит только к уменьшению $h_n(\xi)$.

Для получения более конкретных результатов сделаем дальнейшее предположение о свойствах процесса. Будем считать, что коэффициент корреляции между значениями x_k и x_i, отстоящими друг от друга по времени на величину τ, равен

$$R(\tau) = \exp\left(-|\tau|/\tau_0\right). \qquad (2.5.5)$$

Пусть для простоты t меняется дискретно через одинаковые интервалы T, так что все R_{ki} являются целочисленными степенями некоторого числа $r = \exp(-T / \tau_0)$. Тогда определитель D_n корреляционной матрицы равен [4]

$$D_n = \begin{Vmatrix} 1 & r & r^2 & \dots & r^{n-1} \\ r & 1 & r & \dots & r^{n-2} \\ r^2 & r & 1 & \dots & r^{n-3} \\ \dots & \dots & \dots & \dots & \dots \\ r^{n-1} & r^{n-2} & r^{n-3} & \dots & 1 \end{Vmatrix} = \left(1-r^2\right)^{n-1}. \qquad (2.5.6)$$

С учётом этого,

$$h_n(\xi) = \log\left[\left(2\pi\sigma^2 e\right)^{n/2}\left(1-\exp\left(-2T/\tau_0\right)\right)^{\frac{n-1}{2}}\right]. \qquad (2.5.7)$$

Теперь, желая охарактеризовать процесс в целом, используем предельное соотношение (“энтропия на одну степень свободы”)

$$h(\xi) = \lim_{n\to\infty}\frac{1}{n}h_n(\xi),$$

которое в рассматриваемом случае имеет вид

$$h(\xi) = \frac{1}{2}\log\left(2\pi\sigma^2 e\right) + \lim_{n\to\infty}\left(1-1/n\right)\log\sqrt{1-\exp\left(-2T/\tau_0\right)} =$$

$$= h_1(\xi) + \log\sqrt{1-\exp\left(-2T/\tau_0\right)}. \qquad (2.5.8)$$

Ввиду отрицательности второго слагаемого можно сделать вывод о том, что наличие корреляции между отдельными значениями процесса уменьшает его энтропию.

2.6. ВЗАИМНАЯ ИНФОРМАЦИЯ МЕЖДУ НЕПРЕРЫВНЫМИ СЛУЧАЙНЫМИ ПРОЦЕССАМИ

Взаимная информация между непрерывными случайными величинами была определена в первой главе. Понятие взаимной информации между непрерывными случайными процессами основано на использовании *ортогональных разложений* процессов, в частности, *ортогонального разложения Карунена–Лоэва*. Не вдаваясь в математические тонкости, кратко напомним их основные идеи.

Любой случайный процесс $\xi(t)$ на рассматриваемом временном интервале $[a; b]$ представим в виде ортогонального разложения по системе

$\{\varphi_k(t)\}$ ($k = 1, 2, \ldots$) ортонормированных функций:

$$\xi(t) = \sum_{k=1}^{\infty} \xi_k \varphi_k(t), \qquad (2.6.1)$$

где здесь и далее, если не оговорено иначе, все подобные равенства понимаются в среднеквадратическом смысле, т. е.

$$\lim_{k\to\infty} \mathbf{E}\left[\left(\xi(t) - \sum_{k=1}^{n} \xi_k \varphi_k(t)\right)^2\right] = 0, \qquad (2.6.2)$$

а случайные величины (“координаты” случайного процесса в возникающем функциональном пространстве) ξ_k определяются интегральным соотношением

$$\xi_k = \int_a^b \xi(t)\varphi_k(t)dt\,, k = 1, 2, \ldots\,.$$

Иными словами, случайный процесс $\xi(t)$, изначально задаваемый либо совокупностью всех своих реализаций, либо совокупностью всех конечномерных распределений, можно эквивалентно задать набором (вообще говоря, бесконечным) коэффициентов $\{\xi_k\}$ в представлении (2.6.1).

Например, пусть $\{\xi_k\}$ — совокупность независимых гауссовских случайных величин с нулевым средним и одинаковой дисперсией

$$\mathbf{E}\left[\xi_k^2\right] = N_0/2\,, \quad k = 1, 2, \ldots\,,$$

а в качестве ортогонального базиса рассмотрим систему гармонических функций

$$\varphi_{2k-1}(t) = \sqrt{\frac{(b-a)}{2}}\sin\frac{2\pi kt}{b-a};$$

$$\varphi_{2k}(t) = \sqrt{\frac{(b-a)}{2}}\cos\frac{2\pi kt}{b-a}.$$

В этом случае в представлении (2.6.1) имеем процесс, у которого распределение мощности по спектральным составляющим равномерно — такой процесс называется *белым гауссовским шумом*, а значение $N_0/2$ — *спектральной плотностью средней мощности*, или *интенсивностью* процесса.

Заметим, что если при тех же $\{\xi_k\}$ и $\{\varphi_k(t)\}$ рассматривать усечённый ряд

$$\xi_n(t) = \sum_{k=1}^{n} \xi_k \varphi_k(t), \qquad (2.6.1a)$$

то получается процесс, у которого для всех спектральных составляющих в диапазоне $[-2\pi k / T;\ 2\pi k / T]$ мощность одинакова и равна $N_0 / 2$, а для всех остальных составляющих мощность равна нулю. Такой процесс иногда называют *частотно-ограниченным белым гауссовским шумом*.

Разложение Карунена–Лоэва является частным случаем ортогонального разложения (2.6.1), когда в качестве ортогонального базиса рассматриваются функции $\{\varphi_k(t)\}$, являющиеся решением интегрального уравнения

$$\int_a^b K_\xi(t_1,t_2)\varphi_k(t_1)dt = \lambda_k\varphi_k(t_2), \qquad (2.6.3)$$

где

$$K_\xi(t_1,t_2) = \mathbf{E}\left\{\xi(t_1) - \mathbf{E}\left[\xi(t_1)\right]\right\}\left\{\xi(t_2) - \mathbf{E}\left[\xi(t_2)\right]\right\}$$

является корреляционной функцией процесса $\xi(t)$.

Числа λ_k, при которых уравнение (2.6.3) имеет решения, называются *собственными числами*, а сами функции $\{\varphi_k(t)\}$ — *собственными функциями* уравнения. Из теории интегральных уравнений [10], а также из свойств симметричности и неотрицательной определённости корреляционной функции следует, что все собственные числа $\lambda_1, \lambda_2, \ldots$ вещественны и неотрицательны, а система ортонормированных функция является полной, т. е. справедливо соотношение (2.6.2) среднеквадратической аппроксимации.

Важнейшим свойством разложения Карунена–Лоэва является то, что

$$\mathbf{E}\left[\xi_k\xi_j\right] = \begin{cases}\lambda_k, & k = j;\\ 0, & k \neq j.\end{cases} \qquad (2.6.4)$$

Иными словами, в этом случае различные коэффициенты ряда в (2.6.1) оказываются некоррелированными между собой, а дисперсия k-го коэффициента равна k-му собственному числу.

Обратимся теперь к понятию взаимной информации между непрерывными процессами. Пусть процессы $\xi(t)$ и $\eta(t)$ заданы соответствующими ортогональными разложениями вида (2.6.1), т. е. полностью определяются бесконечными наборами коэффициентов $\{\xi_1, \xi_2, \ldots\}$ и $\{\eta_1, \eta_2, \ldots\}$. Тогда совместное определение процессов $\xi(t)$ и $\eta(t)$ означает задание всех конечномерных совместных распределений $w_{\xi\eta}^{(2n)}(\mathbf{x},\mathbf{y})$, ($n$ = 1, 2, …) с последующим предельным переходом $n \to \infty$. Подчеркнём, что такие распределения

следует отличать от распределений типа (2.5.1), определяющих процессы во временны*ы*х сечениях.

Опираясь на понятие взаимной информации для непрерывных источников, введём сначала взаимную информацию $I_n(\boldsymbol{\xi}, \boldsymbol{\eta})$ для конечных n-элементных наборов коэффициентов:

$$I_n(\boldsymbol{\xi}, \boldsymbol{\eta}) = \int \cdots \int w_{\boldsymbol{\xi}\boldsymbol{\eta}}^{(2n)}(\mathbf{x}, \mathbf{y}) \log \frac{w_{\boldsymbol{\xi}\boldsymbol{\eta}}^{(2n)}(\mathbf{x}, \mathbf{y})}{w_{\boldsymbol{\xi}}^{(n)}(\mathbf{x}) w_{\boldsymbol{\eta}}^{(n)}(\mathbf{y})} d\mathbf{x} d\mathbf{y}\,, \qquad (2.6.5)$$

где интегрирование ведётся по всему $2n$-мерному ансамблю распределений, а маргинальные распределения $w_{\boldsymbol{\xi}}^{(n)}(\mathbf{x})$ и $w_{\boldsymbol{\eta}}^{(n)}(\mathbf{y})$ вычисляются посредством усреднения по "противоположным" составляющим, т. е.

$$w_{\boldsymbol{\xi}}^{(n)}(\mathbf{x}) = \int \cdots \int w_{\boldsymbol{\xi}\boldsymbol{\eta}}^{(2n)}(\mathbf{x}, \mathbf{y}) d\mathbf{y}\,;$$

$$w_{\boldsymbol{\eta}}^{(n)}(\mathbf{y}) = \int \cdots \int w_{\boldsymbol{\xi}\boldsymbol{\eta}}^{(2n)}(\mathbf{x}, \mathbf{y}) d\mathbf{x}\,.$$

Тогда "истинная" взаимная информация $I[\xi(t), \eta(t)]$ между процессами $\xi(t)$ и $\eta(t)$ может быть определена предельным переходом:

$$I\left[\xi(t), \eta(t)\right] = \lim_{n \to \infty} I_n[\boldsymbol{\xi}, \boldsymbol{\eta}]\,, \qquad (2.6.6)$$

если, конечно, такой предел существует.

Чтобы пояснить введённое понятие взаимной информации рассмотрим простой пример. Пусть необходимо найти взаимную информации между процессами $\xi(t)$ и $\eta(t)$, заданными на интервале $[a; b]$, причём процесс $\eta(t)$ представляет собой сумму статистически независимых процессов $\xi(t)$ и $\zeta(t)$. Нетрудно видеть, что данный пример является обобщением примера вычисления взаимной информации для случайных величин, рассмотренный в разд. 1.7.

Из независимости процессов следует, что для любого $n = 1, 2, \ldots$ векторы $\boldsymbol{\xi}$ и $\boldsymbol{\zeta}$, образованные первыми n коэффициентами разложений в ряд (2.6.1), статистически независимы, а условное распределение $\eta(t)$ при заданном $\xi(t)$ определяется распределением $\zeta(t)$, т. е.

$$w_{\boldsymbol{\eta}}^{(n)}(\mathbf{y} / \boldsymbol{\xi}) = w_{\boldsymbol{\zeta}}^{(n)}(\mathbf{z})\,.$$

Тогда, вводя дифференциальные энтропии процессов, имеем:

$$I_n(\boldsymbol{\xi}, \boldsymbol{\eta}) = h_n(\boldsymbol{\eta}) - h_n(\boldsymbol{\eta} / \boldsymbol{\xi}) = h_n(\boldsymbol{\eta}) - h_n(\boldsymbol{\zeta}) =$$

$$= h_n(\boldsymbol{\eta}) + \int \cdots \int w_{\boldsymbol{\zeta}}^{(n)}(\mathbf{z}) \log w_{\boldsymbol{\zeta}}^{(n)}(\mathbf{z}) d\mathbf{z}\,.$$

Разложим процессы $\xi(t)$ и $\eta(t)$ ортогональным образом по Карунену–Лоэву в одном и том же базисе $\{\varphi_k(t)\}$:

$$\eta_k = \int_a^b \eta(t)\varphi_k(t)dt = \int_a^b \xi(t)\varphi_k(t)dt + \int_a^b \zeta(t)\varphi_k(t)dt = \xi_k + \zeta_k .$$

Конкретизируем распределения процессов $\xi(t)$ и $\zeta(t)$, полагая, что они — гауссовские с нулевыми средними. Тогда коэффициенты ζ_k также гауссовские с нулевыми средними и дисперсиями λ_k, являющимися собственными числами интегрального ядра $K_\zeta(t_1, t_2)$ в (2.6.3). Учитывая аддитивность дифференциальной энтропии для независимых величин, запишем

$$h_n[\boldsymbol{\zeta}] = \sum_{k=1}^{n} h[\zeta_k] = \frac{1}{2}\sum_{k=1}^{n} \log\left(2\pi e\lambda_k\right).$$

Такие же рассуждения позволяют записать аналогичное соотношение для $h_n[\boldsymbol{\eta}]$:

$$h[\boldsymbol{\eta}] = \frac{1}{2}\sum_{k=1}^{n} \log\left[2\pi e\left(\mu_k + \lambda_k\right)\right],$$

где μ_k — собственные числа интегрального ядра $K_\xi(t_1, t_2)$.

Таким образом, объединяя слагаемые и осуществляя предельный переход (2.6.6), получим выражение для взаимной информации между гауссовскими процессами $\xi(t)$ и $\eta(t)$:

$$I\left[\xi(t), \eta(t)\right] = \frac{1}{2}\sum_{k=1}^{\infty} \log\left(1 + \frac{\mu_k}{\lambda_k}\right). \qquad (2.6.7)$$

Соотношение (2.6.7) по форме идентично выражению (1.7.10) для взаимной информации между многомерными распределениями. Это и не удивительно, поскольку оба они получены при применении “родственных” ортогональных преобразований: (1.7.10) — как результат ортогонального преобразования корреляционной матрицы для n-элементных последовательностей $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ и $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$, а (2.6.7) — как решение интегрального уравнения (2.6.3), приводящего к ортогональному разложению Карунена–Лоэва.

2.7. ЭНТРОПИЯ В ЕДИНИЦУ ВРЕМЕНИ И ЭНТРОПИЙНАЯ МОЩНОСТЬ СЛУЧАЙНЫХ ПРОЦЕССОВ

В предыдущем разделе было рассмотрено возможное представление непрерывного (и во времени, и по значениям) процесса своими ортого-

нальными некоррелированными координатами. Несмотря на фундаментальность и перспективность этого метода он, тем не менее, не лишён ряда недостатков, главный из которых — сложность в практическом определении собственных чисел и собственных функций при решении интегрального уравнения (2.6.3) для реальных процессов. В данном разделе рассматривается более простой и в некотором смысле альтернативный подход к представлению непрерывного процесса дискретным (счётным) набором значений, основанный на использовании *разложения Котельникова*[1].

Согласно такому разложению, непрерывный процесс $\xi(t)$, энергетический спектр которого ограничен верхней спектральной составляющей F, может быть однозначным образом представлен своими отсчётами, если они берутся с интервалом времени $\Delta t \le 2F$:

$$\xi(t) = \sum_{k=-\infty}^{\infty} \xi\left(k\Delta t\right) \frac{\sin\left[2\pi F\left(k\Delta t\right)\right]}{2\pi F\left(k\Delta t\right)}. \qquad (2.7.1)$$

Отвлекаясь от технических особенностей при реализации самого разложения Котельникова, укажем две причины, существенно снижающие качество его практического использования.

Прежде всего, заметим, что любые используемые на практике сигналы имеют конечную длительность и, следовательно, неограниченную ширину полосы занимаемых частот. В этой связи значение F верхней спектральной составляющей оказывается неопределённым, и его приходится выбирать, с одной стороны, исходя из необходимости воспроизведения как можно большей полосы частот сигнала (и, следовательно, его качества), а с другой — учитывая объёмы порождаемых при этом потоков информации, что в конечном итоге определяет сложность и стоимость всей системы передачи информации в целом.

Кроме того, фигурирующая в (2.7.1) функция

$$\operatorname{sinc}(2\pi Ft) = \frac{\sin(2\pi Ft)}{2\pi Ft}$$

(её иногда называют *котельниковским ядром*) имеет бесконечную длительность, что входит в противоречие с рассмотрением процессов на конечном временном интервале.

[1] В иностранной литературе, особенно американской, это разложение часто называют теоремой Найквиста или теоремой отсчётов.

Тем не менее, определив граничную спектральную составляющую по какому-либо выбранному критерию, например, по критерию заданной концентрации энергии, и рассматривая конечное число слагаемых в сумме (2.7.1), можно трактовать её как приближенное представление процесса $\xi(t)$ с некоторой наперёд заданной точностью.

Рассмотрим гауссовский процесс с равномерным энергетическим спектром в полосе частот F Гц. Согласно теореме Котельникова, возможно неискажённое представление такого процесса его отсчётными значениями, если их число составляет не менее $2F$ в одну секунду. Тогда, определив отсчётный интервал $\Delta t = 1/(2F)$, для такого процесса можно ввести *дифференциальную энтропию на единицу времени* (*дифференциальную энтропию на степень свободы*)

$$h_t = \frac{h}{\Delta t} = 2Fh = F\log\left(2\pi e\sigma^2\right). \qquad (2.7.2)$$

Вычислим теперь дифференциальную энтропию на единицу времени произвольного гауссовского процесса с энергетическим спектром $G(f)$. Для этого разобьём весь частотный диапазон на малые интервалы df, в пределах которых энергетический спектр можно аппроксимировать постоянным значением, а отсчётные временны*е* значения считать независимыми. Тогда дифференциальная энтропия на единицу времени для некоего процесса, спектр которого находится в интервале $[f; f + df]$ в соответствии с (2.7.2) равна

$$dh_t = df\log\left[2\pi eG(f)\right].$$

Поскольку энтропия совокупности независимых процессов равна сумме соответствующих энтропий, для всего исходного процесса дифференциальная энтропия на единицу времени

$$h_t = \int_F \log\left[2\pi G(f)e\right]df\,,$$

или, возвращаясь к "обычной" дифференциальной энтропии,

$$h = \frac{1}{2F}\int_F \log\left[2\pi G(f)e\right]df\,. \qquad (2.7.3)$$

Сравнивая теперь (2.7.3) с выражением (2.5.4), можно получить ряд интересных и полезных соотношений. Например, поскольку, как это следует из свойств сходящихся последовательностей, $h = \lim\limits_{n\to\infty}\left(h_n - h_{n-1}\right)$ и, следовательно,

$$h = \lim_{n\to\infty} \frac{1}{2} \log \left[\left(2\pi\sigma^2 e\right) \frac{D_n}{D_{n-1}} \right], \qquad (2.7.4)$$

сопоставляя (2.5.4), (2.7.3) и (2.7.4), получим

$$h = \lim_{n\to\infty} \frac{1}{2n} \log D_n = \frac{1}{2F} \int_F \log \left[G\left(f\right) / \sigma^2 \right] df \,, \qquad (2.7.5a)$$

$$h = \lim_{n\to\infty} \frac{1}{2} \log \frac{D_n}{D_{n-1}} = \frac{1}{2F} \int_F \log \left[G\left(f\right) / \sigma^2 \right] df \,. \qquad (2.7.5б)$$

Полезность соотношений (2.7.5) проявляется при необходимости оценки величины *n*-мерных корреляционных определителей, трудно поддающихся непосредственному вычислению. Важно отметить, что если случайный процесс, имеющий энергетический спектр $G(f)$ и корреляционную матрицу $||R_{ki}||$, не является гауссовским, то в силу экстремальности энтропии нормального распределения правые части указанных соотношений могут служить верхней оценкой для соответствующих определителей.

Во многих случаях при рассмотрении процессов с непрерывными значениями оказывается удобным пользоваться не непосредственно дифференциальной энтропией или дифференциальной энтропией в единицу времени, а производной от них величиной, названной Шенноном *энтропийной мощностью*. Определим энтропийную мощность $\bar{N}$ некоторого случайного процесса, имеющего ширину спектра F и дифференциальную энтропию h, как среднюю мощность нормального процесса с равномерным энергетическим спектром в такой же полосе и такой же дифференциальной энтропией на единицу времени:

$$\bar{N} = \frac{1}{2\pi e} \exp\left(2h\right). \qquad (2.7.6)$$

Отметим некоторые свойства энтропийной мощности.

Энтропийная мощность случайного процесса с произвольным распределением не превышает его средней мощности: $\bar{N} \le \sigma^2$. Это следует из того, что при заданной средней мощности гауссовский процесс обладает максимальной дифференциальной энтропией.

Энтропийная мощность нормального процесса с произвольным энергетическим спектром $G(f)$ выражается формулой

$$\bar{N} = \exp\left\{ \frac{1}{F} \int_F \log G(f) df \right\}. \qquad (2.7.8)$$

Для доказательства (2.7.8) достаточно подставить (2.7.3) в (2.7.6):

$$\bar{N} = \frac{1}{2\pi e}\exp\left\{\frac{1}{F}\int_F \log\left[2\pi e G(f)\right]df\right\} = \exp\left\{\frac{1}{F}\int_F \log G(f)df\right\}.$$

2.8. ЭПСИЛОН-ЭНТРОПИЯ СЛУЧАЙНЫХ ПРОЦЕССОВ

В разд. 1.9 было введено понятие эпсилон-энтропии источника как минимальной взаимной информации, необходимой для воспроизведения значений источника с определённой погрешностью при заданном критерии качества. Обобщим теперь это понятие применительно к случайным процессам.

Пусть источник X в каждый момент времени выбирает дискретное или непрерывное сообщение, и эти сообщения образуют множество последовательностей вида

$$\mathbf{x} = \left(x^{(1)}, x^{(2)}, \ldots, x^{(n)}\right).$$

Анализ источника X посредством наблюдений за событиями вторичного источника Y будем трактовать как аппроксимацию последовательностей $\mathbf{x}$ последовательностями

$$\mathbf{y} = \left(y^{(1)}, y^{(2)}, \ldots, y^{(n)}\right),$$

т. е. отображение многомерного входного множества X^n на многомерное выходное множество Y^n.

Как и в одномерном случае, существуют различные способы отображения, как вероятностные, так и детерминированные. При этом, исходя из практических соображений, количество возможных аппроксимирующих последовательностей $\mathbf{y}_1, \ldots, \mathbf{y}_M$ ограничено, даже если бесконечно множество входных последовательностей.

Введём в рассмотрение *критерий качества аппроксимации n-элементных последовательностей*

$$d_n(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_{k=1}^{n} d\left(x^{(k)}, y^{(k)}\right), \qquad (2.8.1)$$

где d — расстояние (мера различия) между "элементарными" сообщениями $x^{(k)}$ и $y^{(k)}$. Тогда, задаваясь многомерными входными распределениями $p(\mathbf{x})$ и условными распределения $p(\mathbf{y}/\mathbf{x})$ ($\mathbf{x} \in X^n$, $\mathbf{y} \in Y^n$) для дискретных

источников или соответствующими плотностями вероятностей для непрерывных источников, можно определить среднюю ошибку аппроксимации, равную

$$\Delta_n = \mathbf{E}\left[d_n(\mathbf{x},\mathbf{y})\right] = \sum_{X^n}\sum_{Y^n} d_n(\mathbf{x},\mathbf{y})\,p(\mathbf{y}/\mathbf{x})\,p(\mathbf{x}) \qquad (2.8.2a)$$

в дискретном и

$$\Delta_n = \mathbf{E}\left[d_n(\mathbf{x},\mathbf{y})\right] = \int_{X^n}\int_{Y^n} d_n(\mathbf{x},\mathbf{y})\,w(\mathbf{y}/\mathbf{x})\,w(\mathbf{x})\,d\mathbf{x}d\mathbf{y} \qquad (2.8.2б)$$

в непрерывном случаях.

Линейность математического ожидания позволяет применить соотношения (2.8.2) к одномерным событиям, т. е. к парам $(x^{(k)}, y^{(k)})$:

$$\Delta_n = \frac{1}{n}\sum_{k=1}^{n}\mathbf{E}\left[d\left(x^{(k)},y^{(k)}\right)\right], \qquad (2.8.3)$$

где

$$\mathbf{E}\left[d\left(x^{(k)},y^{(k)}\right)\right] = \sum_{X}\sum_{Y} d\left(x,y\right)p\left(y^{(k)}/x^{(k)}\right)p\left(x^{(k)}\right) \qquad (2.8.4a)$$

и

$$\mathbf{E}\left[d\left(x^{(k)},y^{(k)}\right)\right] = \int_{X}\int_{Y} d\left(x,y\right)w\left(y^{(k)}/x^{(k)}\right)w\left(x^{(k)}\right)dxdy \qquad (2.8.4б)$$

для дискретного и непрерывного случаев соответственно.

В разд. 1.9 в качестве меры различия между элементарными сообщениями $x^{(k)}$ и $y^{(k)}$ рассматривался средний квадрат разности. Несмотря на удобство и важность такого критерия он, разумеется, не является универсальным, и во многих случаях целесообразно оценивать качество аппроксимации по другим критериям.

Пусть, например, источники X и Y состоят из одинакового набора сообщений. Введём меру различия $d(x, y)$ в виде "дискретной дельта-функции":

$$d(x,y) = \begin{cases} 1, & x \neq y; \\ 0, & x = y. \end{cases} \qquad (2.8.5)$$

Тогда средняя ошибка аппроксимации

$$\Delta_n = \frac{1}{n}\sum_{k=1}^{n}\sum_{X}\sum_{Y} d(x,y)\,p(x,y) = \frac{1}{n}\sum_{k=1}^{n} P\left\{x^{(k)} \neq y^{(k)}\right\} \qquad (2.8.6)$$

просто равна средней вероятности того, что символы в последовательностях **x** и **y** не будут совпадать.

В частности, пусть двоичные ($l = 2$) источники $X = Y = \{0, 1\}$ генерируют трехэлементные ($n = 3$) последовательности. При этом входное распределение равномерное, а аппроксимирующее множество состоит из двух последовательностей $\mathbf{y}_1 = (000)$ и $\mathbf{y}_2 = (111)$. В соответствии с (2.8.5) и (2.8.6) значение Δ_n равно одной трети от числа символов, в которых последовательности **x** и **y** не совпадают.

Предложим следующий способ аппроксимации: для каждой последовательности **x** будем выбирать такую из двух последовательностей **y**, которая минимизирует значение d_3. Например, для $\mathbf{x} = (010)$ аппроксимация $\mathbf{y}_1 = (000)$ даёт $d_3 = 1/3$, в то время, как аппроксимация $\mathbf{y}_2 = (111)$ даёт $d_3 = 2/3$, поэтому выбирается первый вариант. В табл. 2.1 приведены входные и аппроксимирующие последовательности, а также соответствующие им значения d_3.

Таблица 2.1

Входные и аппроксимирующие последовательности

Исходный источник **x**	Аппроксимирующий источник **y**	Ошибка $d_3(\mathbf{x}, \mathbf{y})$	Исходный источник **x**	Аппроксимирующий источник **y**	Ошибка $d_3(\mathbf{x}, \mathbf{y})$
000	000	0	100	000	1/3
001	000	1/3	101	111	1/3
010	000	1/3	110	111	1/3
011	111	1/3	111	111	0

Очевидно, что минимизация отдельных значений d_n приводит к минимизации всей средней вероятности ошибки Δ_n, которая для приведённого примера составляет $\Delta_3 = 1/4$.

Введём теперь ряд определений, обобщающих соответствующие определения разд. 1.9 для n-элементных последовательностей.

Наряду со средней ошибкой аппроксимации на $2n$-мерном ансамбле $X^n Y^n$ определим взаимную информацию $I(X^n; Y^n)$ между n-элементными последовательностями ансамблей X^n и Y^n:

$$I(X^n, Y^n) = \int\limits_{X^n} \int\limits_{Y^n} w(\mathbf{y} / \mathbf{x}) w(\mathbf{x}) \log \frac{w(\mathbf{y} / \mathbf{x})}{w(\mathbf{y})} d\mathbf{x} d\mathbf{y}, \qquad (2.8.7)$$

где многомерная плотность вероятностей $w(\mathbf{y})$ на ансамбле Y^n находится по заданным плотностям $w(\mathbf{x})$ и $w(\mathbf{x}/\mathbf{y})$ путём интегрирования по ансамблю X^n:

$$w(\mathbf{y}) = \int_{X^n} w(\mathbf{y}/\mathbf{x})\, w(\mathbf{x})\, d\mathbf{x}. \qquad (2.8.8)$$

Введём класс функций $\Omega_n(\varepsilon)$, куда поместим все условные плотности вероятностей $w(\mathbf{x}/\mathbf{y})$, для которых средняя ошибка аппроксимации Δ_n не превосходит некоторого заданного значения ε:

$$\Omega_n(\varepsilon) = \{ w(\mathbf{y}/\mathbf{x}) : \Delta_n \le \varepsilon \}. \qquad (2.8.9)$$

Тогда функция от ε вида

$$H_n(\varepsilon) = \min_{w(\mathbf{y}/\mathbf{x})} \frac{1}{n} I\left(X^n; Y^n\right), \qquad (2.8.10)$$

где минимум разыскивается по всем условным плотностям $w(\mathbf{x}/\mathbf{y})$, принадлежащим классу $\Omega_n(\varepsilon)$, называется *эпсилон-энтропией источника X относительно меры аппроксимации d при передаче n-элементных последовательностей*.

Свойства эпсилон-энтропии для случайных процессов во многом совпадают со свойствами эпсилон-энтропии для случайных величин. В частности, нетрудно показать, что если в n-элементных последовательностях сообщения независимы, то эпсилон-энтропия (1.8.10) совпадает с эпсилон-энтропией (1.9.6), определённой для отдельных случайных величин.

Действительно, для независимых сообщений в последовательности $\mathbf{x}$ многомерное входное распределение $f(\mathbf{x})$ факторизуется:

$$w(\mathbf{x}) = \prod_{k=1}^{n} w\left(x^{(k)}\right) = w^n(x).$$

Тогда, исходя из свойства аддитивности для дифференциальной энтропии, взаимную информацию можно записать в виде

$$I\left(X^n; Y^n\right) = h\left(X^n\right) - h\left(X^n / Y^n\right) = \sum_{k=1}^{n} h\left(X_k\right) - \sum_{k=1}^{n} h\left(X_k / Y^k X_{k-1} \ldots X_1\right).$$

Здесь представление многомерного алфавита X^k ($k = 1, \ldots, n$) в виде произведения одномерных алфавитов $X_1, \ldots, X_k$ с различающимися индексами (и аналогично для многомерного алфавита Y^k) носит формальный характер; разумеется, в каждый момент времени имеет место один и тот же источник $X_1 = \ldots = X_k = X$.

Поскольку дифференциальная энтропия не уменьшается при исключении части условия,

$$h\left(X_k / Y_k \ldots Y_1 X_{k-1} \ldots X_1\right) \le h\left(X_k / Y_i\right)$$

для любых значений k, $i = 1, \ldots, n$. Отсюда

$$\frac{1}{n} I\left(X^n; Y^n\right) \ge \frac{1}{n} \sum_{k=1}^{n} \left[h\left(X_k\right) - h\left(X_k / Y_k\right) \right] = \frac{1}{n} \sum_{k=1}^{n} I_k(X;Y). \qquad (2.8.11)$$

Индекс k в обозначении взаимной информации $I_k(X; Y)$ означает, что она вычисляется для совместной плотности

$$w_k(x,y) = w\left(y^{(k)} / x^{(k)}\right) w(x) \equiv w_k(y/x) w(x),$$

которая, вообще говоря, может зависеть от k (на каждом шаге может быть своё отображение X в Y). Поэтому, сохраняя общность, применим свойство вогнутости взаимной информации относительно условных распределений. Полагая $\lambda_k = 1/n$ (сумма всех λ_k равна единице), продолжим предыдущее неравенство:

$$\frac{1}{n} I(X^n; Y^n) \ge \sum_{k=1}^{n} \lambda_k I_k(X;Y) \ge I_0(X;Y),$$

где $I_0(X; Y)$ — средняя взаимная информация, вычисленная по входной плотности $w(x)$ и условной плотности

$$w_0(y/x) = \frac{1}{n} \sum_{k=1}^{n} w_k(y/x).$$

При этом

$$\Delta_n = \frac{1}{n} \sum_{k=1}^{n} \mathbf{E}\left[d\left(x^{(k)}, y^{(k)}\right) \right] = \int_X \int_Y d\left(x,y\right) w(x) \frac{1}{n} \sum_{k=1}^{n} w_k\left(y/x\right) dxdy =$$

$$= \int_X \int_Y d\left(x,y\right) w(x) w_0(y/x) w(x) dxdy,$$

и если многомерная условная плотность $w(\mathbf{y}/\mathbf{x})$ входит в класс $\Omega_n(\varepsilon)$, то условная плотность $w_0(y/x)$ входит в класс $\Omega(\varepsilon)$.

Равенства (для нахождения минимума) в записанных неравенствах достигаются в том случае, когда также факторизуется и многомерная условная плотность $w(\mathbf{y}/\mathbf{x})$:

$$w(\mathbf{y}/\mathbf{x}) = \prod_{k=1}^{n} w_k\left(y/x\right) = w^n(y/x), \qquad (2.8.12)$$

т. е. когда п*а*ры источников (X_1Y_1), …, (X_nY_n) независимы и одинаково распределены. Таким образом, показано, что

$$H_n(\varepsilon) = \min_{w(\mathbf{y}/\mathbf{x})} \frac{1}{n} I(X^n;Y^n) = H(\varepsilon) = \min_{w(y/x)} I(X;Y). \qquad (2.8.13)$$

Аналогично одномерному случаю можно показать (см. задачу 2.8), что эпсилон-энтропия $H_n(\varepsilon)$ случайных процессов является выпуклой вниз функцией, равна нулю в некоторой области значений, ограничена для всех $n = 1, 2, \ldots$, а последовательность $\{H_n(\varepsilon)\}$ имеет предельное значение

$$\lim_{n\to\infty} H_n(\varepsilon) = H(\varepsilon).$$

В разд. 1.9 была найдена эпсилон-энтропия для гауссовской случайной величины. Решим теперь подобную задачу для гауссовского случайного процесса. Для этого в конечном итоге необходимо при найденном значении $H_n(\varepsilon)$ между n-элементными последовательностями совершить предельный переход (2.8.10).

Источник X генерирует в общем случае зависимые величины. Поэтому разделим решение поставленной задачи на несколько этапов, рассматривая вначале последовательность независимых гауссовских величин, затем — последовательность зависимых гауссовских величин и, наконец, предельные соотношения для случайного процесса дискретного времени.

Для последовательности независимых гауссовских случайных величин многомерная плотность вероятности $w(\mathbf{x})$ имеет вид

$$w(\mathbf{x}) = \prod_{k=1}^{n} w_k(x) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{x_k^2}{2\sigma_k^2}\right], \qquad (2.8.14)$$

где математическое ожидание каждой из величин равно нулю, что благодаря свойствам взаимной информации, никак не отражается на общности решаемой задачи.

Для нахождения эпсилон-энтропии согласно (2.8.11) и (2.8.13), используя независимость пар (X_1Y_1), …, (X_nY_n), исходное заданное значение ε, определяющее класс $\Omega_n(\varepsilon)$ для n-элементных последовательностей, необходимо заменить совокупностью значений ε_1, …, ε_n, которые определяют классы $\Omega(\varepsilon_1)$, …, $\Omega(\varepsilon_n)$ для одномерных величин и соответствующих взаимных информаций $I_k(X; Y)$ так, чтобы выполнялось условие

$$\frac{1}{n}\sum_{k=1}^{n}\varepsilon_k \le \varepsilon. \qquad (2.8.15)$$

При этом возникает двойная независимая минимизация

$$H_n(\varepsilon) = \min_{w(\mathbf{y}/\mathbf{x})} \frac{1}{n} I\left(X^n; Y^n\right) = \min_{\varepsilon_1,\ldots,\varepsilon_n} \min_{w_k(y/x)} \frac{1}{n}\sum_{k=1}^{n} I_k(X;Y) =$$

$$= \min_{\varepsilon_1,\ldots,\varepsilon_n} \frac{1}{n}\sum_{k=1}^{n} \min_{w_k(y/x)} I_k(X;Y) = \min_{\varepsilon_1,\ldots,\varepsilon_n} \frac{1}{n}\sum_{k=1}^{n} H(\varepsilon_k), \qquad (2.8.16)$$

когда сначала минимизируется взаимная информация $I_k(X; Y)$ по всем условным распределениям $w_k(y / x)$, а затем независимым образом ищется внешний минимум по $\varepsilon_1, \ldots, \varepsilon_n$. Поскольку

$$d_n(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{k=1}^{n} d\left(x^{(k)}, y^{(k)}\right) \le \frac{1}{n}\sum_{k=1}^{n}\varepsilon_k \le \varepsilon,$$

минимизация в каждом классе $\Omega(\varepsilon_k)$ в итоге приводит к минимизации в классе $\Omega_n(\varepsilon)$ для n-элементных последовательностей.

Как показано в разд. 1.9, эпсилон-энтропия $H(\varepsilon_k)$ гауссовской случайной величины, описываемой плотностью $w_k(x)$ согласно (2.8.14), равна

$$H(\varepsilon_k) = \frac{1}{2}\log\max\left\{1, \sigma_k^2/\varepsilon_k\right\},$$

так что

$$H_n(\varepsilon) = \min_{\varepsilon_1,\ldots,\varepsilon_n} \frac{1}{2n}\sum_{k=1}^{n}\log\max\left\{1, \sigma_k^2/\varepsilon_k\right\},$$

и возникает задача минимизации функции n переменных при наличии, на первый взгляд, одного ограничения (2.8.15).

Однако на самом деле имеют место скрытые ограничения-неравенства $\varepsilon_k \le \sigma_k^2$, поскольку ошибки аппроксимации не могут превосходить соответствующих дисперсий, ибо если для некоторого k окажется $\varepsilon_k > \sigma_k^2$, то тогда, положив $\varepsilon_k = \sigma_k^2$ и увеличив какую-то другую компоненту ε_i до значения $\varepsilon_i + \sigma_k^2 - \varepsilon_k$, можно было бы уменьшить эпсилон-энтропию (в пределе — до нуля) без увеличения общей ошибки аппроксимации, что противоречит и формальным свойствам, и здравому смыслу. Таким образом, формулировка задачи минимизации имеет следующий вид:

минимизировать функцию n переменных

$$H(\varepsilon_1,\ldots,\varepsilon_2)=\frac{1}{2n}\sum_{k=1}^{n}\log\frac{\sigma_k^2}{\varepsilon_k}, \qquad (2.8.17)$$

при наличии ограничений

$$\frac{1}{n}\sum_{k=1}^{n}\varepsilon_k\le\varepsilon,\ \varepsilon_k\le\sigma_k^2,\ k=1,\ \ldots,\ n.$$

Условие (2.8.15) является ограничением-неравенством, что, как правило, затрудняет решение оптимизационных задач. Поэтому, используя свойство монотонного невозрастания $H(\varepsilon)$ при увеличении ε, заменим (2.8.15) на ограничение-равенство

$$\frac{1}{n}\sum_{k=1}^{n}\varepsilon_k=\varepsilon.$$

Тогда совокупность ограничений для задачи максимизации имеет следующий вид:

$$\frac{1}{n}\sum_{k=1}^{n}\varepsilon_k=\varepsilon,\ \varepsilon_k\le\sigma_k^2,\ k=1,\ \ldots,\ n. \qquad (2.8.18)$$

Область B векторов $\boldsymbol{\varepsilon}=(\varepsilon_1,\ldots,\varepsilon_n)$ выпукла, поскольку для любых двух допустимых наборов (векторов) $\boldsymbol{\varepsilon}'=(\varepsilon_1',\ldots,\varepsilon_n')$, $\boldsymbol{\varepsilon}''=(\varepsilon_1'',\ldots,\varepsilon_n'')$ и любого числа $\alpha\in[0;1]$ результирующий вектор

$$\boldsymbol{\varepsilon}=\alpha\boldsymbol{\varepsilon}'+(1-\alpha)\boldsymbol{\varepsilon}''=\alpha(\varepsilon_1',\ldots,\varepsilon_n')+(1-\alpha)(\varepsilon_1'',\ldots,\varepsilon_n'')$$

также является допустимым (для его компонент выполняются условия (2.8.16)).

Функция, подвергаемая минимизации, есть сумма вогнутых (т. е. выпуклых вниз) функций и, следовательно, сама является вогнутой. Поэтому сформулированная задача является стандартной задачей минимизации вогнутой функции на выпуклом множестве векторов. Это множество, правда, не является множеством вероятностных векторов, но легко к нему приводится посредством нормировки (деления) компонент ε_k на величину ε. При этом, как нетрудно видеть (см. разд. 1.8), необходимые и достаточные условия экстремума (условия Куна–Таккера) остаются прежними. А именно, если удастся подобрать вектор $\boldsymbol{\varepsilon}_0=(\varepsilon_{01},\ldots,\varepsilon_{0n})$ и число μ такие, что для всех k

$$\frac{\partial H(\boldsymbol{\varepsilon})}{\partial \varepsilon_k}\begin{cases} = \mu, & \varepsilon_k < \sigma_k^2; \\ \geq \mu, & \varepsilon_k = \sigma_k^2, \end{cases} \tag{2.8.19}$$

то этот вектор и будет минимизировать функцию (2.8.17).

Рассмотрим вначале случай $\varepsilon_k < \sigma_k^2$, что соответствует нахождению вектора решения внутри допустимой области. Дифференцируя (2.8.17) по каждой из компонент, имеем $\varepsilon_{0k} = 2\mu/\log e$, и после подстановки в (2.8.18) получаем, что $\varepsilon_{0k} = \varepsilon$ ($k = 1, \ldots, l$). При этом сама эпсилон-энтропия равна

$$H(\varepsilon) = \frac{1}{2n}\sum_{k=1}^{n}\log\frac{\sigma_k^2}{\varepsilon} = \frac{1}{2}\left(\frac{1}{n}\sum_{k=1}^{n}\log\sigma_k^2 - \log\varepsilon\right). \tag{2.8.20}$$

Сравнение (2.8.20) с результатом (1.9.12), полученным для гауссовской случайной величины, показывает, что они отличаются только константами σ^2 и $(1/n)\sum_{k=1}^{l}\sigma_k^2$. При этом значение

$$\overline{\sigma}_k^2 = \frac{1}{n}\sum_{k=1}^{n}\sigma_k^2$$

можно трактовать как среднюю дисперсию последовательности независимых гауссовских величин.

Для того, чтобы удовлетворить (2.8.17) на границе B, выберем, аналогично тому, как это делалось для одномерного случая, в качестве компонент ε_{0k} сами значения дисперсий σ_k^2. При этом общий результат можно записать, используя оригинальную запись, предложенную А. Н. Колмогоровым: пусть θ — корень уравнения

$$\frac{1}{n}\sum_{k=1}^{n}\min\left\{\theta, \sigma_k^2\right\} = \varepsilon. \tag{2.8.21}$$

Тогда вектор решения может быть записан в виде

$$\varepsilon_{0k} = \begin{cases} \theta, & \theta < \sigma_k^2; \\ \sigma_k^2, & \theta = \sigma_k^2. \end{cases} \tag{2.8.22}$$

В таком виде нетрудно показать справедливость условий Куна–Таккера для $\boldsymbol{\varepsilon}_0$ и числа

$$\mu = -\frac{\log e}{2n\theta}.$$

Действительно,

$$\left.\frac{\partial H(\boldsymbol{\varepsilon})}{\partial \varepsilon_k}\right|_{\boldsymbol{\varepsilon}=\boldsymbol{\varepsilon}_0} = \begin{cases} -\dfrac{\log e}{2n\theta} = \mu, & \varepsilon_k < \sigma_k^2; \\ -\dfrac{\log e}{2n\varepsilon_{0k}} \geq -\dfrac{\log e}{2n\theta} = \mu, & \varepsilon_k = \sigma_k^2. \end{cases}$$

Итак, эпсилон-энтропия n-элементной последовательности независимых гауссовских случайных величин равна

$$H(\varepsilon) = \frac{1}{2n}\sum_{k=1}^{n} \log \max\left\{1, \sigma_k^2 / \theta\right\}, \qquad (2.8.23)$$

где θ определяется из уравнения (2.8.21).

Если при этом все случайные величины в последовательности одинаково распределены, то $\sigma_k^2 = \sigma^2$ (k = 1, …, l), и $H(\varepsilon)$, как и следовало ожидать, переходит в выражение (1.9.12) для одномерного случая.

Пусть теперь в рассматриваемой n-элементной последовательности не вся совокупность величин является независимой. В этом случае при вычислении эпсилон-энтропии необходимо декоррелировать элементы последовательности, что для гауссовских величин будет означать совокупную независимость, и после этого задача сводится к уже рассмотренному случаю.

Механизм декорреляции элементов последовательности был рассмотрен в разд. 1.7, и он основан на ортогональном преобразовании корреляционной матрицы $\mathbf{K}_x$ последовательности $\mathbf{x}$, приводящем её к диагональному виду. Если обозначить через $\hat{\mathbf{x}}$ результирующую декоррелированную последовательность, а через $\mathbf{Q}$ — матрицу необходимого ортогонального преобразования, то процесс ортогонализации можно записать в виде

$$\hat{\mathbf{x}} = \mathbf{x}\mathbf{Q}.$$

Далее, декоррелированная последовательность $\hat{\mathbf{x}}$ подвергается аппроксимации в классе $\Omega_n(\varepsilon)$ со средней ошибкой аппроксимации Δ_n, результатом чего является последовательность $\hat{\mathbf{y}}$, также имеющая некоррелированные компоненты. После этого производится обратное преобразование $\mathbf{Q}^T$, возвращающее (разумеется, приближённо, с учётом аппроксимации) исходную корреляцию элементов последовательности.

Ортогональное преобразование является обратимым, и, следовательно, взаимная информация между п*а*рами ($\mathbf{x}$, $\mathbf{y}$) и ($\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$) одинакова:

$$I(\mathbf{x};\mathbf{y}) = I(\hat{\mathbf{x}};\hat{\mathbf{y}}).$$

Кроме того, нетрудно видеть, что оно не изменяет среднюю ошибку аппроксимации Δ_n. Действительно, если записать Δ_n как

$$\Delta_n = \frac{1}{n}\mathbf{E}\left[(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^T\right],$$

то аналогичное выражение для ошибки аппроксимации $\hat{\Delta}_n$ между парой $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ имеет вид

$$\hat{\Delta}_n = \frac{1}{n}\mathbf{E}\left[(\hat{\mathbf{x}}-\hat{\mathbf{y}})(\hat{\mathbf{x}}-\hat{\mathbf{y}})^T\right].$$

Тогда

$$\hat{\Delta}_n = \frac{1}{n}\mathbf{E}\left[(\mathbf{xQ}-\mathbf{yQ})(\mathbf{xQ}-\mathbf{yQ})^T\right] = \frac{1}{n}\mathbf{E}\left[(\mathbf{x}-\mathbf{y})\mathbf{QQ}^T(\mathbf{x}-\mathbf{y})^T\right] =$$

$$= \frac{1}{n}\mathbf{E}\left[(\mathbf{x}-\mathbf{y})\mathbf{U}(\mathbf{x}-\mathbf{y})^T\right] = \frac{1}{n}\mathbf{E}\left[(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^T\right] = \Delta_n,$$

где $\mathbf{U} = \mathbf{QQ}^T$ — единичная матрица.

Таким образом, аппроксимация последовательности $\hat{\mathbf{x}}$ последовательностью $\hat{\mathbf{y}}$ со средней ошибкой аппроксимации $\hat{\Delta}_n$ приводит к аппроксимации последовательности $\mathbf{x}$ последовательностью $\mathbf{y}$ с той же самой ошибкой, т. е. класс $\Omega_n(\varepsilon)$ аппроксимации остаётся одним и тем же. Следовательно, эпсилон-энтропия $H_n(\varepsilon)$ последовательности $\mathbf{x}$ равна эпсилон-энтропии $\hat{H}_n(\varepsilon)$ последовательности $\hat{\mathbf{x}}$. При этом, дисперсии компонент $\hat{\mathbf{x}}$, полученные в результате ортогонального преобразования, равны собственным числам $\lambda_1, \ldots, \lambda_n$ матрицы $\mathbf{K}_x$. Подставляя их в (2.8.23), получим эпсилон-энтропию для произвольной гауссовской последовательности:

$$H(\varepsilon) = \frac{1}{2n}\sum_{k=1}^{n}\log\max\{1, \lambda_k/\theta\}, \qquad (2.8.24)$$

где в данном случае параметр θ определяется из уравнения

$$\frac{1}{n}\sum_{k=1}^{n}\min\{\theta, \lambda_k\} = \varepsilon. \qquad (2.8.25)$$

Итак, эпсилон-энтропия между произвольными n-элементными гауссовскими последовательностями найдена, и для окончательного решения задачи остаётся совершить предельный переход в (2.8.10) при $n \to \infty$. Строгое обоснование данного действия представляет из себя достаточно

сложную математическую задачу, требующую привлечения многих понятий, которые далеко выходят за рамки обсуждаемого материала. Более того, оказывается, что даже для гауссовских процессов сделать это в общем случае весьма затруднительно без наложения на них дополнительных свойств *стационарности* и *эргодичности*. Для того, чтобы не загромождать изложение материала излишней строгостью и подробностью, дадим "инженерное" толкование этих понятий, и на этом же уровне понимания попытаемся обосновать желаемый предельный переход.

Напомним, что процесс называется стационарным (в узком смысле), если все его конечномерные вероятностные характеристики не зависят от произвольного временн*о*го сдвига. Обычно при решении практических задач ограничиваются двумерными распределениями, поэтому зачастую требование инвариантности к произвольному временному сдвигу относят лишь к распределениям, не выше второго порядка, и такие процессы называют *стационарными в широком смысле*. Для процессов, стационарных хотя бы в широком смысле все моменты первого порядка, в частности, математическое ожидание и дисперсия, вообще не зависят от времени, а смешанные моменты второго порядка, в том числе, корреляционная функция, зависят лишь от разности временн*ы*х аргументов.

Понятие эргодичности, уже введённое в разд. 2.3 для дискретных марковских последовательностей, означает практическую идентичность характеристик, полученных при усреднении по ансамблю и по времени, однако, в отличие от дискретного случая, для непрерывных процессов это понятие гораздо шире. В частности, приходится различать понятие эргодичности по отношению к различным вероятностным характеристикам.

Пусть стационарный процесс $\xi(t)$, характеризующийся математическим ожиданием $\mathbf{E}[\xi(t)] = m$ и корреляционной функцией

$$b(t_1, t_2) = b(t_1 - t_2) \equiv b(\tau) = \mathbf{E}\left[\xi(t_1)\xi(t_2)\right],$$

наблюдается на временн*о*м интервале длительностью T. Тогда $\xi(t)$ называется *эргодическим* (*эргодичным*) *по отношению к среднему* (*математическому ожиданию*), *если*

$$\frac{1}{T}\int_0^T \xi(t)dt \underset{T\to\infty}{\to} m. \qquad (2.8.26)$$

Здесь, как обычно, предельный переход и интеграл понимаются в среднеквадратичном смысле, т. е.

$$\lim_{T\to\infty}\frac{1}{T}\mathbf{E}\left[\left(\int_0^T \xi(t)dt - m\right)^2\right] = 0\,.$$

Процесс $\xi(t)$ называется *эргодическим по отношению к корреляционной функции*, *если*

$$\frac{1}{T-\tau}\int_0^{T-\tau} \xi(t+\tau)\xi(t)dt \underset{T\to\infty}{\to} b(\tau)\,. \qquad (2.8.27)$$

Можно, разумеется, дать ещё определения эргодичности по отношению к другим вероятностным характеристикам, но они будут представлять, скорее, академический интерес.

Смысл определений (2.8.26) и (2.8.27) состоит в использовании статистических оценок m_T^* и $b_T^*(\tau)$ для наблюдаемых реализаций $x(t)$ процесса:

$$m_T^* = \frac{1}{T}\int_0^T x(t)dt\,,\ b_T^*(\tau) = \frac{1}{T-\tau}\int_0^{T-\tau} x(t+\tau)x(t)dt\,.$$

При этом оценка m_T^* оказывается состоятельной (т. е. дисперсия ошибки наблюдения стремится к нулю при $T\to\infty$) и несмещённой ($\mathbf{E}[m_T^*] = m$), так что для достаточно больших, но конечных значений T её можно использовать как эквивалент математического ожидания.

Аналогичные свойства справедливы и для оценки $b_T^*(\tau)$ корреляционной функции, однако, к сожалению, оказывается, что для неё не выполняется фундаментальное свойство положительной знакоопределённости, необходимое для совершения процедуры ортогонального преобразования. Поэтому на практике вместо (2.8.27) используют другую оценку

$$b_T^{**}(\tau) = \frac{1}{T}\int_0^{T-\tau} x(t+\tau)x(t)dt\,, \qquad (2.8.27a)$$

имеющую некоторое смещение, но зато характеризуемую положительной знакоопределённостью. При надлежащем соотношении между значениями T и τ оценка $b_T^{**}(\tau)$ оказывается вполне пригодной для практических целей.

На основании введённых понятий эргодичности существуют различные критерии эргодичности, среди которых одним из наиболее часто используемых является *критерий Слуцкого*.

Для того, чтобы стационарный процесс $\xi(t)$ *являлся эргодическим по отношению к среднему, необходимо и достаточно, чтобы выполнялось соотношение*[1]:

$$\frac{1}{T}\int_0^T (1-\tau)b(\tau)d\tau \underset{T\to\infty}{\to} 0. \qquad (2.8.28)$$

Из соотношения (2.8.28) вытекают *достаточные условия эргодичности*, в частности,

$$\frac{1}{T}\int_0^T b(\tau)d\tau \underset{T\to\infty}{\to} 0 \qquad (2.8.29)$$

и совсем простое

$$b(\tau) \underset{\tau\to\infty}{\to} 0. \qquad (2.8.30)$$

Соотношения (2.8.29) и (2.8.30) не являются необходимыми в общем случае для произвольных стационарных процессов. Однако можно показать (см. задачу 2.7), что стационарный гауссовский процесс с нулевым средним обладает свойством эргодичности и по отношению к среднему, и по отношению к корреляционной функции, и для них справедливы соотношения (2.8.29)–(2.8.30). Исходя из этого, при решении практических задач всегда принимается предположение, что *если измеряемый процесс стационарен, то он является эргодическим по отношению к среднему и к корреляционной функции, так что усреднение по ансамблю можно заменить усреднением по времени*.

Приведённые статистические соотношения основаны на предположении о том, что при наблюдении за случайным процессом $\xi(t)$ фиксируются полностью его континуальные реализации $x(t)$. Однако на практике фиксируются, как правило, не сами реализации, а их выборочные значения. Если в выборке X содержится $N+1$ значений $x_0, x_1, \ldots, x_N$, то формируются выборочное (точечное) среднее

$$\overline{x} = \frac{1}{N}\sum_{k=0}^{N} x_k \qquad (2.8.31)$$

и выборочная корреляционная (автокорреляционная) последовательность $\hat{\mathbf{b}}$, элементы которой равны

[1] Здесь и далее предельный переход понимается в обычном смысле.

$$\hat{b}_k = \frac{1}{N}\sum_{i=0}^{N-k} x_{i+k}x_i \,,\ k = 0,\ 1,\ \ldots,\ N. \qquad (2.8.32)$$

Заметим, что оценка $\hat{b}_k$ обладает некоторым смещением, хотя и не существенным при большом объёме выборки.

Элементы $\hat{b}_k$ автокорреляционной последовательности образуют матрицу $\hat{\mathbf{K}}$, являющуюся выборочным аналогом корреляционной матрицы для многомерных случайных величин:

$$\hat{\mathbf{K}} = \begin{pmatrix} \hat{b}_0 & \hat{b}_{-1} & \cdots & \hat{b}_{-N} \\ \hat{b}_1 & \hat{b}_0 & \cdots & \hat{b}_{-N+1} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{b}_N & \hat{b}_{N-1} & \cdots & \hat{b}_0 \end{pmatrix}, \qquad (2.8.33)$$

где элементы с отрицательными индексами являются комплексно-сопряжёнными по отношению к симметричным элементам с положительными индексами: $\hat{b}_{-k} = \hat{b}_k^*$ (т. е. для вещественных выборок эти элементы совпадают).

Вещественная матрица вида (2.8.33) называется *тёплицевой*[1] и обладает тем свойством, что все её элементы, расположенные на любой диагонали, идентичны, т. е.

$$\hat{K}_{ij} = \hat{K}_{(i\pm k)(j\pm k)},\ k \in \mathbb{Z}.$$

Может показаться, что наличие в матрице $\hat{\mathbf{K}}$ конечного числа элементов несколько противоречит тому допущению, что истинная корреляционная функция $b(\tau)$ случайного процесса может иметь бесконечную длительность (см. задачу 2.6.) однако, как уже было сказано выше, для эргодических процессов функция $b(\tau)$ убывает на бесконечности, и всегда можно указать такое значение τ_0, при котором $b(\tau)$ имеет практически нулевое значение. Обычно в качестве τ_0 берётся *время корреляции*, определяемое из условия

$$\tau_0 = \frac{1}{b(0)}\int_0^\infty b(\tau)d\tau\,. \qquad (2.8.34)$$

[1] в честь немецкого математика О. Тёплица

Теперь на основании значений из выборки X и сформированных статистических характеристик (2.8.31)–(2.8.33) можно ввести в рассмотрение *выборочную спектральную плотность средней мощности* $\hat{G}(\omega)$ *автокорреляционной последовательности*, определяемую на интервале [–ω/2; ω/2] как

$$\hat{G}(\omega) = \sum_{k=-N}^{N} \hat{K}_k \exp\left(-jk\omega\right). \qquad (2.8.35a)$$

Соотношение (2.8.35*а*) представляет собой усечённый ряд Фурье, в котором коэффициенты $\hat{K}_k$ равны

$$\hat{K}_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{G}(\omega)\exp\left(jk\omega\right)d\omega. \qquad (2.8.35б)$$

В частности, коэффициент с нулевым индексом

$$\hat{K}_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{G}(\omega)d\omega$$

представляет среднюю мощность дискретного процесса, представленного выборкой X.

Нетрудно видеть, что равенства (2.8.35) представляют собой не что иное, как дискретный статистический аналог *формул Винера–Хинчина*, которые связывают между собой истинные корреляционную функцию $b(\tau)$ и энергетический спектр $G(\omega)$ стационарного случайного процесса парой взаимообратных преобразований Фурье:

$$\begin{aligned} G(\omega) &= \int_{-\infty}^{\infty} b(\tau)\exp\left(-j\omega\tau\right)d\tau; \\ b(\tau) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} G(\omega)\exp\left(j\omega\tau\right)d\omega. \end{aligned} \qquad (2.8.36)$$

Вернёмся к задаче о нахождении эпсилон-энтропии гауссовского процесса, по отношению к которому будем предполагать выполнение свойства эргодичности.

В основе совершения предельного перехода при $n \to \infty$ лежит следующее фундаментальное свойство, связывающее собственные числа корреляционной функции и энергетический спектр процесса.

Если $\left\{\lambda_k^{(n)}\right\}$ *— набор собственных чисел корреляционной матрицы n-го порядка, полученной для n-элементного отрезка стационарного слу-*

чайного процесса с дискретным временем и характеризуемого энергетическим спектром $G(\omega)$, *то для произвольной числовой функции* g *справедливо соотношение*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} g\left(\lambda_k^{(n)}\right)=\frac{1}{2\pi}\int_{-\infty}^{\infty} g\left(G(\omega)\right)d\omega. \qquad (2.8.37)$$

Точная формулировка и доказательство этого свойства достаточно громоздки и утомительны; интересующиеся математическими подробностями читатели могут найти необходимый материал в соответствующей литературе по тёплицевым матрицам.

Считая теперь, что функция g — это фигурирующая в (2.8.4) функция

$$\frac{1}{2}\log\max\left\{1,\lambda_k/\theta\right\},$$

получаем искомое выражение для эпсилон-энтропии гауссовского процесса с энергетическим спектром $G(\omega)$:

$$H(\varepsilon)=\frac{1}{4\pi}\int_{-\infty}^{\infty}\log\max\left\{1,G(\omega)/\theta\right\}d\omega, \qquad (2.8.38)$$

где уровень ограничения θ определяется соотношением

$$\varepsilon=\frac{1}{2\pi}\int_{-\infty}^{\infty}\min\left\{\theta,G(\omega)\right\}d\omega. \qquad (2.8.39)$$

При этом интеграл от всего энергетического спектра даёт среднюю мощность (дисперсию):

$$P\equiv\sigma^2=\frac{1}{2\pi}\int_{-\infty}^{\infty}G(\omega)d\omega. \qquad (2.8.40)$$

Обычно соотношения (2.8.38)–(2.8.39) записывают с использованием линейной частоты:

$$H(\varepsilon)=\frac{1}{2}\int_{-\infty}^{\infty}\log\max\left\{1,G(f)/\theta\right\}df, \qquad (2.8.38a)$$

$$\varepsilon=\int_{-\infty}^{\infty}\min\left\{\theta,G(f)\right\}df. \qquad (2.8.39a)$$

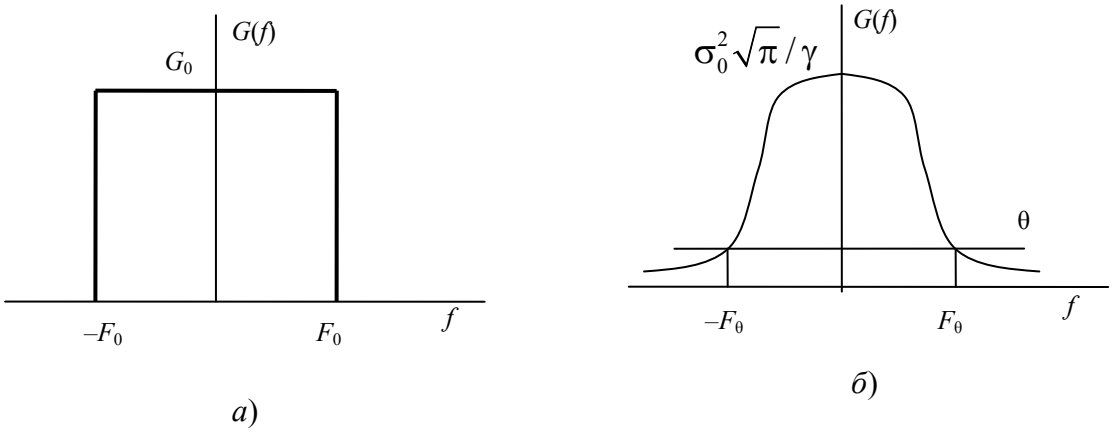
Уровень ограничения θ определяет конечную полосу частот F_θ, которая для частного случая монотонно убывающей функции спектральной плотности мощности (рис. 2.5) находится из уравнения

$$G(F_\theta) = \theta\,. \tag{2.8.41}$$

Тогда расчётные соотношения несколько упрощаются:

$$H(\varepsilon) = \int_0^{F_\theta} \log\bigl(G(f)/\theta\bigr)df\,, \tag{2.8.42}$$

$$\varepsilon = 2\theta F_\theta + 2\int_{F_\theta}^{\infty} G(f)df\,. \tag{2.8.43}$$



а) *б*)

Рис. 2.5. К расчёту ε-энтропии непрерывных сигналов: *а*) спектр с прямоугольной формой; *б*) спектр с экспоненциальной формой

В том случае, когда спектр сигнала имеет строго прямоугольную форму (рис. 2.5, *а*)

$$G(f) = \begin{cases} G_0, & |f| \le F_0 \\ 0, & |f| > F_0 \end{cases},$$

полагая $F_\theta = F_0$, имеем мощность $P = \sigma^2 = 2G_0F_\theta$, а из (2.8.43) $\theta = \varepsilon/(2F_\theta)$, и после подстановки в (2.8.42) получаем соотношение

$$H(\varepsilon) = F_0 \log\bigl(\sigma^2/\varepsilon\bigr), \tag{2.8.44}$$

которое очень похоже на выражение (1.9.12) определяющее эпсилон-энтропию одномерного гауссовского распределения при $\sigma^2 \ge \varepsilon$ с той лишь разницей, что вместо множителя 1/2 фигурирует значение полосы сигнала F_0. Этому можно дать трактовку с позиции теоремы Котельникова как представление гауссовского непрерывного сигнала с равномерным в полосе F_0 спектром последовательностью некоррелированных (а для гауссовского источника — и независимых) отсчётов, следующих со скоростью $v = 2F_0$.

Заметим при этом, что в соотношениях (1.9.12) и (2.8.44) отличается размерность эпсилон-энтропии: в первом случае — это размерность взаимной информации, например, в двоичных единицах (битах), а во втором — количество (взаимной) информации в единицу времени, т. е. в битах в секунду. Исходя из этого, эпсилон-энтропию, фигурирующую в (2.8.44), часто называют *производительностью источника непрерывных сообщений при заданном критерии качества* или *скоростью создания информации непрерывным источником при заданном критерии качества* (далее при рассмотрении передачи информации по каналам связи понятие производительности источника будет рассмотрено подробнее).

Если переобозначить через R производительность источника в (2.8.44) и сохранить обозначение эпсилон-энтропии в (1.9.12), то, считая по-прежнему, что $\sigma^2 \geq \varepsilon$, можно записать:

$$R = 2F_0 H(\varepsilon) = F_0 \log\left(\sigma^2 / \varepsilon\right),$$

откуда получим значение минимально возможной среднеквадратической ошибки восстановления гауссовского сигнала с равномерным финитным спектром:

$$\varepsilon = \sigma^2 2^{-R/F_0}. \qquad (2.8.45)$$

Видно, что при $R = 0$ ошибка максимальна и равна σ^2. При увеличении R становится возможным все более точное восстановление, и в пределе, когда $R \to \infty$, можно достичь сколь угодно малой ошибки восстановления.

Рассмотрим теперь стационарный сигнал, корреляционная функция которого равна

$$b(\tau) = \sigma^2 \exp\left(-\gamma^2 \tau^2\right), \qquad (2.8.46a)$$

где σ^2 — дисперсия сигнала, а параметр γ определяет вероятностные связи во временно*й* области: при $\tau = 1/\gamma$ корреляционная функция убывает в e раз. Нетрудно видеть, что энергетический спектр такого сигнала согласно преобразованиям Винера–Хинчина (2.8.36) также имеет экспоненциальный вид (рис. 2.5, *б*):

$$G(f) = \frac{\sigma^2 \sqrt{\pi}}{\gamma} \exp\left(-\frac{\pi^2 f^2}{\gamma^2}\right). \qquad (2.8.46б)$$

Заметим, что процесс с характеристиками (2.8.46) нетрудно смоделировать, для этого достаточно генерировать независимые импульсы гаус-

совской формы и с гауссовским распределением амплитудных значений, поскольку в таких случаях форма усреднённого энергетического спектра совпадает (с точностью до множителя) с формой спектра одиночных импульсов [11].

Задавшись уровнем ограничения θ, из (2.8.43) получим связь между значениями ε и F_θ в виде трансцендентного уравнения

$$\varepsilon = 2\theta F_\theta + \sigma^2 \operatorname{erfc}\left(\pi\sqrt{2}F_\theta / \gamma\right) =$$

$$= 2F_\theta\left(\sigma^2\sqrt{\pi}/\gamma\right)\exp\left(-\pi^2 F_\theta^2/\gamma^2\right) + \sigma^2\operatorname{erfc}\left(F_\theta\pi/\gamma\right), \qquad (2.8.47)$$

где $\operatorname{erfc}(z) = \left(2/\sqrt{\pi}\right)\int_z^\infty \exp\left(-t^2\right)dt$ — *дополнительная функция ошибок*[1].

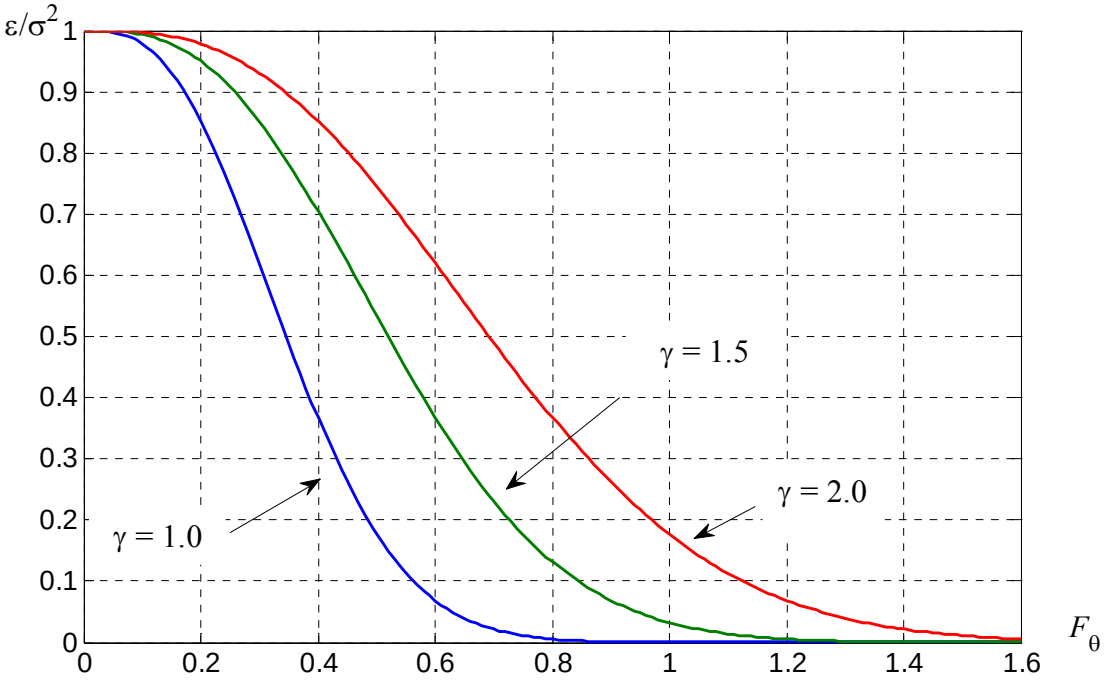


Рис. 2.6. Зависимость ε / σ² от F_θ для сигналов с гауссовским спектром

Теперь, подставив в (2.8.42) значение

$$\theta = \left(\sigma^2\sqrt{\pi}/\gamma\right)\exp\left(-\pi^2 F_\theta^2/\gamma^2\right),$$

и проинтегрировав, получим, что

[1] Эта функция дополняет до единицы обычную функцию ошибок:

$$\operatorname{erfc}(z) = 1 - \operatorname{erf}(z) = 1 - \left(2/\sqrt{\pi}\right)\int_0^z \exp\left(-t^2\right)dt\,.$$

$$H(\varepsilon)=\frac{2}{3}\frac{\pi^2}{\gamma^2}F_\theta^3(\varepsilon)\,. \qquad (2.8.48)$$

На рис. 2.6 представлена зависимость ε/σ^2 от частоты F_θ в соответствии с (2.8.47) при различных значениях параметра γ. Видно, что при больших F_θ значения ε/σ^2 стремятся к нулю, что соответствует снижению уровня ограничения θ. При $F_\theta \to 0$ значение ε/σ^2 независимо от γ оказывается равным единице. Это и понятно, поскольку в этом случае

$$\varepsilon=\int_{-\infty}^{\infty}\min\{\theta,G(f)\}\,df\equiv\int_{-\infty}^{\infty}G(f)df=\sigma^2\,.$$

Теперь, задаваясь значениями ε (или нормированными значениями ε/σ^2) можно по зависимости (2.8.47) определить соответствующие значения F_θ, а затем — согласно (2.8.48) — искомые значения ε-энтропии.

На рис. 2.7 представлена зависимость ε-энтропии от нормированного аргумента ε/σ^2 на интервале изменения ε от 0 до σ^2 для некоторых значений параметра γ. В тех случаях, когда возможна "угроза" $\varepsilon \geq \sigma^2$, как обычно, для реконструкции сигнала достаточно воспроизвести статистически независимые выборки аппроксимирующего гауссовского источника с дисперсией $\varepsilon - \sigma^2$, имея гарантию того, что среднеквадратическая ошибка воспроизведения вновь станет меньше σ^2.

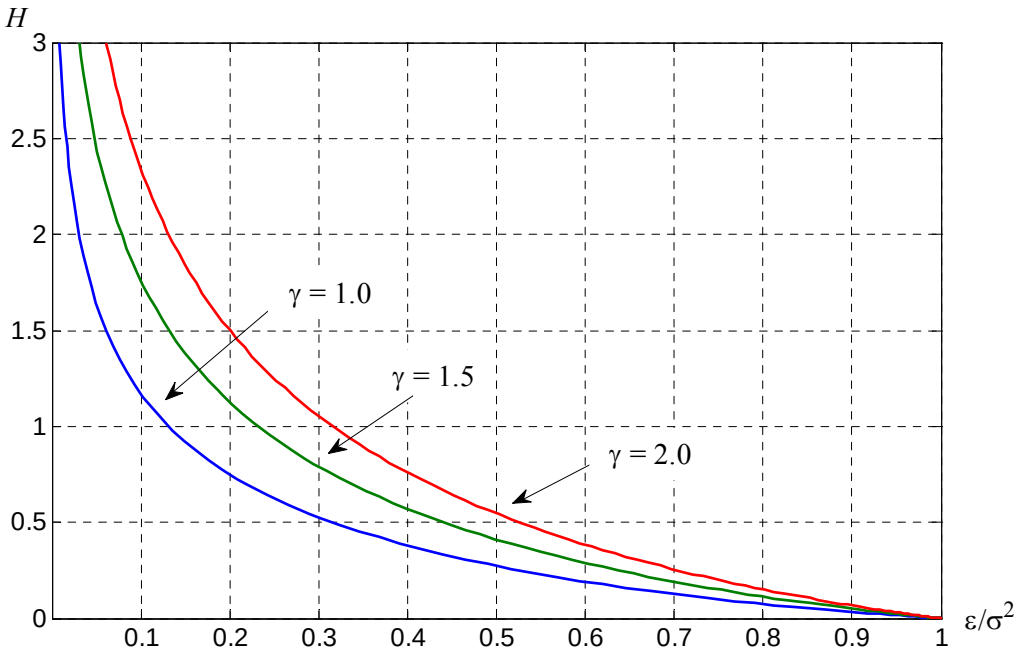


Рис. 2.7. Зависимость ε-энтропии от нормированного аргумента ε/σ^2

Представляет интерес сравнение ε-энтропии, определяемой (2.8.48), с аналогичной зависимостью для гауссовских сигналов с прямоугольной формой энергетического спектра. Для этого, прежде всего, необходимо определить условную граничную частоту[1] F_0 гауссовского сигнала.

Выберем критерий концентрации энергии (средней мощности), согласно которому в качестве граничной частоты F_0 принимается значение, обеспечивающее заданную величину

$$\eta = \frac{\int_0^{F_0} G(f)df}{\int_0^{\infty} G(f)df} = \mathrm{erf}\left(\frac{\pi F_0}{\gamma}\right),$$

т. е.

$$F_0 = \frac{\gamma}{\pi}\mathrm{erf}^{-1}(\eta), \qquad (2.8.49)$$

где $\mathrm{erf}^{-1}(z)$ обозначает функцию, обратную $\mathrm{erf}(z)$. Практически значения η должны находиться в пределах 0,90…0,95, поэтому граничная частота F_0 составляет (0,37…0,44)γ.

Теперь можно записать (2.8.48) в форме, аналогичной (2.8.44):

$$H(\varepsilon) = F_0 \frac{2\pi^3}{3\gamma^3 \mathrm{erf}^{-1}(\eta)} F_\theta^3(\varepsilon). \qquad (2.8.50)$$

На рис. 2.8 представлены зависимости нормированной (к граничной частоте F_0) эпсилон-энтропии от нормированной (к дисперсии сигнала σ^2) среднеквадратической ошибки воспроизведения для гауссовских сигналов с прямоугольной формой энергетического спектра (кривая 3), а также с гауссовской формой при двух значениях концентрации энергии: η = 0,90 (кривая 2) и η = 0,95 (кривая 1). Как и следовало ожидать, переход от прямоугольной к гауссовской форме приводит к снижению эпсилон-энтропии. Для сигналов с гауссовской формой наблюдается некоторая чувствительность к параметру η: бо́льшая для малых значений ε и меньшая — для значений ε/σ^2 вблизи единицы.

[1] Необходимо различать условную частоту F_0, определяющую ширину спектра сигнала, и частоту F_θ, значение которой зависит от уровня ограничения θ, определяемого, в свою очередь, выбранным значением погрешности ε. Фактически F_θ — это некоторая промежуточная переменная, необходимая для представления зависимости $H(\varepsilon)$ от ε/σ^2.

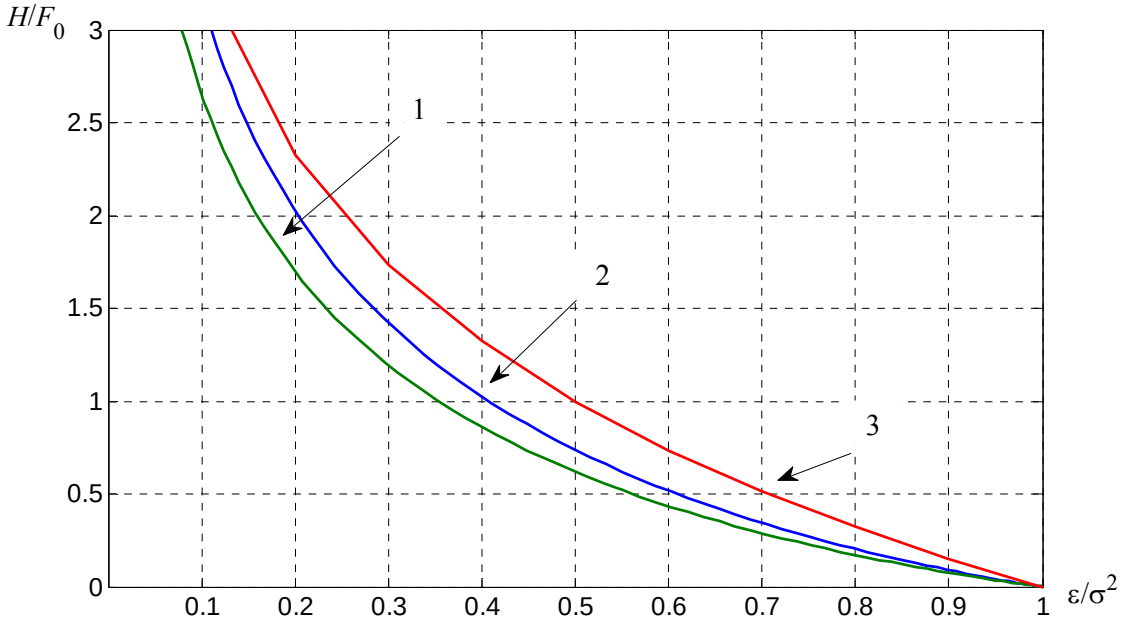


Рис. 2.8. Нормированная эпсилон-энтропия для гауссовских сигналов с различной формой энергетического спектра

Рассмотренные примеры касались случая монотонного убывания энергетического спектра сигналов с ростом частоты. Если поведение спектра немонотонно, например, как показано на рис. 2.9, то нахождение эпсилон-энтропии несколько усложняется, хотя и непринципиально.

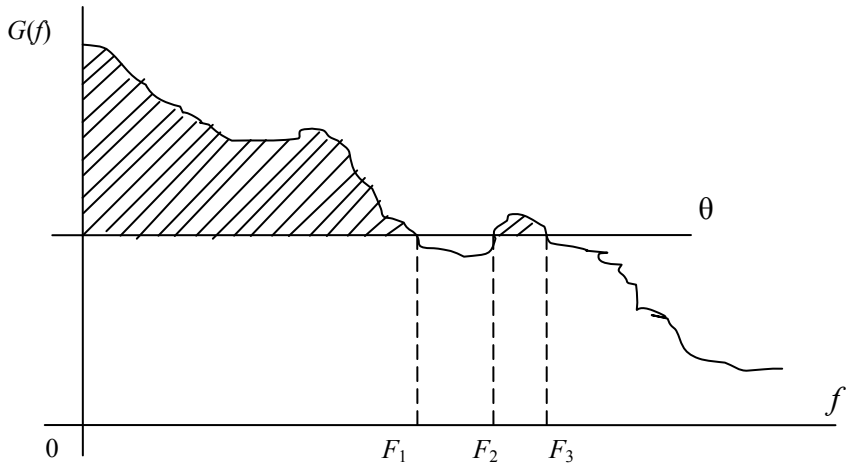


Рис. 2.9. К вычислению. ε-энтропии для сигналов с немонотонным убыванием энергетического спектра

Обычно методике расчёта ε-энтропии дают следующую геометрическую трактовку с помощью рассуждений “о разливании воды в сосуде”. Представим себе цилиндрический сосуд с крышкой, сечение которого задаётся формой энергетического спектра $G(f)$. Тогда жидкость, полный объем которой равен ε, после опускания крышки установится на уровне θ, и для нахождения эпсилон-энтропии необходимо проинтегрировать функцию $\log G(f)/\theta$ по той области частот, в которой $G(f) \geq \theta$. Так, на рис. 2.9 это надо сделать по интервалам $[0; F_1]$ и $[F_2; F_3]$, удваивая затем результат для учёта симметричной отрицательной области.

2.9. ИНФОРМАЦИОННЫЕ ХАРАКТЕРИСТИКИ ИСТОЧНИКОВ РЕЧЕВЫХ СООБЩЕНИЙ

В разд. 1.3 были приведены некоторые простые примеры, касающиеся информационных характеристик сообщений. Обратимся к реальным типам сообщений и рассмотрим некоторые информационные характеристики письменной и устной речи.

Проводя анализ информационных характеристик письменной и устной речи, ограничимся сообщениями на русском и английском языках. При этом статистическим аналогом распределения вероятностей сообщений является частота встречаемости отдельных букв (или элементов речи), определяемая как отношение количества появлений каждой буквы к общему числу букв в сообщении.

Таблица 2.2

Частоты встречаемости букв русского языка

–	О	Е, Ё	А	И	Т	Н	С
0,175	0,090	0,072	0,062	0,062	0,053	0,053	0,045
Р	В	Л	К	М	Д	П	У
0,040	0,038	0,035	0,028	0,026	0,025	0,023	0,021
Я	Ы	З	Ь, Ъ	Б	Г	Ч	Й
0,018	0,016	0,016	0,014	0,014	0,013	0,012	0,010
Х	Ж	Ю	Ш	Ц	Щ	Э	Ф
0,009	0,007	0,006	0,006	0,003	0,003	0,003	0,002

В табл. 2.2 и 2.3 приведены усреднённые значения частоты встречаемости букв русского и английского алфавитов, включая пробел (–) между словами. При этом для русского алфавита буквы Е и Ё, а также Ъ и Ь не различаются.

Таблица 2.3

Частоты встречаемости букв английского языка

–	E	T	O	A	N	I	R	S
0,200	0,105	0,072	0,064	0,063	0,059	0,055	0,054	0,052
H	D	L	C	F	U	M	P	Y
0,047	0,035	0,029	0,023	0,022	0,022	0,021	0,018	0,012
W	G	B	V	K	X	J	Q	Z
0,012	0,011	0,009	0,007	0,003	0,002	0,001	0,001	0,001

Следует заметить, что представленные в табл. 2.2 и 2.3 данные являются усреднёнными по большому количеству источников и могут значительно различаться, если рассматривать произведения различных авторов, пишущих в различных стилях (художественная проза, поэзия, документы, научно-техническая литература и др.).

В качестве примера можно назвать начало стихотворения К.Д. Бальмонта “Камыши”

Полно*ч*ной порою в болотной глу*ш*и
*Ч*уть слы*ш*но, бес*ш*умно *ш*урш*а*т камы*ш*и…,

целиком построенное на обыгрывании шипящих звуков *ч* и *щ*, средняя частота встречаемости которых в русском языке весьма мала.

Готтлоб Бурманн, немецкий поэт (1737–1805), написал 130 поэм, содержащих 20000 слов, и ни разу при этом не использовал букву г. Более того, в последние семнадцать лет своей жизни Бурманн вовсе изъял эту букву из своей речи.

В 1939 г. Эрнест Винсент Райт опубликовал роман “Гэтсби”. На 267 страницах этой книги ни разу не была использована наиболее частая буква английского языка e. Ниже процитирован один абзац этого из романа.

Upon this basis I am going to show you how a bunch of bright young folks did find a champion; a man with boys and girls of his own; a man of so dominating and happy individuality that Youth is drawn to him as is a fly to a sugar bowl. It is a story about a small town. It is not a gossipy yarn; nor is it a dry, monotonous account, full of such customary “fill-ins” as “romantic moonlight casting murky shadows down a long, winding country road”. Nor will it say anything about tinklings lulling distant folds; robins caroling at twilight, nor any “warm glow of lamplight” from a cabin window. No. It is an account of up-and-doing activity; a vivid portrayal of Youth as it is today; and a practical discarding of that wornout notion that “a child don’t know anything”.

Обратимся к значениям частот встречаемости букв русской речи. Если считать, что все буквы выбираются независимо друг от друга, то энтропия

русского алфавита согласно (1.3) составляет $H = 4{,}357$ бит, в то время как максимальная энтропия источника с таким же объёмом $l = 32$ равна $H_{\max} = \log_2 32 = 5{,}000$. Следовательно, избыточность источника в данном случае составляет

$$\rho = 1 - \frac{4{,}357}{5{,}000} = 0{,}129 .$$

Учёт вероятностной зависимости появления букв в осмысленном тексте при вычислении соответствующих значений энтропии приводит к значительным трудностям, тем большим, чем больше глубина зависимости. Для вычисления значений энтропии $H^{(n)}$ ($n > 1$) литературных источников, описываемых цепью Маркова n-го порядка необходимо иметь таблицы частот встречаемости различных многобуквенных сочетаний. В табл. 2.4 приведены наиболее частые трехбуквенные сочетания в русском языке.

Таблица 2.4

Частоты встречаемости некоторых трёхбуквенных сочетаний

Сочетание	Частота 10^{-4}	Сочетание	Частота 10^{-4}	Сочетание	Частота 10^{-4}
_и_	82	го_	50	но_	40
_не	71	_он_	49	_бы	39
то_	68	_в_	45	ли_	36
на_	65	не_	45	ала	31
_по	59	что	44	ом_	31
ть_	56	_чт	41	ост	30
_на	55	_пр	41	_ко	30
ла_	51	_то	40	его	29

Величины энтропий $H^{(2)}$ и $H^{(3)}$ для английского языка подсчитаны ещё К. Шенноном на основании имеющихся таблиц частот двух- и трёхбуквенных сочетаний. Кроме того, поскольку средняя длина английского слова составляет 4–5 букв, располагая данными о частотах появления различных слов в английском языке, Шеннон приближённо оценил и значения величины $H^{(5)}$. В результате им был получен ряд чисел, представленных в табл. 2.5.

Таблица 2.5

Значения энтропий различного приближения

$H^{(0)}$	$H^{(1)}$	$H^{(2)}$	$H^{(3)}$	$H^{(5)}$
4,76	4,03	3,32	3,10	2,10

Из данных, приведённых в табл. 2.5, можно заключить, что избыточность английского языка не меньше, чем 0,56. Оценка величин $H^{(n)}$ при больших значениях n возможна, по-видимому, только косвенными методами. Рассмотрим один из таких методов, предложенный самим К. Шенноном [1].

Условная энтропия $H^{(n)}$ представляет собой меру степени неопределённости опыта α_n, состоящего в определении n-й буквы текста, при условии, что предыдущие $(n-1)$ букв известны. Одним из возможных практических способов определения степени такой неопределённости является, как это не парадоксально, простое угадывание букв. Эксперимент по отгадыванию n-й буквы может быть легко поставлен: для этого достаточно выбрать $(n-1)$-буквенный отрывок осмысленного текста и предложить кому-либо отгадать следующую букву. Подобный опыт может быть повторен многократно. При этом трудность отгадывания n-й буквы оценивается с помощью среднего значения Q числа попыток, требующихся для нахождения правильного ответа. Очевидно, что среднее число попыток Q с возрастанием n может только уменьшаться; прекращение этого уменьшения будет свидетельствовать о том, что соответствующая им условная энтропия $H^{(n)}$ достигла предельного значения $H^{(\infty)}$.

Исходя из этих простых, но конструктивных соображений, К. Шеннон произвёл ряд подобных экспериментов, в которых n принимало значения 1, 2, …, 14, 15 и 100. При этом он обнаружил, что отгадывание 100-й буквы по 99 предыдущим является заметно более простой задачей, чем отгадывание 15-й буквы по 14 предыдущим. Отсюда можно сделать вывод, что $H^{(15)}$ ощутимо больше, чем $H^{(100)}$, т. е. условную энтропию $H^{(15)}$ ещё нельзя отождествить с предельным значением $H^{(\infty)}$.

Впоследствии аналогичные опыты были проведены для $n = 1, 2, 4, 8, 16, 32, 64$ и 128. Из результатов этих опытов следовало, что величины $H^{(32)}$, $H^{(64)}$ и $H^{(128)}$ практически не отличаются, в то время как условная энтропия $H^{(16)}$ ещё заметно больше этих величин. Таким образом, можно предполо-

жить, что величина $H^{(n)}$ убывает вплоть до значений $n \approx 30$ и при дальнейшем увеличении n практически не меняется.

Эксперименты по отгадыванию букв не только позволяют судить о сравнительной величине $H^{(n)}$, но также дают возможность оценить их значение на основании верхней и нижней границ:

$$\sum_{k=1}^{l} q_n^k \log q_n^k \le H^{(n)} \le \sum_{k=1}^{l} k\left(q_n^k - q_n^{k+1}\right)\log k\,, \qquad (2.9.1)$$

где q_n^1, q_n^2, …, q_n^l — вероятности (частоты) того, что буква будет правильно угадана с 1-й, 2-й, …, l-й попытки (l — число букв в алфавите), при этом полагается, что $q_n^{l+1} = 0$.

Однако указанные границы, к сожалению, не сближаются неограниченно при увеличении n (более того, как уже было сказано, начиная с $n \approx 30$, они вообще перестают от n зависеть), поэтому желательно иметь более точные методы определения $H^{(n)}$. В этой связи представляет интерес предложенное А. Н. Колмогоровым дальнейшее развитие метода отгадывания. Идея такого метода состоит в том, что отгадывающему предлагается каждый раз не перечислять по порядку те буквы, которые, как он думает, должны появиться, а сразу назвать все условные вероятности p_1^n, p_2^n, …, p_l^n того, что появится 1-я, 2-я, …, l-я буква алфавита, при условии, что предшествующие $n - 1$ букв известны. Если этот опыт повторить много раз, подсчитывая при этом величину $-\log p_i^n$ ($i = 1, \ldots, l$), то среднее значение подсчитываемой величины $-\log p_i^n$ (т. е. сумма всех таких величин, определённых в большом числе M опытов, делённая на M) при неограниченном увеличении M будет приближаться к величине $H^{(n)}$.

Разумеется, в такой постановке этот метод практически нереализуем, поскольку невозможно требовать от отгадывающего, чтобы он каждый раз указывал весь набор условных вероятностей появления всевозможных букв и при этом никогда не ошибался. Однако следует заметить, что любые ошибки в значениях таких вероятностей приводят лишь к возрастанию соответствующей суммы значений $-\log p_i^n$. Поэтому вполне допустимо заранее ограничить множество распределений вероятностей, которые может назвать отгадывающий, а сумма полученных таким образом значений $-\log p_i^n$, разделённая на число M опытов, будет оценкой сверху величины $H^{(n)}$.

В реальных опытах отгадывающему предлагалось делать следующие предсказания:

1. *Следующей буквой наверняка будет одна определённая i-я буква алфавита*.

2. *Следующей буквой наверняка будет одна из указываемых отгадывающим двух или трёх букв алфавита.*

3. *Следующей буквой возможно (но не наверняка) будет одна определённая i-я буква алфавита.*

4. *Следующей буквой возможно (но не наверняка) будет одна из указываемых отгадывающим двух или трёх букв алфавита.*

5. *Нельзя сказать, какой будет следующая буква алфавита.*

При этом считалось, что каждое из этих утверждений равносильно выбору следующего условного распределения вероятностей для следующей буквы текста:

1. *Некоторая i-я буква имеет фиксированную относительно большую вероятность p; для других k-х букв* $\left(k \neq i\right)$ *вероятность их появления* p'_k *принимается равной* $p'_k = p_k\left(1-p\right)/\left(1-p_i\right)$, *где* p_i *и* p_k *— безусловные вероятности появления i-й и k-й букв.*

2. *Выбранные две или три буквы имеют одинаковую условную вероятность p/2 или p/3; остальные буквы имеют указанные вероятности* p'_k.

3. *Некоторая i-я буква имеет фиксированную вероятность q < p; для других k-х букв* $\left(k \neq i\right)$ *вероятность их появления* p'_k *принимается равной* $p'_k = p_k\left(1-q\right)/\left(1-p_i\right)$.

4. *Выбранные две или три буквы имеют одинаковую условную вероятность p/2 или p/3, а остальные буквы — вероятности, пропорциональные их безусловной вероятности.*

5. *Условная вероятность появления k-й буквы принимается равной её безусловной вероятности для всех значений k.*

Неизвестные вероятности p и q могут быть подобраны по известным результатам опытов так, чтобы сумма всех величин $-\log p_i^n$, где p_i^n — предсказанная условная вероятность реально появившейся буквы, была бы возможно меньшей.

Окончательная оценка условной энтропии $H^{(n)}$ при проведении M опытов, в которых M_1 — число предсказаний 1 или 2, M_2 — число пред-

сказаний 3 или 4, M_1' — число предсказаний 2 или 4, в которых предсказывается одна из двух букв, M_2' — число предсказаний 2 или 4, в которых предсказывается одна из трёх букв, даётся формулой:

$$H^{(n)} = \left(1/M\right)\left(M_1 h_1 + M_2 h_2 + M_1' + M_2' \log 3 + S\right), \qquad (2.9.2)$$

где

$$h_1 = \left(m_1/M_1\right)\log\left(m_1/M_1\right) - \left(1 - m_1/M_1\right)\log\left(1 - m_1/M_1\right),$$

$$h_2 = \left(m_2/M_2\right)\log\left(m_2/M_2\right) - \left(1 - m_2/M_2\right)\log\left(1 - m_2/M_2\right),$$

а m_1 и m_2 — число ошибок в предсказаниях 1 или 2 и 3 или 4 соответственно. Величина *S* есть сумма величин $-\log p_k''$, где p_k'' — или безусловная вероятность p_k реально появившейся буквы (в случае предсказания 5), или предсказанная вероятность p_k', разделённая на $1-p$ (в случае предсказаний 1 или 2), или, наконец, предсказанная вероятность p_k', разделённая на $1-q$ (в случае предсказаний 3 или 4).

Проводившийся по предложенной методике анализ произведений русской прозы С. Т. Аксакова (“Детские годы Багрова-внука”) и И. А. Гончарова (“Литературный вечер”) позволил дать не отличающуюся от $H^{(50)}$ оценку $H^{(\infty)}$ в пределах 1,0–1,2 бит. Соответственно этому, для избыточности литературного языка русской классической прозы получается значение, близкое к 80%.

Разумеется, само применение к литературным текстам стандартных теоретико-информационных представлений, базирующихся на вероятностно-статистических понятиях и игнорирующих вопрос о смысловом содержании сообщения, вызывает известные сомнения. А.Н. Колмогоров, ставя вопрос о возможности различных подходов к понятию количества информации, писал: “*...Практически можно считать, например, вопрос об* “*энтропии*” *потока поздравительных телеграмм ... корректно поставленным в его вероятностной трактовке и при обычной замене вероятностей эмпирическими частотами... Но какой реальный смысл имеет, например, говорить о* “*количестве информации*”*, содержащейся в тексте* “*Войны и мира*”? *Можно ли включить разумным образом этот роман в совокупность* “*возможных романов*”*, да ещё постулировать существование в этой совокупности некоторого распределения вероятностей*? *Или следу-*

ет считать отдельные сцены "Войны и мира" образующими случайную последовательность с достаточно быстро затухающими на расстоянии нескольких страниц "стохастическими связями" ? ".

Сущность предложенного Колмогоровым *комбинаторного* подхода к определению энтропии заключается в следующем. Обычную, шенноновскую энтропию текста H можно определить условием, что для l-буквенного алфавита число N-буквенных текстов, удовлетворяющих заданным статистическим ограничениям, равно не $l^N = 2^{N\log l} = 2^{NH_{\max}}$, как было бы в случае рассмотрения *любых* наборов их N последовательных букв, а всего лишь $M = 2^{NH}$. В соответствии с этим, введя понятие "осмысленный текст", можно определить комбинаторную энтропию текста H_{komb} как

$$H_{\text{komb}} = \lim_{N\to\infty} \frac{1}{N} \log M\left(N\right), \tag{2.9.3}$$

где $M(N)$ — число всевозможных осмысленных текстов длины N, которое можно оценить с помощью подсчёта числа возможных *продолжений* текста. А именно.

Пусть Ø — "пустое" слово, вовсе не содержащее букв, а d_k — некоторые буквы языка. Обозначим через $\Omega\left(d_1 d_2 \ \ldots \ d_k\right)$ число всевозможных *осмысленных продолжений* последовательности букв $d_1 d_2 \ \ldots \ d_k$, т. е. число x таких букв $d_{i_1} d_{i_2} \ldots d$, что полученная последовательность

$$d_1 d_2 \ldots d_k d_{i_1} d_{i_2} \ldots d_{i_x}$$

может быть продолжена до осмысленного текста. В таком случае значение величины

$$\tilde{M} = \Omega\left(\varnothing\right)\Omega\left(\varnothing d_1\right)\Omega\left(\varnothing d_1 d_2\right)\cdots\Omega\left(\varnothing d_1 d_2 \ldots d_{N-1}\right),$$

усреднённое по большому числу последовательностей букв, можно рассматривать как оценку величины $M(N)$.

В упомянутой работе А. Н. Колмогорова были намечены некоторые пути определения количества информации, содержащейся в сообщениях, без привлечения теоретико-вероятностных понятий. К сожалению, эти идеи, по-видимому, не получили дальнейшего развития.

Статистический анализ литературных источников порождает решение и обратной задачи: возможность моделирования письменной речи, исходя из присущих какому-либо языку закономерностей.

Пусть, например, будет создана последовательность букв для русского и английского языков. Предположим, что появление любой буквы равновероятно — это можно сделать, например, выбирая наугад буквы из раскрытой книги[1]. При этом будет получено, по терминологии Шеннона, *приближение нулевого порядка*:

XFOML RXKHRJFFJ ZLKWCBTDRQ
EMPDLDPNEZPOBSDQPMYSDAZPWLGE

или

УСУРХЗРОБЬЩДГКУЦЕ РЦТЩВ
СЫЮХЙКЬЧПЯРЮЖТЧУКБГХП,

где в данном случае перенос строки означает пробел между словами.

Для получения *приближения первого порядка*, когда буквы появляются независимо и с частотой, соответствующей данным из табл. 2.9.1 и 2.9.2, можно поступить следующим образом: положить в урну 1000 карточек, из которых 175 содержат пробел, 90 — букву “О”, и т.д.; затем вытаскивать по одной карточке, каждый раз записывая вытянутую букву и возвращая карточку снова в урну. Проведя опыт, можно получить “фразы”, подобные следующим:

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA QTH EEI
ALHENHTTRA OOBTTVA

или

ЫЕНТ ЦИЕА ОЕРВ ОДГН АУЕМЛОЛЙК ЗБЯ ЪЕНВТША.

Эти “фразы” несколько более похожи на осмысленный текст, чем предыдущие (наблюдается сравнительно правдоподобное распределение числа гласных и согласных и близкая к обычной средняя длина слова), но и они, разумеется, ещё очень далеки от разумного текста, прежде всего, ввиду отсутствия зависимости между буквами. Так, если очередной буквой явилась гласная, то значительно возрастает вероятность появления на следующем месте согласной буквы. В русском языке Ъ никак не может следовать за пробелом; с другой стороны, в английском языке нет ни одной пары букв, начинающейся с Q, кроме сочетания QU. Следовательно, необходим механизм, исключающий подобные образования.

[1] Описываемые опыты по моделированию текста, разумеется, с гораздо большим эффектом могут быть реализованы с использованием средств вычислительной техники.

Наиболее простой способ извлечения букв с вероятностными связями состоит в следующем. Берётся набор из 32 (или 27 — по числу букв в языке) урн, в каждой из которых находятся карточки, содержащие *диграммы*, т. е. двухбуквенные сочетания, начинающиеся с обозначенной на урне буквы в количестве, пропорциональном их относительным частотам. Опыт состоит в многократном извлечении карточек из урн и выписыванием с них последней буквы. При этом каждый раз (начиная со второго) карточка извлекается из той урны, которая содержит диграммы, начинающиеся с последней выписанной буквы, а после того, как буква выписана, карточка возвращается в урну.

Как и в предыдущем случае, этот же опыт допускает “книжную” (и, разумеется, компьютерную) модификацию, заметно упрощающий его проведение: вместо урн воспользоваться какой-либо книгой, в которой надо лишь, начиная каждый раз с выбранного наугад места, отыскивать первое появление последней уже выписанной буквы и следующую за ней букву из книги дописывать к уже имеющемуся тексту.

Подобными опытами получается следующее *приближение второго порядка*:

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

или

УМАРОНО КАЧ ВСВАННЫЙ РОСЯ МНЫХ КОВКРОВ НЕДАРЕ ОННЕДА.

По звучанию эти “фразы” заметно ближе к естественному языку, чем предыдущие (имеется не только правдоподобное соотношение числа гласных и согласных букв, но и близкое к привычному их чередование).

Подобным же образом, оперируя с *триграммами*, т. е. трёхбуквенными сочетаниями, и располагая урнами в количестве $32^2 = 1024$, в каждой из которых лежат карточки, начинающиеся на одни и те же диграммы (или, что эквивалентно, опыт с книгой, в которой много раз наудачу отыскивается первое повторение последней уже выписанной диграммы и выписывается следующая за ней буква), можно получить *приближение третьего порядка*:

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

или

ПОКАК ПОТ ДУРНОСКАКА НАКОНЕПНО ЗНЕ СТВОЛОВИЛ СЕ ТВОЙ ОБНИЛЬ.

В последних предложениях помимо “удобопроизносимости” всей фразы в целом, можно уже обнаружить “настоящие” слова.

Дальнейшие приближения к языку связаны с привлечением четырёх- и более буквенных сочетаний, что даже с использованием средств вычислительной техники приводит к практически нереализуемым опытам.

Рассмотрим теперь количество информации и энтропию для устной речи. Наиболее простым подходом к решению данной задачи было бы определение среднего числа букв, произносимых за единицу времени, например, за 1 секунду. Если исходить из данных физиологической акустики, то приближенная оценка количества информации, сообщаемой при разговоре за 1 с, составляет 5–6 бит. Однако эта оценка количества передачи информации при разговоре относится лишь к “смысловой информации”, которую можно извлечь из записи сказанных слов. На самом деле устная речь содержит, кроме того, ещё довольно значительную дополнительную информацию, которая часто может и противоречить “смысловой информации” (настроение говорящего, его отношении к сказанному, громкость устной речи и т. д.). Количественная оценка всей этой информации представляет собой очень сложную задачу, требующую значительно больших знаний о языке, чем имеется в настоящее время.

В этом многообразии сравнительно решённым является вопрос о логических ударениях, подчёркивающих в фразе отдельные слова; эти ударения также несут определённую информационную нагрузку, которую в случае звукового восприятия можно оценить количественно. Если вероятность того, что данное слово w_r находится под ударением обозначить через q_r, то среднее количество информации, заключающееся в сведениях о наличии или отсутствии ударения на этом слове, будет, очевидно, равно

$$-q_r \log q_r - (1-q_r)\log(1-q_r).$$

Пусть теперь $p_1, p_2, \ldots, p_K$ — вероятности (частоты) всех слов w_1, w_2, …, w_K, где K — общее число всех употребляемых слов. Тогда для сред-

него количества информации H, заключённого в логическом ударении, можно написать следующую формулу:

$$H = \sum_{i=1}^{K} p_i \left[-q_i \log q_i - (1-q_i) \log (1-q_i) \right]. \qquad (2.9..4)$$

Согласно исследованиям, средняя информация, которая получается при выяснении, на какие слова падает логическое ударение, близка к 0,65 бит/слово. Однако эта цифра имеет лишь условный характер, поскольку в действительности во время разговора произносятся звуки, значительно отличающиеся от букв. При этом основным элементом устной речи (в том же смысле, в каком буква является основным элементом письменной речи) считается отдельный звук — *фонема*. Поэтому гораздо больший интерес составляет не энтропия одной произнесённой буквы (являющейся условным понятием), а энтропия одной реально произнесённой фонемы.

Таблица 2.6

Энтропии различного приближения для устной речи

$H^{(0)}$	$H^{(1)}$	$H^{(2)}$	$H^{(3)}$
$\log_2 42 = 5,38$	4,77	3,62	0,70

Список фонем данного языка, разумеется, не совпадает со списком букв алфавита. Общее число фонем заметно превышает число букв, так как одна и та же буква в разных случаях может звучать по-разному. Следует заметить, что до сих пор не существует общепринятой точки зрения по вопросу количества фонем в том или ином разговорном языке, что, разумеется, приводит к неизбежному разногласию между авторами. Так, если выделить в русском языке 42 различных фонемы и подсчитать частоты отдельных фонем, а также различных комбинаций двух и трёх следующих друг за другом фонем, то получится ряд значений энтропии, представленных в табл. 2.6.

Как следует из сравнения данных, представленных в табл. 2.5 и 2.6, убывание ряда условных энтропий для фонем происходит заметно быстрее, чем в случае букв письменного текста.

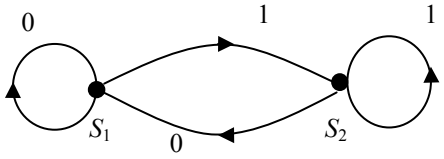
ВОПРОСЫ И ЗАДАНИЯ К ГЛАВЕ 2

2.1. Простой марковский источник, порождающий сообщения, состоящие из четырёх букв алфавита X: x_1, x_2, x_3 и x_4, описывается следующими распределениями вероятностей:

	x_1	x_2	x_3	x_4
x_1	0,00	0,20	0,40	0,40
x_2	0,20	0,20	0,30	0,30
x_3	0,25	0,00	0,25	0,50
x_4	0,20	0,40	0,40	0,00

Вычислить энтропии $H(X)$, $H(X/X)$, $H(X/X^2)$, $H(X/X^\infty)$ и найти соответствующие избыточности

2.2. Марковский источник задан следующими диаграммой и переходной матрицей, описывающими появление одиночных символов:

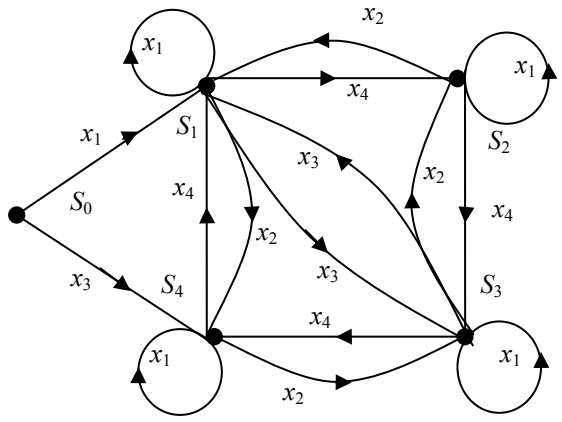


$$P(0/S_1)=0,8 \quad P(1/S_1)=0,2$$
$$P(0/S_2)=0,6 \quad P(1/S_2)=0,4$$

а) Найти стационарное значение энтропии источника на символ.

б) Построить диаграмму и найти переходные вероятности для пар символов.

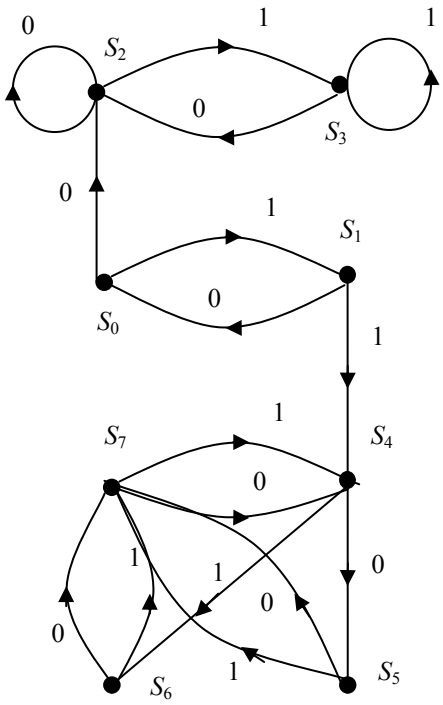
2.3. Диаграмма состояний марковского источника и вероятности появления символов в каждом из состояний имеют следующий вид:



	S_0	S_1	S_2	S_3	S_4
x_1	0,300	0,700	0,300	0,500	0,300
x_2	0,000	0,125	0,500	0,100	0,500
x_3	0,700	0,075	0,000	0,100	0,000
x_4	0,000	0,100	0,200	0,300	0,200

Полагая, что начальным состоянием является состояние S_0, найти финальные вероятности состояний и вычислить энтропию источника.

2.4. Марковский источник, представленный своей диаграммой, находясь в начальном состоянии S_0, начинает генерировать последовательности двоичных символов в соответствии с указанной таблицей.



	$P(0/S_r)$	$P(1/S_r)$
S_0	1/2	1/2
S_1	1/2	1/2
S_2	1/2	1/2
S_3	1/4	3/4
S_4	1/4	3/4
S_5	1/4	3/4
S_6	1/2	1/2
S_7	1/2	1/2

а) Указать, какие состояния источника являются возвратными и невозвратными.

б) Указать все неразложимые подцепи и определить период всех периодических подцепей.

в) Найти предельное значение вероятности каждой неразложимой подцепи.

г) Найти асимптотическое значение энтропии на символ для каждой неразложимой подцепи.

2.5. *Рангом* символа называется положение символа в алфавите источника в порядке убывания вероятностей. Пусть $P(r)$ — вероятность появления на выходе источника символа r-го ранга и пусть

$$q_r = P(r) - P(r+1).$$

а) Показать, что для условной энтропии $H(X/X^{\infty})$ непериодического l-элементного марковского источника X справедлива следующая оценка:

$$\sum_{r=1}^{l} r q_r \log r \le H\left(X / X^{\infty}\right) \le -\sum_{r=1}^{l} P(r) \log P(r).$$

б) Рассмотрим l-элементный стационарный непериодический источник, для которого каждый выходной символ зависит от ν предыдущих символов и представим такой источник при помощи диаграммы с l^{ν} состояниями. Построим *редуцированную* диаграмму, получаемую объединением в одно состояние те l состояний, для которых предшествующие $(\nu - 1)$ выходных символов совпадают. В такой редуцированной диаграмме принимаются во внимание зависимость только от $(\nu - 1)$ предшествующих символов. Показать справедливость следующих неравенств

$$\sum_{r=1}^{t} P(r) \geq \sum_{r=1}^{t} P'(r), \qquad 1 \leq t \leq l;$$

$$-\sum_{r=1}^{l} P(r)\log P(r) \leq \sum_{r=1}^{l} P'(r)\log P'(r);$$

$$\sum_{r=1}^{l} rq_r \log r \leq \sum_{r=1}^{l} rq'_r \log r,$$

где $P'(r)$ — вероятность появления на выходе источника символа r-го ранга для редуцированной диаграммы, а $q'_r = P'(r) - P'(r+1)$.

2.6. Найти корреляционную функцию для следующих процессов:

а) *пуассоновского сигнала* $\pi(t)$ — целочисленного случайного процесса с независимыми приращениями, для которого вероятность того, что будет принято значение, равное k ($k = 0, 1, \ldots$) определяется выражением

$$P\left[\pi(t) = k\right] = \frac{(\lambda t)^k}{k!}\exp\left(-\lambda t\right),\ \lambda > 0.$$

б) *телеграфного сигнала*

$$\alpha(t) = \left(-1\right)^{\pi(t)},$$

где $\pi(t)$ — пуассоновский сигнал, рассмотренный выше, а α — не зависящая от него случайная величина, равновероятно принимающая значения ± 1.

в) *случайное колебание*

$$\xi(t) = \alpha\cos\left(\omega_0 t + \varphi_0\right),$$

где α и φ_0 — случайные независимые параметры, причём φ_0 распределено равномерно на интервале $[0; 2\pi]$.

Являются ли такие процессы стационарными? Если да, то найти энергетический спектр.

2.7. Покажите, что стационарный гауссовский процесс с нулевым средним обладает свойством эргодичности и по отношению к среднему, и по отношению к корреляционной функции.

2.8. Покажите, что эпсилон-энтропия $H_n(\varepsilon)$ случайных процессов является выпуклой вниз функцией, равна нулю в некоторой области значений, ограничена для всех $n = 1, 2, \ldots$, а последовательность $\{H_n(\varepsilon)\}$ имеет предельное значение $\lim\limits_{n\to\infty} H_n(\varepsilon) = H(\varepsilon)$.

2.9. Источник порождает независимые символы из троичного алфавита $X = \{A, B, C\}$ с вероятностями $P(A) = 0{,}4$; $P(B) = 0{,}4$; $P(C) = 0{,}2$. Для троичного аппроксимирующего источника мера искажения имеет вид (2.8.5). Найти эпсилон-энтропию источника X и изобразить её графически.

2.10. Гауссовский стационарный сигнал характеризуется энергетическим спектром вида

$$G(\omega) = \frac{1}{1 + \left(\omega/\omega_0\right)^{2n}},\ \omega_0 > 0,\ n = 1, 2, \ldots .$$

а) Найти корреляционную функцию сигнала.

б) Найти ε-энтропию сигнала при квадратичном критерии качества и сравнить со значением ε-энтропии гауссовского сигнала, имеющего такую же дисперсию и прямоугольную форму энергетического спектра.

ГЛАВА 3. ДИСКРЕТНЫЕ КАНАЛЫ И ИХ ХАРАКТЕРИСТИКИ

В предыдущих главах достаточно подробно изучались основные понятия теории информации, в частности, характеристики количества информации, содержащейся в сообщениях, главной из которой является энтропия — в различных её вариациях. Однако само по себе знание энтропийных характеристик сообщений хотя и является полезным в теоретико-информационном плане, но не должно являться самоцелью, поскольку основной задачей, стоящей перед разработчиками телекоммуникационных систем, является передача информации по реальным каналам связи. Особенностью большинства таких каналов является наличие в них помех, имеющих спорадическую природу и, следовательно, не могущих быть полностью учтёнными на передающей стороне. Таким образом, передача сообщений в общем случае связана с некоторой потерей информации. Рассмотрим основные характеристики дискретных каналов и попытаемся определить те возможности и ограничения, которые возникают при использовании различных моделей.

3.1. ПРОПУСКНАЯ СПОСОБНОСТЬ КАНАЛА

Дискретный канал определяется следующими параметрами:

- алфавитом кодовых символов y_i ($i = 1, \ldots, m$), поступающих на его вход;
- алфавитом кодовых символов y'_j ($j = 1, \ldots, m'$), снимаемых с его выхода;
- количеством v кодовых символов, пропускаемых каналом в единицу времени (т. е. технической скоростью передачи);
- значениями вероятностей переходов $p(y'_j / y_i)$, т. е. вероятностей того, что на выходе появится символ y'_j, если на вход подан символ y_i.

Часто в определении канала фигурирует не техническая скорость передачи, а обратная ей величина $T = 1/v$ — время, затрачиваемое на подачу в канал одного символа. При этом приходится различать *источники с фиксированной скоростью* и *источники с управляемой скоростью*. Различие

между этими двумя случаями заключается в трактовке понятия *производительности* H', т. е. среднего количества информации, выдаваемого в единицу времени:

$$H'(Y) = \frac{H(Y)}{T}.$$

В первом случае рассматривается производительность источника с фиксированной скоростью, и она не может быть изменена при проектировании системы передачи информации. Во втором случае говорят о производительности передающего устройства, которая выбирается проектировщиком в соответствии с различными техническими и экономическими требованиями, предъявляемыми к системе.

Следует особо подчеркнуть, что скорость передачи v и переходные вероятности являются связанными между собой величинами: уменьшение или увеличение скорости передачи символов, вообще говоря, изменяет механизмы прохождения сигналов по каналу (описываемые, например, свёрткой входного сигнала с эквивалентной импульсной характеристикой канала) и, как следствие, приводит к изменению вероятностей правильной или ошибочной передачи.

Вероятности $p(y'_j / y_i)$, вообще говоря, зависят от того, какие символы передавались и принимались ранее. Если вероятности перехода $p(y'_j / y_i)$ для каждой пары (j, i) остаются постоянными и не зависят от того, какие символы передавались и принимались ранее, то такой канал называется *постоянным* или *однородным*. (Иногда применяют также названия *канал без памяти* или *канал с независимыми ошибками*.) Если к тому же в канале алфавиты на входе и выходе одинаковы, и для любой пары $i \neq j$ вероятности $p(y'_j / y_i) = p = \text{const}$, то такой канал называют *симметричным*.

Обозначим через $y^{(k)}$ произвольный k-й (в порядке следования) символ, а через $y'^{(k)}$ — соответствующий ему выходной символ, так что n-элементному сообщению $\mathbf{y} = (y^{(1)}, \ldots, y^{(n)})$ на входе соответствует выходное n-элементное сообщение $\mathbf{y}' = (y'^{(1)}, \ldots, y'^{(n)})$. Тогда, если на n-мерном множестве Y^n задано совместное распределение $p(\mathbf{y}) \equiv P(Y^n)$ вероятностей входных символов, то совместное распределение $p(\mathbf{y}') \equiv P(Y'^n)$ соответствующих выходных символов равно

$$p(\mathbf{y}') = \sum_{Y^n} P(\mathbf{y}) P(\mathbf{y}' / \mathbf{y}), \qquad (3.1.1)$$

где суммирование ведётся по всему входному ансамблю сообщений.

Выражение для среднего количества взаимной информации между входными и выходными последовательностями сообщений имеет вид

$$I\left(Y'^n;Y^n\right)=H\left(Y'^n\right)-H\left(Y'^n/Y^n\right), \qquad (3.1.2)$$

где, как обычно, ансамбли Y^n и Y'^n являются произведением n алфавитов, т. е. если, например, алфавит Y_k соответствует сообщению $y^{(k)}$ ($k = 1, \ldots, n$), то

$$Y^n = Y_1 Y_2 \ldots Y_k \ldots Y_n,$$

и аналогично для ансамбля Y'^n. Конечно, в каждый момент времени $Y_i = Y$ и $Y_i' = Y'$.

Значение среднего количества взаимной информации (3.1.2) зависит как от распределения вероятностей последовательностей входных символов, поскольку вероятности $p(\mathbf{y}')$ согласно (3.1.1) зависят от $p(\mathbf{y})$, так и от условных распределений $p(y'_j / y_i)$, характеризующих канал. Фактически величина $I(Y'^n; Y^n)$ является мерой среднего количества взаимной информации, содержащейся в n символах на выходе канала относительно n символов на входе канала. Если символы поступают в канал со средней скоростью v, то целесообразно определить характеристику

$$I'(Y'^n;Y^n)=\frac{v}{n}I(Y'^n;Y^n), \qquad (3.1.3)$$

являющуюся скоростью передачи информации по каналу при использовании n-элементных последовательностей.

Выбор того или иного входного многомерного распределения $p(\mathbf{y}')$ приводит к возможности реализации (хотя бы в принципе) б*о*льших или меньших скоростей передачи информации по заданному каналу. Желая получить не зависящую от статистических свойств сообщений **y** характеристику канала, введём фундаментальное для всей теории информации понятие *пропускной способности канала*[1] как максимальное значение скорости передачи информации, найденное по всем возможным входным распределениям $p(\mathbf{y})$.

[1] Понятие пропускной способности канала было введено К.Э. Шенноном англоязычным термином “capacity”, перевод которого имеет достаточно широкую вариативность. Так, некоторые авторы (например, [7]) для названия характеристики (3.1.4) используют термин “ёмкость канала”, что представляется не совсем удачным.

Так, рассматривая всевозможные входные n-элементные последовательности $\mathbf{y} = (y^{(1)}, \ldots, y^{(n)})$, определим пропускную способность C_n дискретного канала при его однократном использовании как

$$C_n = \sup_{P(Y^n)} \frac{v}{n} I(Y'^n; Y^n). \qquad (3.1.4)$$

В этом соотношении фигурирует супремум множества. Математически это означает, что искомое распределение $p(\mathbf{y}')$, максимизирующее значение скорости передачи информации, может и не принадлежать исходному множеству, на котором рассматриваются всевозможные многомерные распределения сообщений[1]. Однако с практической точки зрения такая постановка задачи вряд ли оправдана, и более разумным представляется поиск не супремума, а максимума:

$$C_n = \max_{P(Y^n)} \frac{v}{n} I\left(Y'^n; Y^n\right). \qquad (3.1.4a)$$

Эта характеристика хотя и максимизирована по всем входным распределениям, но, вообще говоря, зависит от длины передаваемых последовательностей. Поэтому разумно определить инвариантную по отношению к длине n последовательности сообщений характеристику пропускной способности канала посредством предельного перехода:

$$C = \lim_{n\to\infty} C_n = \lim_{n\to\infty}\left[\max_{P(Y^n)} \frac{v}{n} I\left(Y'^n; Y^n\right)\right]. \qquad (3.1.5)$$

Задачи нахождения пропускной способности каналов представляют собой целое теоретико-информационное направление, они характеризуются большой сложностью как в аналитическом, так и в вычислительном плане. Достаточно сказать, что для большинства более или менее адекватных моделей каналов не удаётся быть уверенным даже в существовании предела, фигурирующего в соотношении (3.1.5). Тем не менее, существует класс каналов, для которых пропускную способность удаётся найти сравнительно несложными методами. Это класс *постоянных дискретных каналов*.

---

[1] Разницу между максимумом и экстремумом множества можно пояснить на примере задачи о нахождении максимума элементарной функции $f(x) = x^2$ на открытом интервале (0; 2). Очевидно, что на этом множестве рассматриваемая функция не имеет максимума, ибо она принимает значения, сколь угодно близкие к 4, являющимся супремумом, но никогда не достигает его.

3.2. ПРОПУСКНАЯ СПОСОБНОСТЬ ПОСТОЯННОГО ДИСКРЕТНОГО КАНАЛА

Из большого разнообразия возможных моделей каналов самый элементарный класс образуют *односторонние дискретные постоянные каналы*, т. е. каналы с одним входом, одним выходом и с дискретными входными и выходными алфавитами, связанными заданными условными распределениями вероятностей. Такие каналы полностью описываются заданием входного $Y = \{y_i, i = 1, \ldots, m\}$ и выходного $Y' = \{y'_j, j = 1, \ldots, m'\}$ алфавитов, технической скоростью следования символов v и совокупностью условных вероятностей $p(y'_j / y_i)$, часто называемых также *переходными* вероятностями.

Как следует из определения постоянного канала, условное распределение $p\left(y_j'^{(k)} / y_i^{(k)}\right)$ остаётся одним и тем же для любого значения k независимо от передачи всех предшествующих сообщений, следовательно[1],

$$p\left(\mathbf{y}' / \mathbf{y}\right) = \prod_{k=1}^{n} p\left(y'^{(k)} / y^{(k)}\right). \qquad (3.2.1)$$

Например, при $m = m' = 2$ и $n = 2$ входной ансамбль имеет вид

$$Y^2 = \left\{y_1y_1,\ y_1y_2,\ y_2y_1,\ y_2y_2\right\},$$

тогда на основании (1.1.1) и (1.2.1) имеем для Y'^2:

$$p\left(\mathbf{y}'\right) = p\left(y_1'y_1' / y_1y_1\right)p\left(y_1y_1\right) + p\left(y_1'y_2' / y_1y_2\right)p\left(y_1y_2\right) +$$

$$+ p\left(y_2'y_1' / y_2y_1\right)p\left(y_2y_1\right) + p\left(y_2'y_2' / y_2y_2\right)p\left(y_2y_2\right) =$$

$$= p^2\left(y_1' / y_1\right)p\left(y_1y_1\right) + p\left(y_1' / y_1\right)p\left(y_2' / y_2\right)\left[p\left(y_1y_2\right) + p\left(y_2y_1\right)\right] +$$

$$+ p^2\left(y_2' / y_2\right)p\left(y_2y_2\right).$$

Если к тому же входные символы независимы и равновероятны, то

$$p\left(\mathbf{y}'\right) = p^2\left(y_1' / y_1\right)\cdot\frac{1}{4} + p\left(y_1' / y_1\right)p\left(y_2' / y_2\right)\cdot\left(\frac{1}{4} + \frac{1}{4}\right) + p^2\left(y_2' / y_2\right)\cdot\frac{1}{4} =$$

$$p\left(\mathbf{y}'\right) = \frac{1}{4}\left[p^2\left(y_1' / y_1\right) + 2p\left(y_1' / y_1\right)p\left(y_2' / y_2\right) + p^2\left(y_2' / y_2\right)\right].$$

Поставим задачу о нахождении пропускной способности постоянного дискретного канала.

---

[1] Для упрощения записи нижние индексы, отражающие позицию символа в алфавите, опущены.

Фигурирующие в (3.1.2) условную и безусловную энтропии можно выразить в виде суммы условных энтропий при заданных предшествующих событиях. Действительно, поскольку для произвольных дискретных источников A и B справедливо соотношение

$$H\left(AB\right)=H\left(A\right)+H\left(B/A\right),$$

последовательно применяя его, имеем:

$$H\left(Y'^{n}\right)\equiv H\left(Y'_1,Y'_2,...,Y'_n\right)=H\left(Y'_1\right)+H\left(Y'_2,...,Y'_n/Y'_1\right)=$$

$$=H\left(Y'_1\right)+H\left(Y'_2/Y'_1\right)+H\left(Y'_3,...,Y'_n/Y'_2Y'_1\right)=...=$$

$$=H\left(Y'_1\right)+H\left(Y'_2/Y'_1\right)+H\left(Y'_3/Y'_2Y'_1\right)+...+H\left(Y'_n/Y'_{n-1}Y'_{n-2}\ldots Y'_1\right), \quad (3.2.2)$$

где

$$H\left(Y'_k/Y'_{k-1}\ldots Y'_1\right)=-\sum_{Y'^{k}}P\left(Y'^{k}\right)\log p\left(y'^{(k)}/y'^{(k-1)}\ldots y'^{(1)}\right),$$

а $P(Y'^{n})$ — совместная вероятность k-элементных последовательностей символов $(y'^{(1)}, y'^{(2)}, \ldots, y'^{(k)})$ на выходе канала.

С другой стороны,

$$H\left(Y'^{n}/Y^{n}\right)=-\sum_{Y^{n}}\sum_{Y'^{n}}p\left(\mathbf{y}',\mathbf{y}\right)\log p\left(\mathbf{y}'/\mathbf{y}\right)=$$

$$=-\sum_{Y^{n}}\sum_{Y'^{n}}p\left(y'^{(1)},y^{(1)},y'^{(2)},y^{(2)},\ldots,y'^{(n)},y^{(n)}\right)\log\prod_{k=1}^{n}p\left(y'^{(k)}/y^{(k)}\right)=$$

$$=-\sum_{Y^{n}}\sum_{Y'^{n}}p\left(y'^{(1)},y^{(1)},y'^{(2)},y^{(2)},\ldots,y'^{(n)},y^{(n)}\right)\sum_{k=1}^{n}\log p\left(y'^{(k)}/y^{(k)}\right).$$

Структура последнего выражения такова, что совместные вероятности $p(\mathbf{y}', \mathbf{y})$ суммируются по n-мерным входному и выходному алфавитам Y^n и Y'^n со своеобразными “окнами”-множителями $\log p(y'^{(k)} / y^{(k)})$, которые в итоге оставляют суммирование по соответствующим одномерным алфавитам Y_k и Y'_k. Например, для уже рассмотренного случая $m = m' = 2$ и $n = 2$ имеем:

$$H\left(Y'^{2}/Y^{2}\right)=-\sum_{Y'^{2}}\sum_{Y^{2}}p\left(y'^{(1)},y^{(1)},y'^{(2)},y^{(2)}\right)\log\prod_{k=1}^{2}p\left(y'^{(k)}/y^{(k)}\right)=$$

$$=-\sum_{Y'^{2}}\sum_{Y^{2}}p\left(y'^{(1)},y^{(1)},y'^{(2)},y^{(2)}\right)\left[\log p\left(y'^{(1)}/y^{(1)}\right)+\log p\left(y'^{(2)}/y^{(2)}\right)\right]=$$

$$=-\sum_{Y_1}\sum_{Y'_1}p\left(y'^{(1)},y^{(1)}\right)\log p\left(y'^{(1)}/y^{(1)}\right)-\sum_{Y_2}\sum_{Y'_2}p\left(y'^{(2)},y^{(2)}\right)\log p\left(y'^{(2)}/y^{(2)}\right)=$$

$$= H\left(Y_1'/Y_1\right) + H\left(Y_2'/Y_2\right).$$

Поэтому в общем случае как нетрудно видеть,

$$H\left(Y'^n / Y^n\right) = -\sum_{k=1}^{n}\left(\sum_{Y_k'}\sum_{Y_k} p\left(y^{(k)}, y'^{(k)}\right)\log p\left(y'^{(k)}/y^{(k)}\right)\right) = \sum_{k=1}^{n} H\left(Y_k'/Y_k\right). \quad (3.2.3)$$

Тогда, подставляя (3.2.2) и (3.2.3) в (3.1.2), имеем:

$$I\left(Y'^n; Y^n\right) = \sum_{k=1}^{n}\left[H\left(Y_k'/Y_{k-1}'\ldots Y_1'\right) - H\left(Y_k'/Y_k\right)\right], \quad (3.2.4)$$

где при $k=1$ условная энтропия $H\left(Y_k'/Y_{k-1}'Y_{k-2}'\ldots Y_1'\right)$ полагается равной $H\left(Y_1'\right)$.

При нахождении максимума (3.2.4) учтём, что наличие вероятностных связей может лишь уменьшить величину энтропии, т. е. справедлива усугубляющая оценка

$$H\left(Y_k'/Y_{k-1}'\ldots Y_1'\right) \le H\left(Y_k'\right),$$

а поскольку

$$H\left(Y_k'\right) - H\left(Y_k'/Y_k\right) = I\left(Y_k'; Y_k\right) = I\left(Y'; Y\right),\ k = 1,\ldots n,$$

имеем

$$C = \lim_{n\to\infty}\max_{P(Y^n)}\left[\frac{v}{n} nI\left(Y'; Y\right)\right] = v\max_{P(Y)} I\left(Y'; Y\right). \quad (3.2.5)$$

Итак, *максимальное значение скорости передачи информации для постоянного дискретного канала достигается в том случае, когда сообщения на его выходе статистически независимы. При этом для определения пропускной способности канала при передаче последовательностей произвольной длины достаточно произвести максимизацию лишь по одномерному входному распределению* $P(Y)$.

Отметим, что вычисление пропускной способности согласно (3.2.5) включает в себя максимизацию функции $I(Y, Y')$ по m переменным с двумя ограничениями, одно из которых — ограничение-равенство в виде условия нормировки

$$\sum_{i=1}^{m} p(y_i) = 1,$$

а другое — ограничение-неравенство

$$p(y_i) \ge 0,\ i = 1,\ldots, m.$$

Более того, зачастую практический смысл имеет максимизация по ненулевым вероятностям, так что может оказаться, что ограничение-неравенство является более жёстким:

$$p(y_i) > 0,\ i = 1, \ldots, m,$$

и если в алфавите, на котором реализуется пропускная способность, содержатся нулевые вероятности, то это означает вырожденность решения.

Для дальнейших вычислений необходимо конкретизировать параметры канала. Прежде, чем перейти к методам нахождения пропускной способности произвольных постоянных каналов рассмотрим частный случай постоянного канала — симметричный постоянный канал.

3.3. ПРОПУСКНАЯ СПОСОБНОСТЬ ПОСТОЯННОГО СИММЕТРИЧНОГО ДИСКРЕТНОГО КАНАЛА

Как уже было сказано выше, для любого постоянного канала, вероятности переходов $p(y'_j / y_i)$ не зависят от последовательностей ранее передаваемых символов, т. е. в произвольный момент времени матрица переходных вероятностей имеет вид

$$\mathbf{P} = \begin{pmatrix} p(y'_1/y_1) & p(y'_2/y_1) & \ldots & p(y'_{m'}/y_1) \\ p(y'_1/y_2) & p(y'_2/y_2) & \ldots & p(y'_{m'}/y_2) \\ \ldots & \ldots & \ldots & \ldots \\ p(y'_1/y_m) & p(y'_2/y_m) & \ldots & p(y'_{m'}/y_m) \end{pmatrix}, \qquad (3.3.1)$$

где сумма элементов любой строки равна единице.

Следует заметить, что указанный вид канальной матрицы соответствует случаю описания канала со стороны источника (наблюдатель, находясь на передающей стороне, знает, что был послан, например, символ y_1, и он с единичной вероятностью перейдёт в один из выходных символов $y'_1, \ldots, y'_{m'}$.

Если же канал описывается со стороны приёмника, то с получением сообщения y'_j предполагается, что было послано какое-то из сообщений $y_1, \ldots, y_m$. При этом канальная матрица имеет вид

$$\mathbf{P} = \begin{pmatrix} p(y'_1/y_1) & p(y'_1/y_2) & \ldots & p(y'_1/y_m) \\ p(y'_2/y_1) & p(y'_2/y_2) & \ldots & p(y'_2/y_m) \\ \ldots & \ldots & \ldots & \ldots \\ p(y'_{m'}/y_1) & p(y'_{m'}/y_2) & \ldots & p(y'_{m'}/y_m) \end{pmatrix}, \qquad (1.3.1a)$$

и единице должны равняться суммы условных вероятностей не по строкам, а по столбцам.

Будем различать следующие разновидности симметричных каналов.

Дискретный канал называется *симметричным по выходу*, если все строки матрицы переходных вероятностей образованы перестановками элементов первой строки. Для такого канала условная энтропия $H(Y'/Y)$ не зависит от распределения вероятностей на входе, поскольку

$$H(Y'/y_i) = -\sum_{j=1}^{m'} p(y'_j/y_i)\log p(y'_j/y_i) = -\sum_{j=1}^{m'} p_j \log p_j \ , \ i = 1,\ldots,m,$$

и правая часть не зависит от индекса i, а набор вероятностей p_j не меняет своих значений и лишь перегруппировывается от строки к строке.

Дискретный канал называется *симметричным по входу*, если все столбцы матрицы переходных вероятностей образованы перестановками элементов первого столбца. Легко видеть, что если канал симметричен по входу, и символы на его входе равновероятны, то также равновероятны и символы на его выходе. Действительно, пусть $p(y_i) = 1/m$ $(i = 1,\ldots,m)$, тогда

$$p(y'_j) = \sum_{i=1}^{m} p(y'_j/y_i)\,p(y_j) = \frac{1}{m}\sum_{i=1}^{m} q_i \ ,$$

и правая часть не зависит от индекса j, а набор вероятностей q_i аналогично p_j не меняет своих значений и лишь перегруппировывается от столбца к столбцу.

Дискретный канал называется *симметричным*, если он симметричен по входу и выходу[1]. Иными словами, и строки и столбцы матрицы переходных вероятностей образованы перестановками одного набора чисел.

Исходя из указанных свойств, найдём пропускную способность симметричного (по входу и выходу) дискретного канала. Имеем:

$$\frac{C}{v} = \max_{p(Y)} I(Y',Y) = \max_{p(Y)} H(Y') - H(Y'/Y) =$$

$$= \log m' + \sum_{j=1}^{m'} p_j \log p_j \ , \qquad (3.3.2)$$

[1] Иногда в литературе встречаются термины *полусимметричный канал* (по входу или выходу) и *строго симметричный канал*.

и, как это следует из рассмотренного выше, пропускная способность достигается при равновероятных входных символах.

В частности, рассмотрим симметричный канал, для которого алфавиты Y и Y' полностью совпадают ($m = m'$), а переходные вероятности определяются выражением

$$p\left(y'_j / y_i\right) = \begin{cases} 1-p, & j = i; \\ p/(m-1), & j \neq i, \end{cases} \tag{3.3.3}$$

где p — так называемая *средняя вероятность ошибки*.

Тогда

$$\frac{C}{v} = \log m + (1-p)\log(1-p) + p\log\frac{p}{m-1}. \tag{3.3.4}$$

На рис. 3.1 представлены нормированные к величине v зависимости пропускной способности канала C от вероятности ошибки p, вычисленные по формуле (3.3.4) для различных значений m объёма входного алфавита Y. Беря производную от функции C по переменной p, нетрудно убедиться в том, что в точках $p = (m-1)/m$ пропускная способность принимает минимальное, нулевое значение.

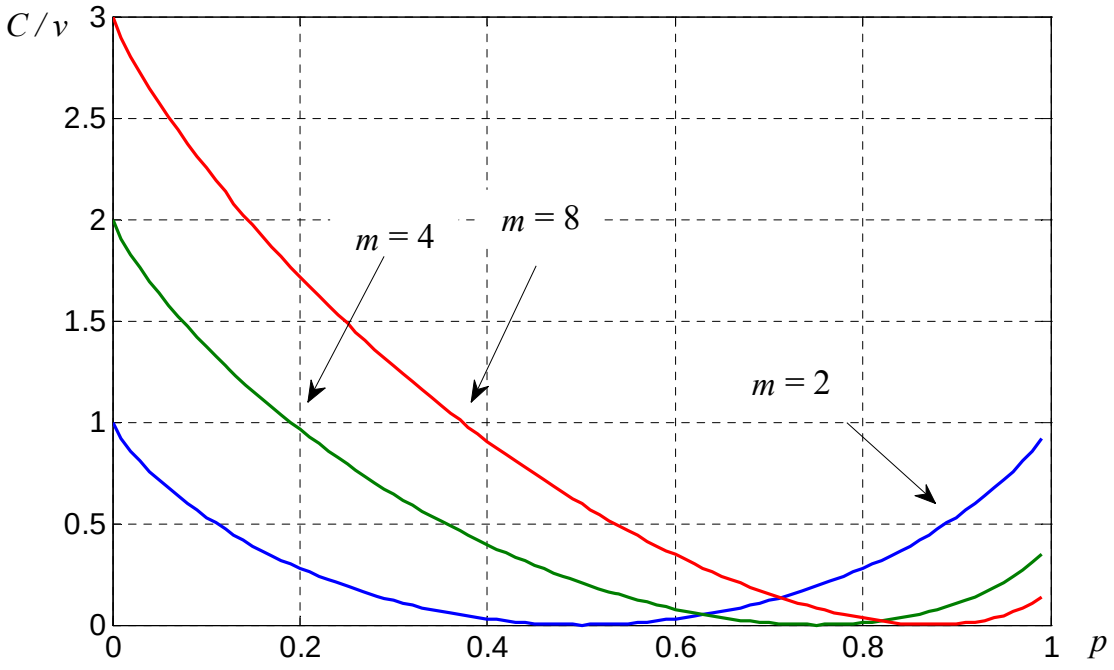


Рис. 3.1. Пропускная способность постоянного симметричного канала

В качестве другого примера найдём пропускную способность *двоичного стирающего канала*, выходной алфавит которого $Y' = \{0, 1, Ø\}$ отличается от входного алфавита $Y = \{0, 1\}$ наличием символа Ø — “стирание”, появление которого означает отказ от выноса решения в пользу любого из элементов 0 или 1 (рис. 3.2). Матрица переходных вероятностей для такого канала имеет следующий вид:

$$\mathbf{P} = \begin{pmatrix} 1-p-q & p & q \\ p & 1-p-q & q \end{pmatrix}, \tag{1.3.5}$$

где обозначено

$$p(0/0) = p(1/1) = 1-p-q,$$
$$p(0/1) = p(1/0) = p, \quad p(\varnothing/0) = p(\varnothing/1) = q.$$

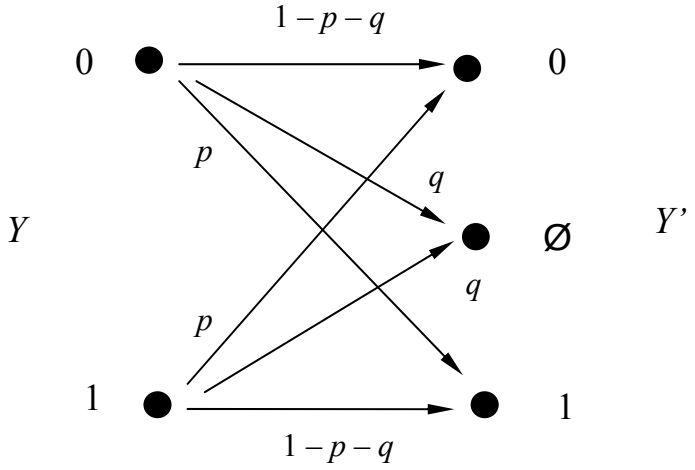


Рис. 3.2. Двоичный симметричный канал со стиранием

Очевидно, что двоичный стирающий канал симметричен по входу, но не симметричен по выходу, поэтому нельзя непосредственно воспользоваться выражением (3.1.2), и максимизацию придётся осуществлять “прямым” методом, находя частные производные и приравнивая их к нулю.

Безусловные вероятности символов на выходе канала равны

$$p(y_1') = p(y_1'/y_1)p(y_1) + p(y_1'/y_2)p(y_2);$$
$$p(y_2') = p(y_2'/y_1)p(y_1) + p(y_2'/y_2)p(y_2);$$
$$p(y_3') = p(y_3'/y_1)p(y_1) + p(y_3'/y_2)p(y_2).$$

Обозначив для простоты $p_1 = p(y_1)$ и $p_2 = p(y_2)$, можно записать выражение для энтропии $H(Y')$ источника на выходе канала:

$$H\left[Y'\left(p_1,p_2\right)\right]=-\left[p_1\left(1-p-q\right)+p_2p\right]\log\left[p_1\left(1-p-q\right)+p_2p\right]-$$
$$-\left[p_1p+p_2\left(1-p-q\right)\right]\log\left[p_1p+p_2\left(1-p-q\right)\right]-$$
$$-\left(p_1q+p_2q\right)\log\left(p_1q+p_2q\right), \qquad (3.3.6)$$

которая в данной записи зависит только от вероятностей входных символов.

Для нахождения значений вероятностей входных символов, дающих максимум энтропии на выходе канала, продифференцируем (3.3.6) частным образом по p_i (i = 1, 2) и приравняем нулю полученные производные:

$$-\frac{\partial H\left[Y'\left(p_1,p_2\right)\right]}{\partial p_1}=\left(1-p-q\right)\log\left[p_1\left(1-p-q\right)+p_2p\right]+\frac{\left(1-p-q\right)}{\ln 2}+$$

$$+\frac{p}{\ln 2}+p\log\left[p_1p+p_2\left(1-p-q\right)\right]+q\log\left(p_1q+p_2q\right)+\frac{q}{\ln 2}=0, \qquad (3.3.7)$$

$$-\frac{\partial H\left[Y'\left(p_1,p_2\right)\right]}{\partial p_2}=p\log\left[p_1\left(1-p-q\right)+p_2p\right]+\frac{p}{\ln 2}+\frac{\left(1-p-q\right)}{\ln 2}+$$

$$+\left(1-p-q\right)\log\left[p_1p+p_2\left(1-p-q\right)\right]+q\log\left(p_1q+p_2q\right)+\frac{q}{\ln 2}=0. \qquad (3.3.8)$$

Тогда, вычитая (3.3.8) из (3.3.7), имеем:

$$\left(1-2p-q\right)\log\frac{p_1\left(1-p-q\right)+p_2p}{p_1p+p_2\left(1-p-q\right)}=0.$$

Поскольку в общем случае соотношение между вероятностями p и q произвольно, множитель $(1-2p-q)$ следует считать не равным нулю, а обнулить логарифм, т. е.

$$p_1\left(1-p-q\right)+p_2p=p_1p+p_2\left(1-p-q\right),$$

откуда следует, что $H\left(Y'\right)$ максимизируется при равновероятных входных символах. При этом, как нетрудно видеть,

$$p\left(y=0\right)=p\left(y=1\right)=\frac{1-q}{2},\ p\left(y=\varnothing\right)=q, \qquad (3.3.9)$$

а энтропия выходного алфавита равна

$$H\left(Y'\right)=-q\log q-\left(1-q\right)\log\frac{1-q}{2}.$$

Поскольку стирающий канал симметричен по входу, для него

$$H\left(Y'/Y\right)=-p\log p-q\log q-\left(1-p-q\right)\log\left(1-p-q\right),$$

так что пропускная способность двоичного симметричного канала имеет

следующий вид:

$$\frac{C}{v} = -(1-q)\log\frac{1-q}{2} + p\log p + (1-p-q)\log(1-p-q). \qquad (3.3.10)$$

В частном случае при $p = 0$ (помехи в канале отсутствуют, не считая стираний) пропускная способность

$$\frac{C}{v} = 1 - q, \qquad (3.3.11)$$

т. е. равна пропускной способности двоичного симметричного канала при отсутствии в нем шума (1 бит/с) минус вероятность стирания q.

Полностью несимметричные каналы на практике встречаются редко, поэтому для них мало разработаны методы нахождения пропускной способности. Общим для таких каналов является то, что максимальное значение скорости передачи информации реализуется на неравномерном входном распределении, когда чаще используются те символы, которые принимаются с большей верностью.

Рассмотрим более подробно простейший двоичный несимметричный постоянный канал, для которого входным и выходным алфавитами являются

$$Y = \{y_1 = 0; y_2 = 1\},\ Y' = \{y_1' = 0; y_2' = 1\},$$

входные вероятности равны

$$p(y_1) = p(0) = p,\ p(y_2) = p(1) = q = 1 - p,$$

а канальные вероятности имеют вид

$$p(y_1'/y_2) = p(y' = 0/y = 1) = p_1,$$
$$p(y_2'/y_1) = p(y' = 1/y = 0) = p_2 \neq p_1. \qquad (3.3.12)$$

Заметим, что для двоичного канала достаточно задать только две канальных вероятности, поскольку в канальной матрице

$$\mathbf{P} = \begin{pmatrix} p(y_1'/y_1) & p(y_2'/y_1) \\ p(y_1'/y_2) & p(y_2'/y_2) \end{pmatrix}$$

сумма вероятностей в каждой строке равна 1, и, тем самым, незаданные вероятности $p(y_1'/y_1)$ и $p(y_2'/y_2)$ выражаются через (3.3.12) как

$$p(y_1'/y_1) = 1 - p(y_2'/y_1) = 1 - p_2,$$
$$p(y_2'/y_2) = 1 - p(y_1'/y_2) = 1 - p_1.$$

Отсюда могут быть найдены безусловные вероятности $p(y_1')$ и $p(y_2')$ выходных символов:

$$p(y_1') = p\left(y_1'/y_1\right)p(y_1) + p\left(y_1'/y_2\right)p(y_2) = (1-p_2)p + p_1 q;$$

$$p(y_2') = p\left(y_2'/y_1\right)p(y_1) + p\left(y_2'/y_2\right)p(y_2) = p_2 p + (1-p_1)q.$$

Тогда взаимная информация для такого канала имеет следующий вид:

$$I\left(Y,Y'\right) = H(Y') - H(Y'/Y) =$$

$$= -\left[p(1-p_2) + qp_1\right]\log\left[p(1-p_2) + qp_1\right] -$$

$$-\left[pp_2 + q(1-p_1)\right]\log\left[pp_2 + q(1-p_1)\right] + p(1-p_2)\log(1-p_2) +$$

$$+ pp_2\log p_2 + qp_1\log p_1 + q(1-p_1)\log(1-p_1). \qquad (3.3.13)$$

Поскольку в данном случае вероятность p фактически является единственной переменной (учитывая, что $q = 1 - p$), для нахождения оптимального значения p_{opt} достаточно продифференцировать обыкновенным образом полученное выражение по p и приравнять производную нулю. В результате получим следующее выражение:

$$p_{\text{opt}} = \frac{1}{1-p_1-p_2}\left\{\frac{1}{1+\exp\left[-a/(1-p_1-p_2)\right]} - p_1\right\}, \qquad (3.3.14)$$

где введено обозначение

$$a = p_2\log p_2 + (1-p_2)\log(1-p_2) - p_1\log p_1 - (1-p_1)\log(1-p_1).$$

Подстановка (3.3.14) в (3.3.13) даёт значение пропускной способности канала при произвольном соотношении между входными вероятностями $p(y_1)$ и $p(y_2)$. В общем случае получается весьма громоздкое выражение, приведение которого здесь вряд ли целесообразно. Заметим лишь, что пропускная способность двоичного несимметричного канала равна нулю, если канальные вероятности связаны соотношением $p_1 + p_2 = 1$, ибо при этом априорные вероятности совпадают с апостериорными, и принятые символы не содержат никакой информации о переданных символах. Рассмотрим некоторые частные случаи.

При $p_1 = p_2$ имеет место двоичный симметричный канал, для которого оптимальная вероятность равна 1/2, а пропускная способность определяется выражением (3.3.4).

На рис. 3.3 показан случай, когда $p_2 = 0$, т. е. один из символов (в данном случае — y_1) всегда принимается правильно. Выражение для оптимальной вероятности имеет следующий вид:

$$p_{\text{opt}} = \frac{1}{1 - p_2 + \left(1/p_2\right)^{p_2/(1-p_2)}}. \qquad (3.3.15)$$

По терминологии Л. М. Финка символ y_1 является “надёжным”, поскольку он всегда принимается безошибочно, а символ y_2 — “достоверным”, так как приём именно данного символа означает достоверность передачи (приняв его, можно с полной достоверностью утверждать, что именно этот символ передавался).

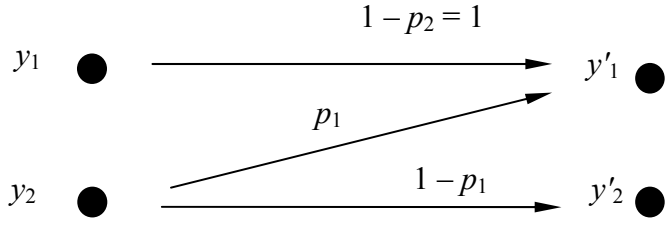


Рис. 3.3. Двоичный несимметричный канал

Например, при $p_1 = 0$ и $p_2 = 0{,}5$ оптимальными входными вероятностями оказываются $p(y_1) = 0{,}4$ и $p(y_2) = 0{,}6$. При этом пропускная способность такого канала составляет $0{,}322v$ бит / с.

3.4. ПРОПУСКНАЯ СПОСОБНОСТЬ ПРОИЗВОЛЬНОГО ДИСКРЕТНОГО ПОСТОЯННОГО КАНАЛА

Перейдём к вычислению пропускной способности дискретного постоянного канала, имеющего алфавиты $Y = \{y_1, \ldots, y_m\}$ на входе и $Y' = \{y'_1, \ldots, y'_{m'}\}$ на выходе соответственно.

Целью является нахождение максимального значения средней взаимной информации

$$I(Y', Y) = \sum_{Y'} \sum_{Y} p\left(y'_j / y_i\right) p(y_i) \log \frac{p\left(y'_j / y_i\right)}{\sum_{Y} p\left(y'_j / y_i\right) p(y_i)} \qquad (3.4.1)$$

по всем возможным распределениям $p(Y)$, являющимся точками в m-мерном пространстве с декартовыми координатами $p(y_1)$, …, $p(y_m)$ при наличии ограничений

$$\sum_{i=1}^{m} p(y_i) = 1,$$

и

$$p(y_i) \geq 0,\ i = 1, \ldots, m.$$

С математической точки зрения задача сводится к поиску максимума выпуклой функции в выпуклой области, образованной вероятностными распределениями на входе канала. Общие алгоритмы для численного вычисления с заданной точностью максимума выпуклой функции для случая конечного числа аргументов при наличии ограничений-равенств и ограничений-неравенств хорошо известны, и, в принципе, они могут быть использованы для решения поставленной задачи. Однако учёт специфических свойств средней взаимной информации позволяет вывести алгоритм, который оказывается существенно проще стандартных алгоритмов нахождения экстремума функций многих переменных.

В данном разделе будет рассмотрен аналитический подход, позволяющий в некоторых случаях найти решение задачи в замкнутой форме.

Согласно методу неопределённых множителей Лагранжа необходимым условием того, что функция достигает своего экстремума (может быть, локального), является система из m уравнений

$$\frac{\partial}{\partial p(y_i)}\left[I(Y', Y) + \lambda\left(\sum_{a=1}^{m} p(y_a) - 1 \right) \right] = 0, \quad i = 1, \ldots, m, \tag{3.4.2}$$

где λ — неизвестная постоянная, определяемая в дальнейшем из условия нормировки. Имеем[1]:

$$\frac{\partial H\left[Y'(Y) \right]}{\partial p(y_i)} = \sum_{j=1}^{m'} \frac{\partial H(Y')}{\partial p(y'_j)} \frac{\partial p(y'_j)}{\partial p(y_i)} =$$

$$= -\sum_{j=1}^{m'} \left[\log p(y'_j) + \log e \right] p(y'_j / y_i) = -\left[\log e + \sum_{j=1}^{m'} p(y'_j / y_i) \log p(y'_j) \right],$$

[1] Применение метода Лагранжа для отыскания экстремальных точек, строго говоря, возможно лишь для внутренних точек областей существования переменных, и его использование по отношению к граничным точкам, в частности, нулевых значений вероятностей $p(y_i)$ (см. далее) требует дополнительного обоснования, выходящего за рамки данного пособия.

$$\frac{\partial H\left[Y'/Y\right]}{\partial p\left(y_i\right)} = -\sum_{j=1}^{m'} p\left(y'_j/y_i\right)\log p\left(y'_j/y_i\right),\ \frac{\partial}{\partial p\left(y_i\right)}\sum_{a=1}^{m} p\left(y_a\right) = 1,$$

и после подстановки частных производных в соотношение (3.4.2) получаем набор уравнений

$$\sum_{j=1}^{m'} p\left(y'_j/y_i\right)\log\frac{p\left(y'_j/y_i\right)}{p\left(y'_j\right)} + \lambda - \log e =$$

$$= I\left(Y', y_i\right) + \lambda - \log e = 0,\ i = 1,\ldots,m. \qquad (3.4.3)$$

Для определения неизвестной константы $\lambda - \log e$ умножим (3.4.3) на $p\left(y_i\right)$ и просуммируем по всем *i*, тогда

$$\sum_{i=1}^{m} I\left(Y', y_i\right) p\left(y_i\right) = I\left(Y', Y\right) = \log e - \lambda, \qquad (3.4.4)$$

т. е. $\lambda - \log e$ является значением количества взаимной информации в точке максимума, которая по определению есть величина *C*/*v* — нормированное значение пропускной способности канала.

Соотношение (3.4.4) означает, что распределение вероятностей, максимизирующее среднюю взаимную информации между входом и выходом, должно обладать тем свойством, что для каждого входного символа количество взаимной информации между таким символом и всем выходным источником равняется одному и тому же числу. Необходимость этого условия интуитивно понятна, ибо если бы некоторый символ имел б*о*льшую взаимную информацию с выходом канала, то среднее количество взаимной информации могло быть увеличено за счёт более частого использования такого символа.

Соотношение (3.4.3) удобно переписать в виде

$$\sum_{j=1}^{m'} p\left(y'_j/y_i\right)\left[C/v + \log p(y'_j)\right] = -H\left(Y'/y_i\right),\ i = 1,\ldots,m, \qquad (3.4.5)$$

где вероятности выходных символов

$$p(y'_j) = \sum_{i=1}^{m} p\left(y'_j/y_i\right)p(y_i),\ j = 1,\ldots,m'$$

подчиняются условию нормировки

$$\sum_{j=1}^{m} p\left(y'_j\right) = 1.$$

Итак, для того, чтобы определить входное распределение $P(Y)$, на котором реализуется пропускная способность канала необходимо решить систему, содержащую $m + m' + 1$ уравнений с таким же числом неизвестных (с учётом пока неизвестного значения C / v).

С формальной точки зрения эта система хотя и является в общем случае трансцендентной, но, тем не менее, при современном уровне развития численных методов не вызывает принципиальных затруднений, поскольку показано, что она имеет единственное решение относительно C/v. Однако при этом существует один важный момент, связанный с тем, что решением должны являться неотрицательные (или, даже, строго положительные) числа. Таким образом, если в процессе решения все полученные значения $p(y_1), \ldots, p(y_m)$ неотрицательны, то соответствующее значение $I(Y; Y')$ и будет являться пропускной способностью канала.

Если же хотя бы одно из решений $p(y_k)$ ($k = 1, \ldots, m$) оказалось меньше нуля, то это означает, что взаимная информация $I(Y; Y')$ принимает наибольшее допустимое значение на границе пересечения гиперплоскости, определяемой условием нормировки, и гиперплоскости $p(y_k) = 0$, т. е. достигается на вырожденном — с некоторыми нулевыми входными вероятностями — распределении, причём не существует никакого простого способа определения того, какой из символов должен быть исключён; это не обязательно должен быть символ, для которого вероятность оказалась отрицательной. В этом случае для отыскания пропускной способности канала надо последовательно исключать символы из алфавита, приравнивая нулю каждую из вероятностных компонент в получившемся оптимальном распределении, и решать m систем вида (3.4.5), в каждой из которых одна из вероятностей $p(y_k)$ равна нулю. Тогда наибольшее среди всех систем решение и будет являться пропускной способностью канала, опять-таки, при условии, что все полученные вероятности неотрицательны. Если же решения для всех m систем содержат хотя бы по одному отрицательному компоненту $p(y_k)$, то необходимо вернуться назад к первоначальной системе и приравнять нулю одновременно по две компоненты в оптимальном распределении и т. д. Процедура продолжается до тех пор, пока не будет найдено оптимальное распределение или показано, что его не существует.

Описанный процесс нахождения пропускной способности существенно упрощается для одного частного, но практически важного случая, когда

число символов во входном и выходном алфавитах совпадает ($m = m'$), и канальная матрица **P** является невырожденной, т. е. существует обратная ей матрица

$$\mathbf{Q} = \mathbf{P}^{-1} = \| q_{ji} \| = \frac{\| T_{ij} \|}{\det(\mathbf{P})}, \tag{3.4.6}$$

где det(**P**) — определитель, а T_{ij} — алгебраические дополнения[1] элементов $p\left(y'_j / y_i\right)$ канальной матрицы.

Перепишем систему (3.4.5) в матричном виде:

$$\mathbf{P}\left[C / v + \log \mathbf{P}_{Y'}\right] = -\mathbf{H}\left(Y' / y\right),$$

где введена условная запись матричных элементов $\log \mathbf{P}_{Y'}$ и $-\mathbf{H}\left(Y' / y\right)$, компонентами которых являются числа $\log p(y'_j)$ и $-H\left(Y' / y_i\right)$ соответственно.

Умножим это матричное уравнение на обратную матрицу $\mathbf{P}^{-1}$ слева, получим

$$C / v + \log \mathbf{P}_{Y'} = -\mathbf{P}^{-1}\mathbf{H}\left(Y' / y\right),$$

или, возвращаясь к прежней форме,

$$C / v + \log p(y'_j) = -\sum_{i=1}^{m} q_{ji} H\left(Y' / y_i\right),\ i = 1,\ldots,m. \tag{3.4.7}$$

Потенцируя полученные уравнения и складывая их, имеем следующее выражение для пропускной способности канала:

$$C / v = \log_a \sum_{j=1}^{m} \exp_a \left(-\sum_{i=1}^{m} q_{ji} H\left(Y' / y_i\right) \right), \tag{3.4.8}$$

где, в целях упрощения записи формул введено, согласно [7], обозначение показательной функции $\exp_a(\zeta) \equiv a^{\zeta}$ ($a > 0$). В частности, потенцируя по основанию 2 (для того, чтобы оперировать с битами), имеем

$$C / v = \log_2 \sum_{j=1}^{m} \exp_2 \left(-\sum_{i=1}^{m} q_{ji} H\left(Y' / y_i\right) \right). \tag{3.4.8a}$$

---

[1] Алгебраическим дополнением матричного элемента a_{ji} называется число $A_{ji} = (-1)^{j+i} \overline{M_i^j}$, где $\overline{M_i^j}$ — *матричный минор*, т. е. определитель, полученный из исходного определителя матрицы посредством вычёркивания *j*-й строки и *i*-го столбца. Необходимые сведения по матричным вычислениям можно найти, например, в [4].

Аналогичным образом можно найти и вид входного распределения, на котором реализуется, пропускная способность.

Перепишем соотношение

$$p(y'_j) = \sum_{i=1}^{m} p\left(y'_j / y_i\right) p(y_i),\ j = 1, \ldots, m$$

в матричной форме:

$$\mathbf{P}_{Y'} = \mathbf{P}\mathbf{P}_Y .$$

Умножая это матричное уравнение слева на обратную канальную матрицу $\mathbf{P}^{-1}$, получаем

$$\mathbf{P}^{-1}\mathbf{P}_{Y'} = \mathbf{P}_Y ,$$

или

$$p(y_i) = \sum_{j=1}^{m} q_{ij}\, p(y'_j) =$$

$$= \exp_a(-C/v) \sum_{j=1}^{m} q_{ij} \exp_a\left[-\sum_{k=1}^{m} q_{jk} H\left(Y'/y_k\right) \right], \quad i = 1, \ldots, m. \quad (3.4.9)$$

Как нетрудно видеть, подстановка в выражение (3.4.8) канальных вероятностей вида (3.3.3), описывающих постоянный симметричный канал, даёт в итоге знакомое выражение (3.3.4).

Пусть, к примеру, постоянный канал задаётся одинаковыми троичными алфавитами на входе и выходе и канальной матрицей

$$\mathbf{P} = \begin{pmatrix} 6/8 & 1/8 & 1/8 \\ 1/4 & 2/4 & 1/4 \\ 1/8 & 2/8 & 5/8 \end{pmatrix}.$$

Матрица **Q**, обратная к канальной, имеет вид

$$\mathbf{P} = \begin{pmatrix} 1{,}455 & -0{,}273 & -0{,}182 \\ -0{,}727 & 2{,}636 & -0{,}909 \\ 0{,}000 & -1{,}000 & 2{,}000 \end{pmatrix},$$

а условные энтропии

$$\mathbf{H}\left(Y'/y\right) = -\sum_{j=1}^{m} \mathbf{P}\left(y'_j / y\right) \log \mathbf{P}\left(y'_j / y\right)$$

равны

$$H(Y'/y_1) = 1{,}063;\ H(Y'/y_2) = 1{,}500;\ H(Y'/y_3) = 1{,}299.$$

Произведя необходимые вычисления по (3.4.8) и (3.4.9), получим значение нормированной пропускной способности

$$C/v = 0{,}326$$

и входное распределение, на котором реализуется пропускная способность канала:

$$p(y_1) = 0{,}478;\ p(y_2) = 0{,}036;\ p(y_3) = 0{,}487.$$

3.5. КАНАЛЫ С ПАМЯТЬЮ

В предыдущих разделах данной главы изучались свойства постоянных дискретных каналов, т. е. каналов без памяти, в которых вероятности передачи символов в данный момент времени не зависят от того, какие символы передавались по каналу ранее. Значительно больший практический интерес представляют каналы с памятью, к которым относятся подавляющее большинство используемых на практике каналов, а постоянные каналы, в которых распределение числа ошибок подчиняется биномиальному распределению, являются лишь их модельными приближениями. Физической причиной того, что распределение числа ошибок в реальных каналах оказывается отличным от биномиального являются замирания сигналов, атмосферные, промышленные и коммутационные помехи, а также другие факторы естественного и искусственного происхождений.

Изучение каналов с памятью затруднено тем, что для их описания недостаточно знать только фиксированное число параметров, как это имеет место в постоянных каналах (для описания постоянного симметричного канала зачастую достаточно задать один параметр), а нужно уметь определять вероятности любых сочетаний ошибок в пределах последовательностей произвольной длины. Наиболее естественным представляется создание моделей каналов с памятью на основе экспериментальных измерений, однако обобщение полученных данных затрудняется тем, что не всегда удаётся подобрать удобное аналитическое представление: модель получается либо с очень сложной функциональной зависимостью, либо с большим количеством параметров. Исходя из этого, пытаются строить такие математические модели каналов с памятью, которые определяются небольшим числом параметров и соответствующий подбор которых позволяет хотя бы приближённо описать поведение реальных каналов.

Рассмотрим основные особенности каналов с памятью.

Если в постоянном канале условная вероятность $p_e(i+1/i)$ ошибочного приёма $(i + 1)$-го символа при условии, что i-й символ также принят ошибочно, равна безусловной вероятности ошибки p_e:

$$p_e(i+1/i) = p\,, \tag{3.5.1}$$

то в каналах с памятью она может быть как больше, так и меньше этого значения.

Подавляющее большинство реальных каналов удовлетворяет условию

$$p_e(i+n/i) \geq p\,, \tag{3.5.2}$$

для натуральных значений n, означающее тенденцию группирования ошибок, причём с ростом n неравенство (3.5.2) обычно приближается к равенству. Исходя из этого, такие каналы называются *каналами с группированием ошибок*.

Одной из задач, решаемых при рассмотрении характеристик каналов с группированием ошибок, является сравнение вероятности $p_e(i+n/i)$ с величиной $(m - 1)/m$, определяющей нулевое значение пропускной способности постоянного дискретного канала (1.3.4). Как правило, выполняется соотношение

$$p(i+n/i) < (m-1)/m\,, \tag{3.5.3}$$

и тогда канал называется *нормальным* в отличие от *аномального* канала, в котором вероятность может превышать величину $(m - 1)/m$.

Значительно реже каналов с группированием ошибок встречаются *каналы с рассредоточенными ошибками*, в которых

$$p_e(i+n/i) < p\,. \tag{3.5.4}$$

Примером такого канала может служить канал, в котором причиной ошибок является источник импульсных помех, обладающий тем свойством, что каждый импульс поражает только один символ, а вероятность появления следующего импульса сразу после предыдущего мала и возрастает со временем.

Кроме упомянутых типов каналов с памятью встречаются и такие, у которых для одних значений n справедливо неравенство (3.5.2), а для других — (3.5.3). Так, если (3.5.2) выполняется при нечётных n, а (3.5.3) — при чётных n, то в канале имеется тенденция к сдваиванию ошибок.

Простейшей моделью канала с памятью, описывающей (хотя бы в первом приближении) указанные свойства, является представление после-

довательности ошибок в виде простой марковской цепи, когда вероятность того, что некоторой символ будет принят ошибочно, равна некоторой величине p_1, при условии, что предыдущий символ был принят правильно, и другой величине p_2, если предыдущий символ был принят ошибочно. При $p_1 < p_2$ имеет место группирование ошибок; при $p_1 > p_2$ — их рассредоточение. Безусловная вероятность ошибки p в таком канале удовлетворяет уравнению

$$p = pp_2 + (1 - p)p_1$$

откуда

$$p = \frac{p_1}{1 + p_1 - p_2}. \qquad (3.5.5)$$

При использовании простой марковской модели легко вычисляется вероятность любого сочетания ошибок, однако такая модель чрезвычайно упрощённо воспроизводит свойства реальных каналов с группированием ошибок, поэтому в настоящее время она практически вышла из употребления.

Аналогично тому, как это делалось при моделировании письменной речи, представляется перспективным использование марковских цепей более высоких порядков, когда вероятность ошибочного приёма данного символа определяется n предыдущими символами. Однако, как уже говорилось, при больших значениях n сложность вычислений даже при современном уровне развития средств вычислительной техники становится непреодолимой, и вряд ли ситуация принципиально изменится в ближайшее время. Более успешной в использовании оказалась *модель Джильберта*.

Согласно модели Джильберта, канал может находиться в двух состояниях[1]: S_1 и S_2. В состоянии S_1 ошибок не происходит, а в состоянии S_2 происходит порождение пакетов ошибок, причём внутри пакета ошибки возникают независимо с вероятностью p_0. При передаче очередного символа известны вероятности q_{12} перехода из состояния S_1 в S_2 и q_{21} — перехода из состояния S_2 в S_1. Таким образом, в модели Джильберта простую марковскую цепь образует не последовательность ошибок, а последовательность состояний. Подчеркнём, что независимость ошибок внутри пакета является определяющим фактором для всей джильбертовской модели.

[1] В англоязычной литературе эти состояния обычно обозначаются буквами *G* (good — хороший) и *B* (bad — плохой или burst — пакет).

Если влияние канала с шумом трактовать как действие некоторого шумового генератора, порождающего дискретные символы z_i ($i = 1, \ldots, m$), то символы на выходе канала можно записать в виде (сложение осуществляется по модулю 2)

$$y_i' = y_i + z_i, \tag{3.5.6}$$

где в состоянии S_1 всегда $z_i = 0$, а в состоянии S_2 длительные пакеты из нулей должны чередоваться с относительно короткими пакетами из единиц. Чтобы это действительно имело место, необходима ситуация, когда цепь, находясь в состояниях S_1 или S_2, имела тенденцию оставаться в прежнем состоянии, для чего необходимо, чтобы вероятности переходов q_{12} и q_{21} были бы относительно малыми, а вероятности $p_1 = 1 - q_{12}$ и $p_2 = 1 - q_{21}$ остаться в прежних состояниях — относительно большими.

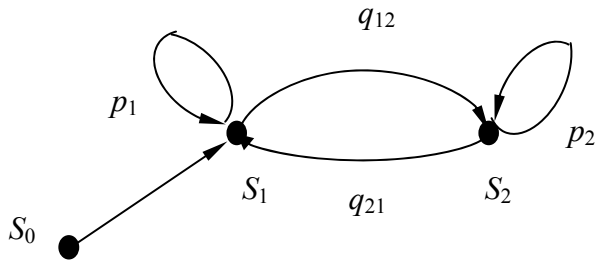


Рис. 3.4. Марковская цепь для шумового генератора в модели Джильберта

На рис. 3.4 представлена схема переходов для марковской цепи по модели Джильберта. Состояние S_0 является начальным, и разумно предположить, что из начального состояния цепь переходит в состояние с отсутствием ошибок, т. е. в состояние S_1, порождая при этом нулевой символ.

Распределение длин пакетов ошибок может быть, вообще говоря, произвольным. В простейшем случае можно предположить, что различные помехи, являющиеся причиной ошибок в реальных каналах, не связаны друг с другом.

Типичная 500-элементная выборка с параметрами $q_{12} = 0{,}03$, $q_{21} = 0{,}25$ и $p_0 = 0{,}5$, реализованная генераторами случайных чисел на основе описанной модели, имеет следующий вид [27]:

$$0^{62}110^{17}10^{46}110101110^{11}110^{15}10^{42}10^{28}110^{90}10^{37}110^{5}10010^{35}\ldots$$
$$1011010^{23}110^{4}10^{18}10^{15}11011101101110^{5} \tag{3.5.7}$$

Верхние индексы являются индикаторами длин пакетов, т. е., например, 0^{42} означает пакет из 42 последовательно идущих нулей. Как и предполагалось, достаточно длительные серии, состоящие из неискажённых символов, разделяются пакетами ошибок.

Нетрудно подсчитать вероятности пребывания канала в состояниях S_1 и S_2 для простой марковской цепи:

$$p_1 \equiv P(S_1) = \frac{q_{21}}{q_{12} + q_{21}},\ p_2 \equiv P(S_2) = \frac{q_{12}}{q_{12} + q_{21}}, \tag{3.5.8}$$

а безусловная вероятность ошибки p оказывается равной

$$p = p_0 P(S_2) = p_0 \frac{q_{12}}{q_{12} + q_{21}}. \tag{3.5.9}$$

Поставим задачу о нахождении пропускной способности канала для модели Джильберта.

При максимизации средней взаимной информации

$$I_n = I\left(Y^n; Y'^n\right) = H\left(Y'^n\right) - H\left(Y'^n / Y^n\right),$$

когда передаются n-элементные последовательности двоичных символов, важно понимать то, что значение условной энтропии $H\left(Y'^n / Y^n\right)$ выходных последовательностей $\mathbf{y}'$ при известных входных последовательностях $\mathbf{y}$ определяется многомерным распределением шумовых символов z внутри последовательности $\mathbf{z} = \left(z^{(1)}, \ldots, z^{(n)}\right)$ и является одним и тем же для всех $\mathbf{y}$, т. е. от $\mathbf{y}$ не зависит:

$$H\left(Y'^n / Y^n\right) = H\left(Z^n\right) = \sum_{Z^n} P(\mathbf{z}) \log P(\mathbf{z}),$$

так что поиск максимума можно осуществлять только по безусловной энтропии:

$$\frac{C}{v} = \lim_{n\to\infty} \frac{1}{n} \left\{ \max_{P\left(Y^n\right)} H\left[Y'^n \left(Y^n\right) \right] - H\left(Z^n\right) \right\}.$$

Как уже было показано, в таких случаях максимальное значение $H\left[Y'^n \left(Y^n\right) \right]$ реализуется на независимых и равновероятных n-элементных последовательностях $\mathbf{y}'$, а сама энтропия принимает максимальное значение, равное единице. Итак,

$$\frac{C}{v} = \lim_{n\to\infty} \frac{1}{n}\left\{1 - H\left(Z^n\right)\right\}, \tag{3.5.10}$$

и задача практически сводится к нахождению отношения

$$-H_n = \frac{1}{n}\sum_{Z^n} P(\mathbf{z}) \log P(\mathbf{z}) \tag{3.5.11}$$

при больших значениях n, ибо аналитическое решение задачи о вычислении предела (3.5.10) даже для простейших случаев рассматриваемой модели, по-видимому, невозможно.

Учитывая, что в случае независимых и равновероятных шумовых символов энтропия появления очередного шумового символа $z^{(n+1)}$ при всех известных предшествующих символах $z^{(n)}$, …, $z^{(1)}$ равна единице, запишем логарифм совместной вероятности $\log P(\mathbf{z})$ следующим образом:

$$\log P(\mathbf{z}) = \log P\left(z^{(1)},\ldots,z^{(n)}\right) =$$

$$= \sum_{z^{(n+1)}=0}^{1} P\left(z^{(n+1)} / z^{(1)},\ldots,z^{(n)}\right) \log P\left(z^{(n+1)} / z^{(1)},\ldots,z^{(n)}\right). \tag{3.5.12}$$

При этом, если в произвольном месте пакета генерируемый символ $z^{(i)} = 1$ $(i = 1,\ldots,n)$, то это означает, что цепь остаётся в состоянии S_2 — в состоянии простой марковской цепи, когда не играет никакой роли вероятностная зависимость от всех предыдущих символов $z^{(1)}$, …, $z^{(i-1)}$. Следовательно, для всех $j \geq 1$ справедливо соотношение

$$P\left(z^{(i+1)},\ldots,z^{(i+j)} / z^{(1)},\ldots,z^{(i)},\ 1\right) = P\left(z^{(i+1)},\ldots,z^{(i+j)} /\ 1\right),$$

в частности, для вероятности

$$P\left(z^{(1)}\ldots,z^{(n)}\right) = P\left(z^{(n+1)} / z^{(1)},\ldots,z^{(i)},\ 1, z^{(i+1)},\ldots,z^{(n)}\right) =$$

$$= P\left(z^{(n+1)} /\ 1, z^{(i+1)},\ldots,z^{(n)}\right).$$

Иными словами, именно число последовательных нулей в конце (после единицы) блока $(z^{(1)},\cdots,z^{(n)})$ полностью определяет многомерное распределение $P(\mathbf{z})$, и энтропия H_n содержит 2^n слагаемых, каждое из которых есть одно из n чисел

$$P(1)\log P(1),\ P(10)\log P(10),\ \ldots,\ P\left(10^{n-1}\right)\log P\left(10^{n-1}\right),$$

где верхние показатели, как и прежде, означают длины серий из нулей. Используя этот факт, запишем энтропию H_n в виде

$$-H_n = \frac{1}{n}\sum_{i=0}^{n-1} P\left(10^i\right)\log P\left(10^i\right), \qquad (3.5.13)$$

и задача теперь сводится к вычислению вероятностей серий из нулей.

Прежде чем непосредственно заняться вычислением искомых вероятностей, перепишем выражение (3.5.13) в виде, в котором его анализ и расчёт вероятностей может существенно облегчиться.

Обозначим через $u(i)$ условную вероятность того, что вслед за единицей появится i нулей, т. е. $u(i) = P\left(0^i/1\right)$. Поскольку по предположению из начального состояния S_0 цепь попадает в состояние S_1 с отсутствием ошибок, условимся считать $u(0) = 1$. Тогда фигурирующие в (3.5.13) вероятности $P\left(10^i\right)$ можно представить как

$$P\left(10^i\right) = u(i)P(1) = u(i)p,$$

где $P(1) = p$ — безусловная вероятность ошибки, задаваемая (3.5.9).

Условная вероятность $u(i + 1)$ появления вслед за единицей серии из $i + 1$ нуля равна произведению условных вероятностей $u(i)$ и $P(0^i/1)$:

$$u(i+1) \equiv P\left(0^{i+1}/1\right) = P\left(0/10^{i+1}\right)P\left(0^i/1\right) = P\left(0/10^{i+1}\right)u(i),$$

откуда

$$P\left(0/10^{i+1}\right) = \frac{u(i+1)}{u(i)}.$$

Теперь, используя представление (3.5.12) для логарифма вероятности серии, можно записать:

$$-H_n = \frac{p}{n}\sum_{i=0}^{n-1}\left\{u(i+1)\log\frac{u(i+1)}{u(i)} + \left[u(i) - u(i+1)\right]\log\frac{u(i) - u(i+1)}{u(i)}\right\}. \qquad (3.5.14)$$

Здесь $\log u(0) = \log 1 = 0$, а также $u(n) = P\left(0^n/1\right) = 0$, поскольку в n-элементной последовательности вслед за единицей никак не могут идти n нулей (максимум $n - 1$ нуль), поэтому, после приведения подобных, выражение для $-H_n$ принимает следующий вид:

$$-H_n = \frac{p}{n}\sum_{i=0}^{n-1} v(i)\log v(i), \qquad (3.5.15)$$

где введена функция

$$v(i) = u(i) - u(i-1). \qquad (3.5.16)$$

Нетрудно видеть, что $v(i)$ является ничем иным, как вероятностью "времени возвращения единицы" $P(0^i1/1)$, т. е. вероятностью того, что, сгенерировав единицу в состоянии S_2, источник выйдет из этого состояния, а затем через i шагов вернётся в него и снова сгенерирует единицу. При этом нужно понимать, что на протяжении i шагов возвращений в состоянии S_2 может быть несколько, и генерирование единицы не обязательно должно произойти в первое же возвращение.

Для дальнейших вычислений удобно использовать аппарат производящих функций.

Для произвольной последовательности $\{a_k\}$, $k = 0, 1, \ldots$ производящая функция $F(z)$ определяется в виде степенного ряда

$$F(z) = \sum_{k=0}^{\infty} a_k z^k ,$$

где z — произвольное, вообще говоря, комплексное число. Если последовательность $\{a_k\}$ ограничена, то этот ряд сходится в круге $|z| < 1$.

В вероятностных приложениях в качестве $\{a_k\}$ обычно фигурирует совокупность вероятностей $\{p_k\}$, $k = 0, 1, \ldots$, с которыми дискретный случайный процесс $\xi(t)$ принимает значения ξ_k. Оперирование с производящими функциями зачастую позволяет существенно облегчить нахождение характеристик случайных процессов.

Так, если стационарный дискретный процесс имеет математическое ожидание $\mathbf{E}[\xi] = m$ и дисперсию $\mathbf{D}[\xi] = \sigma^2$, то нетрудно получить ряд соотношений, связывающих параметры процесса со значениями его производящей функции:

$$F(1) = \sum_{k=0}^{\infty} p_k = 1 ;$$

$$F'(1) = \sum_{k=0}^{\infty} k p_k = m ;$$

$$\sigma^2 = F''(1) + F'(1) - \left(F'(1)\right)^2 .$$

Весьма эффективным является применение производящих функций к *рекуррентным событиям*, обладающим, говоря нестрого, тем свойством, что после первого появления некоторого свойства Ω эксперимент продолжается как бы заново. При этом промежутки времени между последовательными появлениями Ω — взаимно независимые одинаково распределённые случайные величины.

Важным свойством рекуррентных событий является то, что они факторизуют производящие функции, т. е. если последовательность вероятностей $\{f_n\}$ того, что Ω впервые происходит при n-м испытании, имеет производящую функцию $F(z)$, то последовательность вероятностей $\left\{f_n^{(r)}\right\}$ того, что r-е появление свойства Ω происходит при n-м испытании, имеет производящую функцию $[F(z)]^r$.

Более подробно с теорией рекуррентных событий можно ознакомиться, например, в [9].

Применим аппарат производящих функция для нахождения вероятностей, фигурирующих в (3.5.13)–(3.5.16).

Прежде всего, найдём вероятности времён возвращения в состояние S_2. Пусть f_i означает условную вероятность первого возвращения в состояние S_2 на i-м шаге при условии, что в начальный момент имело место также состояние S_2:

$$f_i = P\left(S_1^{i-1} S_2 / S_2\right),$$

где S_1^{i-1} означает $(i-1)$-кратное пребывание в состоянии S_1.

Нетрудно видеть, что

$$f_1 = P\left(S_2 / S_2\right) = p_2;$$

$$f_2 = P\left(S_1 S_2 / S_2\right) = q_{21} q_{12};$$

$$f_i = P\left(S_1^{i-1} S_2 / S_2\right) = q_{21} p_1^{i-2} q_{12},\ i \geq 2.$$

Введём производящую функцию $F(t)$ для вероятностей f_i:

$$F(t) = \sum_{i=1}^{\infty} f_i t^i = p_2 t + q_{21} q_{12} \sum_{i=2}^{\infty} p_1^{i-2} t^i .$$

Поскольку $\sum_{i=0}^{\infty} \left(p_1 t\right)^i = 1/\left(1 - p_1 t\right)$ — сумма убывающей геометрической прогрессии, производящая функция имеет следующий вид:

$$F(t) = p_2 t + \frac{q_{21} q_{12} t^2}{1 - p_1 t}. \qquad (3.5.17)$$

Теперь через коэффициенты функции $F(t)$ можно выразить и другие вероятности поведения источника. Поскольку событие, состоящее в том, что s-е возвращение в состояние S_2 произойдёт на i-м шаге, является рекуррентным по отношению к первому возвращению в состояние S_2 на i-м шаге, вероятность $f_i^{(s)}$ такого события имеет производящую функцию

$$\sum_{i=1}^{\infty} f_i^{(s)} t^i = \left[F(t)\right]^s. \qquad (3.5.18)$$

Тогда, учитывая, что вероятность невозвращения в состояние S_2 за k шагов равна $q_{21} p_1^{k-1}$, вероятность $g(i, s)$ того, что окажется строго s возвращений в состояние S_2 за i шагов (при этом возвращение не обязательно должно быть именно на i-м шаге) равна

$$g(i,s) = f_i^{(s)} + \sum_{k=1}^{i-s} f_{i-k}^{(s)} q_{21} p_1^{(k-1)},$$

а соответствующая производящая функция имеет вид

$$\sum_{i=1}^{\infty} g(i,s) t^i = \left(1 + \frac{q_{21} t}{1 - p_1 t}\right) \left[F(t)\right]^s. \qquad (3.5.19)$$

Обратимся теперь к вероятностям времён возвращения единицы. После появления единицы в состоянии S_2 и выхода из этого состояния следующая единица должна сгенерироваться также в состоянии S_2, но не обязательно при первом же возвращении в S_2. Вероятность того, что следующая единица появится при s-м возвращении в S_2 на i-м шаге, есть

$$p_0 (1 - p_0)^{s-1} f_i^{(s)}.$$

Тогда различные вероятности времени возвращения единицы за $(i - 1)$ шаг равны

$$v(i-1) \equiv P\left(0^{i-1} 1 / 1\right) = \sum_{s=1}^{\infty} p_0 \left(1 - p_0\right)^{s-1} f_i^{(s)}.$$

Соотношение (1.5.18) даёт возможность получения производящей функции $V(t)$ для последовательности вероятностей $v(i)$. Умножая последнее равенство на $f_i^{(s)}$ и суммируя по всем i, имеем:

$$\sum_{i=0}^{\infty} \sum_{s=1}^{\infty} p_0 \left(1 - p_0\right)^{s-1} f_i^{(s)} t^i = p_0 \sum_{s=1}^{\infty} \left(1 - p_0\right)^{s-1} F^s(t) =$$

$$= p_0 F(t) \sum_{s=0}^{\infty} \left[\left(1 - p_0\right) F(t)\right]^s = \frac{p_0 F(t)}{1 - \left(1 - p_0\right) F(t)}.$$

С другой стороны[1],

$$\sum_{i=0}^{\infty} v(i-1) t^i = t \sum_{i=0}^{\infty} v(k) t^i = t V(t).$$

Таким образом, получена связь между производящими функциями $F(t)$ и $V(t)$ для вероятностей времён возвращения в состояние S_2 и возвращения единицы:

$$tV(t) = \frac{p_0 F(t)}{1 - (1 - p_0) F(t)}. \qquad (3.5.20)$$

[1] В производящей функции $V(t)$ формально фигурирует нулевая вероятность $v(-1)$.

Аналогичным образом можно получить связь между производящими функциями $F(t)$ и $U(t) = \sum_{i=1}^{\infty} u(i)t^i$.

Поскольку $u(i)$ есть вероятность того, что единица не появится в течение i шагов, она может быть представлена как

$$u(i) = \sum_{s=1}^{\infty} g(i,s)(1-p_0)^s .$$

Тогда, проводя для $u(i)$ такие же операции, которые были проведены для $v(i)$, получим:

$$U(t) = \frac{1+(q_{21}-p_1)t}{(1-p_1t)\left[1-(1-p_0)F(t)\right]}, \qquad (3.5.21)$$

и осталось только подставить вместо $F(t)$ ее выражение (3.5.16) через параметры p_1, p_2, q_{12} и q_{21}. Например, для $U(t)$ имеем:

$$U(t) = \frac{1+(q_{21}-p_1)t}{1-\left[p_1+p_2(1-p_0)\right]t-(1-p_0)(q_{21}-p_1)t^2}. \qquad (3.5.22)$$

Теперь, если разложить функцию $U(t)$ в ряд по степеням t^i, то соответствующие коэффициенты и будут искомыми значениями $u(i)$. Для того, чтобы осуществить степенное разложение, стоящий в знаменателе квадратичный трёхчлен целесообразно представить в виде произведения сомножителей $(1 - A_1t)\,(1 - A_2t)$, где[1]

$$2A_{1,2} = p_1 + p_2(1-p_0) \pm \sqrt{\left[p_1+p_2(1-p_0)\right]^2 + 4(1-p_0)(p_{21}-p_1)},$$

и производящая функция $U(t)$ принимает следующий вид:

$$U(t) = \frac{1+(q_{21}-p_1)t}{A_1-A_2}\left(\frac{A_1}{1-A_1t} - \frac{A_2}{1-A_2t}\right). \qquad (3.5.22a)$$

Вновь используя представление суммы убывающей геометрической прогрессии, имеем

$$U(t) = \frac{1+(q_{21}-p_1)t}{A_1-A_2}\left(A_1\sum_{i=0}^{\infty}(A_1t)^i - A_2\sum_{i=0}^{\infty}(A_2t)^i\right) =$$

$$= \frac{1}{A_1-A_2}\left(A_1-A_2+\sum_{i=1}^{\infty}\left[(A_1+q_{21}-p_1)A_1^i-(A_2+q_{21}-p_1)A_2^i\right]t^i\right) =$$

[1] Минус перед корнем относится к меньшему, а плюс — к большему значениям.

$$= \frac{1}{A_1 - A_2} \sum_{i=0}^{\infty} \left[\left(A_1 + q_{21} - p_1 \right) A_1^i - \left(A_2 + q_{21} - p_1 \right) A_2^i \right] t^i .$$

Итак, вероятности $u(i)$ того, что единица не появится на протяжении i шагов, определяются выражением

$$u(i) = \frac{1}{A_1 - A_2} \left[\left(A_1 + q_{21} - p_1 \right) A_1^i - \left(A_2 + q_{21} - p_1 \right) A_2^i \right] \qquad (3.5.23).$$

При их практическом вычислении (особенно, для больших i) целесообразно использовать рекуррентные формулы, связывающие значения u в различные моменты. Для их нахождения запишем (3.5.22) в виде

$$\sum_{i=0}^{\infty} u(i) t^i \left\{ 1 - \left[p_1 + p_2 \left(1 - p_0 \right) \right] t - \left(1 - p_0 \right) \left(q_{21} - p_1 \right) t^2 \right\} = 1 + \left(p_2 - p_1 \right) t ,$$

затем разложим обе части в степенной ряд и сравним соответствующие коэффициенты при t^i, после чего имеем следующие соотношения:

$$u(0) = 1 \text{ (уже известное значение);}$$

$$u(1) - \left[p_1 + p_2 (1 - p_0) \right] u(0) = q_{21} - p_1 ;$$

$$u(k) - \left[p_1 + p_2 (1 - p_0) \right] u(k-1) - (1 - p_0)(p_2 - p_1) u(k-2) = 0 .$$

Таким образом, при начальных значениях

$$u(0) = 1,\ u(1) = q_{21} + p_2 (1 - p_0)$$

имеет место рекуррентная связь

$$u(i) = \left[p_1 + p_2 (1 - p_0) \right] u(i-1) + (1 - p_0)(p_2 - p_1) u(i-2) . \qquad (3.5.24)$$

Аналогичным образом получаются формулы для вероятности $v(i)$ времени возвращения единицы:

$$v(i) = \frac{p_0}{A_1 - A_2} \left[\left(p_2 A_1 + q_{21} - p_1 \right) A_1^i - \left(p_2 A_2 + q_{21} - p_1 \right) A_2^i \right], \qquad (3.5.25)$$

при этом рекуррентное соотношение имеет такой же вид

$$v(i) = \left[p_1 + p_2 (1 - p_0) \right] v(i-1) + (1 - p_0)(p_2 - p_1) v(i-2) . \qquad (3.5.26)$$

но с другими начальными значениями:

$$v(0) = p_0 p_2,\ v(1) = p_0 \left(q_{21} q_{12} + p_2^2 (1 - p_0) \right).$$

Итак, наконец получены явные выражения (3.5.23)–(3.5.26) для расчёта вероятностей серий из нулей. Эти формулы содержат достаточное большое число параметров, которые определяют и количественное и качественное поведение канала. Задаваясь определёнными значениями пара-

метров, можно получить бесконечное число формальных вероятностных моделей, которые могут оказаться в той или иной мере адекватными реальным анализируемым каналам. Например, для каналов, в которых превалируют коммутационные помехи, обычно полагают $p_0 = 1/2$, трактуя состояние S_2 как обрыв связи, а состояние S_1 — свободным от ошибок.

Другим способом является метод статистической подгонки, когда по наблюдаемым измерениям в реальных каналах делается (усреднённая) статистическая оценка параметров. Однако в этом случае параметры модели p_0, q_{21} и q_{12} непосредственно не могут наблюдаться, и их приходится оценивать косвенным образом через какие-то наблюдаемые параметры, а затем производить пересчёт. В качестве таких наблюдаемых параметров удобно взять следующие статистики:

$$\alpha = P(1);\ \beta = P(1/1);\ \gamma = P(111)/(P(101) + P(111)).$$

Здесь α определяет (как статистический эквивалент) безусловную вероятность появления единицы, β — условную вероятность того, что вслед за единицей вновь появится единица, а γ — условную вероятность того, что между двумя единицами также будет находиться единица. Можно показать, что в модели Джильберта эта вероятность равна

$$\frac{p_0 p_2^2}{p_2^2 + q_{21} q_{12}}.$$

Наблюдаемые параметры α, β и γ связаны с ненаблюдаемыми параметрами модели p_0, q_{21} и q_{12} следующими соотношениями:

$$q_{21} = 1 - \frac{\alpha\gamma - \beta^2}{2\alpha\gamma - \beta(\alpha + \gamma)};$$

$$p_0 = \frac{b}{1 - q_{12}};\ q_{12} = \frac{\alpha q_{21}}{p_0 - \alpha}. \qquad (3.5.27)$$

Заметим, что если положить $p_0 = 1/2$, то $1 - q_{21} = 2\beta$, и необходимость в измерении параметра γ отсутствует.

Например, в 500-элементной выборке (3.5.7) содержится 38 единиц, 15 двойных единиц, 7 комбинаций 101 и 3 комбинации 111. Тогда, учитывая, что в 15 случаях из 38 вслед за единицей вновь следует единица, для параметров α, β, γ имеем следующие статистические оценки:

$$\alpha = \frac{38}{500},\ \beta = \frac{15}{38},\ \gamma = \frac{3}{10}.$$

Однако в этом случае выясняется, что соотношения (3.5.27) дают бессмысленные значения параметров p_0, q_{21} и q_{12}, ибо q_{21} оказывается отрицательной. Причина этого заключается в том, что 500 символов образуют слишком малую выборку. В частности, оценка $\gamma = 3/10$, основанная на 10 наблюдениях, далека от истинного значения 0,49, изначально заложенного в модель.

Если теперь допустить, что $p_0 = 1/2$, то оценки получаются вполне приемлемыми: $q_{21} = 0{,}210$ и $q_{12} = 0{,}036$ (напомним точные значения: $q_{21} = 0{,}250$ и $q_{12} = 0{,}030$).

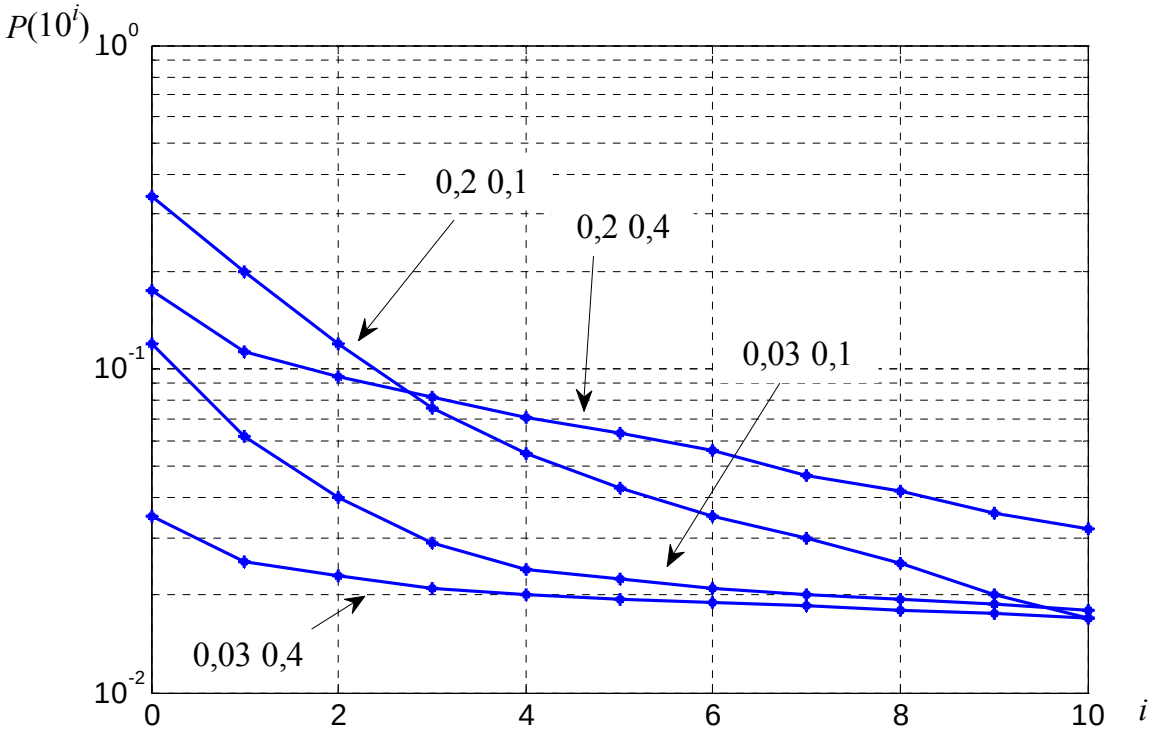


Рис. 3.5. Вероятности $P(10^i)$ при $p_0 = 1/2$

На рис. 1.5 показаны кривые вероятностей $P(10^i) = P(1)u(i)$ в зависимости от i для некоторых серий нулей с различными параметрами q_{21} и q_{12}. Видно, что эти кривые при больших значениях i имеют асимптоты, которые, как следует из (1.5.23), определяются величиной A_1.

Действительно, в соответствии с (1.5.24) и (1.5.26) вероятности $u(i)$ и $v(i)$ при больших i ведут себя подобно A_1^i:

$$u(i)=\frac{\left(A_1+q_{21}-p_1\right)A_1^i}{A_1-A_2}\left(1-\frac{\left(A_2+q_{21}-p_1\right)}{\left(A_1+q_{21}-p_1\right)}\left(\frac{A_2}{A_1}\right)^i\right)\approx$$

$$\approx\frac{\left(A_1+q_{21}-p_1\right)A_1^i}{A_1-A_2}.$$

При этом, как уже было замечено, для практически интересных случаев вероятность q_{12} мала, и A_1 близко к $p_1\approx 1$ (разумеется, всегда $A_1\geq p_1$). Это означает, что ряд (1.5.14) сходится достаточно медленно. Однако отношение

$$\frac{u(i+1)}{u(i)}=A_1\frac{1-\frac{\left(A_2+q_{21}-p_1\right)}{\left(A_1+q_{21}-p_1\right)}\left(\frac{A_2}{A_1}\right)^{i+1}}{1-\frac{\left(A_2+q_{21}-p_1\right)}{\left(A_1+q_{21}-p_1\right)}\left(\frac{A_2}{A_1}\right)^{i}}$$

стремится к значению A_1 ($A_1>A_2$, и стоящие в числителе и знаменателе дроби стремятся к нулю). Отсюда следует, что в выражении (3.5.14) компоненты

$$h_i=\frac{u(i+1)}{u(i)}\log\frac{u(i+1)}{u(i)}+\left[1-\frac{u(i+1)}{u(i)}\right]\log\left[1-\frac{u(i+1)}{u(i)}\right]$$

имеют своим пределом величину

$$h_0=-A_1\log A_1-\left(1-A_1\right)\log\left(1-A_1\right),$$

причём значения, близкие к предельным, начинают появляться достаточно быстро, практически при $i=i_0\approx 10$. Таким образом, можно считать, что $h_i=h_0$ для всех $i\geq 10$, и погрешность вычисления остатка ряда составляет порядка

$$\frac{h_0 p}{i_0}\frac{\left(A_1+q_{21}-p_1\right)A_1^{i_0}}{A_1-A_2}.$$

Итак, нахождение пропускной способности канала с памятью, описываемого моделью Джильберта, состоит в вычислении конечной суммы (3.5.13), где входящие в каждое слагаемое компоненты

$$\frac{u(i+1)}{u(i)}=A_1\frac{1-\frac{\left(A_2+q_{21}-p_1\right)}{\left(A_1+q_{21}-p_1\right)}\left(\frac{A_2}{A_1}\right)^{i+1}}{1-\frac{\left(A_2+q_{21}-p_1\right)}{\left(A_1+q_{21}-p_1\right)}\left(\frac{A_2}{A_1}\right)^{i}},$$

и $P(1)=p$ — безусловная вероятность ошибки, задаваемая (3.5.9), могут быть найдены по формулам (3.5.23) или (3.5.24).

Отметим, что данные о сериях нулей можно использовать в качестве других оценок параметров p_0, q_{21} и q_{12}, поскольку доля серий длины i является оценкой для $u(i)$. Тогда, в соответствии с (1.5.22), можно найти значения A_1, A_2 и B, такие, что

$$u(i) = BA_1^i + (1-B)A_2^i, \tag{3.5.28}$$

и они легко могут быть найдены посредством подбора кривой вида (1.5.28), аппроксимирующей измеренную серию. Конкретно, методика подбора следующая: сначала выбираются значения B и A_1, чтобы правильно определить поведение слагаемого BA_1^i для больших i, а затем подбирается значение A_2, чтобы улучшить соответствие при малых i.

Выражения p_0, q_{21} и q_{12} через A_1, A_2 и B имеют следующий вид:

$$p_0 = 1 - \frac{A_1 A_2}{A_1 - B(A_1 - A_2)},\ q_{12} = \frac{(1-A_1)(1-A_2)}{p_0},$$

$$q_{21} = B(A_1 - A_2) + (1-A_1)\frac{A_2 - (1-p_0)}{p_0}. \tag{3.5.29}$$

На рис. 3.6 представлены распределения долей нулевых серий для двух реальных телефонных каналов, по которым передавались двоичные данные.

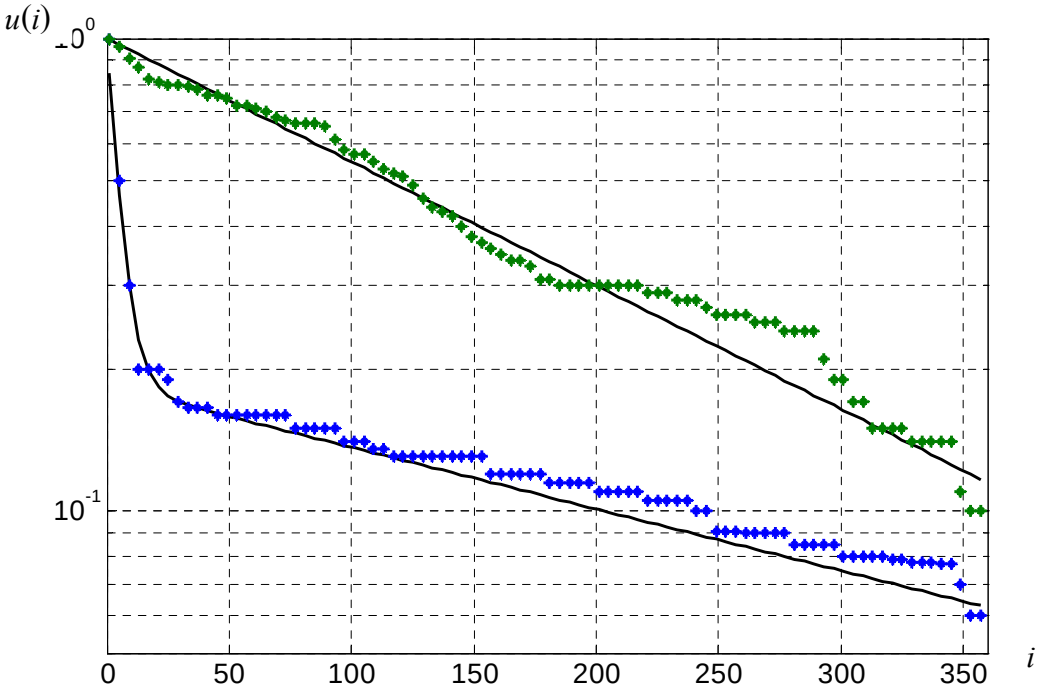


Рис. 3.6. Экспериментальные распределения нулевых серий

Экспериментальные точки представляют собой доли серии нулей длины i из выборки, содержащей около 130 последовательных серий из нулей для каждого канала, а

линии — это кривые вида (3.5.28), подобранные для этих распределений с соответствующими значениями параметров.

Для случая канала 1 значение $u(i) = 0{,}995^i$ (т. е. одно слагаемое в выражении (3.5.28)) обеспечивает вполне приемлемую аппроксимацию. Результаты для канала 2 в большей степени соответствуют данным, приведённым на рис. 3.6, где прямолинейная асимптота, выбираемая так, чтобы аппроксимировать данные для больших i — это функция BA_1^i с параметрами $B = 0{,}184$ и $A_1 = 0{,}997$, а значение параметра $A_2 = 0{,}810$ обеспечивает соответствие кривой (3.5.28) для малых значений i.

Вычисление согласно (3.5.29) приводит к следующим оценкам:

$$p_0 = 0{,}160;\ q_{12} = 0{,}003;\ q_{21} = 0{,}034.$$

Заметим также, что канал 1 также неплохо аппроксимируется двоичным симметричным каналом со значением средней вероятности ошибки $p = 0{,}995$.

В заключение данного раздела заметим, что в настоящее время существуют обобщения модели Джильберта, например *модель Беннета–Фрелиха*, согласно которой ошибки в канале возникают в виде пакетов, где первый и последний символы приняты ошибочно, а между ними могут располагаться и правильно, и ошибочно принятые символы. При этом предполагается, что пакеты возникают независимо друг от друга с заданной вероятностью p_1. Кроме того, канал характеризуется вероятностью p_2 появления ошибок внутри пакета и распределением $p(k)$ числа символов k в пачке. Подбирая значения p_1 и p_2, а также вид распределения $p(k)$, часто удаётся получить модельное описание канала, удовлетворительно согласующееся с экспериментальными данными. И модель Беннета–Фрелиха, и модель Джильберта, а также ряд других известных в литературе моделей [12] относятся к *формальным моделям*, т. е. таким, при построении которых не рассматриваются физические причины, вызывающие пакеты ошибок, а просто под наблюдаемые экспериментальные данные подбирается некоторая вероятностная схема. В отличие от формальных моделей существуют *физические модели*, в которых распределение ошибок выводится из вероятностных свойств самого канала.

ВОПРОСЫ И ЗАДАНИЯ К ГЛАВЕ 3

3.1. В чем различие между постоянными и непостоянными дискретными каналами?

3.2. Дайте определение пропускной способности произвольного дискретного канала.

3.3. Выберете правильный вариант ответа и обоснуйте выбор.

Дискретный канал задан если:

а) заданы множества X и Y на входе и выходе канала, и заданы условные распределения вероятностей $p(y/x)$, $y \in Y$, $x \in X$;

б) заданы множества X и Y на входе и выходе канала, и для любого натурального n заданы условные распределения $p(\mathbf{y}/\mathbf{x})$, $\mathbf{y} \in Y^n$, $\mathbf{x} \in X^n$;

в) заданы множества X и Y на входе и выходе канала, и для любого натурального n заданы условные распределения $p(\mathbf{y}/\mathbf{x})$, $\mathbf{y} \in Y^n$, $\mathbf{x} \in X^n$, и задано распределение $p(\mathbf{x})$;

г) среди вариантов *а*)–*в*) нет правильных.

3.4. В чем проявляется марковское свойство каналов?

3.5. Опишите вероятностные соотношения появления канальных ошибок в модели Джильберта.

3.6. Что такое канал со стиранием?

3.7. Априорные вероятности появления символов двоичного источника без памяти равны $p(x_1) = 0{,}1$ и $p(x_2) = 0{,}9$. Последовательность из 400 таких символов передаётся по постоянному дискретному каналу связи с матрицей

$$p(Y / X) = \begin{pmatrix} 0{,}98 & 0{,}02 \\ 0{,}3 & 0{,}7 \end{pmatrix}.$$

где $Y = (y_1, y_2)$ — символы на выходе канала. Определить объем переданной по каналу связи информации.

3.8. Рассматривается двоичный канал, т. е. канал, в котором символы на входе Y и выходе Y' принимают значения 0 и 1.

а). Показать, что канал, в котором многомерные канальные вероятности $p(\mathbf{y}'/\mathbf{y}) = 2^{-n}$, $\mathbf{y} \in Y^n$, $\mathbf{y}' \in Y'^n$ (n — произвольное натуральное число) является каналом без памяти и удовлетворяет условию стационарности.

б). Показать, что канал, в котором многомерные канальные вероятности $p(\mathbf{y}' / \mathbf{y}) = 2^{-n+1}$, $\mathbf{y} \in Y^n$, $\mathbf{y}' \in Y'^n$ и $y^{(1)} = 0$ является каналом без памяти, но не удовлетворяет условию стационарности.

3.9. Входные символы двух идентичных двоичных каналов искажаются в канале независимо с вероятностью p. Найти результирующую пропускную способность, если:

а) каналы соединены последовательно, т. е. символ, полученный на выходе одного канала, далее подаётся на вход другого канала;

б) каналы соединены параллельно, т. е. один и тот же символ передаётся по обоим каналам, хотя при этом оба принятых символа не обязательно должны совпадать.

3.10. Вычислить пропускную способность дискретных каналов, задаваемых следующими канальными матрицами:

$$а)\begin{pmatrix} 1-p & p & 0 \\ 0 & 1-p & p \\ p & 0 & 1-p \end{pmatrix};\ б)\ \frac{1}{2}\begin{pmatrix} 1-p & 1-p & p & p \\ p & p & 1-p & 1-p \end{pmatrix};$$

$$в)\begin{pmatrix} 1-p_1-p_2 & p_1 & p_2 \\ p_2 & 1-p_1-p_2 & p_1 \\ p_1 & p_2 & 1-p_1-p_2 \end{pmatrix};\ г)\begin{pmatrix} 1-p_1 & p_1 \\ p_2 & 1-p_2 \end{pmatrix};$$

$$д)\begin{pmatrix} 1-p & p & 0 & 0 \\ p & 1-p & 0 & 0 \\ 0 & 0 & 1-p & p \\ 0 & 0 & p & 1-p \end{pmatrix};\ е)\begin{pmatrix} 1-p & p & 0 \\ p & 1-p & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

3.11. Пусть C_1, C_2, …, C_n — пропускные способности n дискретных постоянных каналов. Определим суммарный канал как канал с входным и выходным алфавитами, являющимися объединением алфавитов отдельных каналов. Иными словами, наличие суммарного канала означает возможность использования в любой момент времени любого канала. Показать, что пропускная способность C суммарного канала равна

$$C = \log \sum_{k=1}^{n} 2^{C_k}.$$

3.12. Рассмотрим модель *двоичного симметричного канала* (ДСК) *с ограничением*. Пусть p — вероятность ошибки в ДСК с алфавитом $\{0, 1\}$ и $\delta < 0{,}5$ — некоторое заданное число. Найти пропускную способность канала C_δ, полученную путём максимизации средней взаимной информации по всем входным распределениям, для которых априорная вероятность $P\{1\} \le \delta$. Исследовать зависимость $C_\delta(p)$ и сравнить с аналогичной зависимостью для ДСК без ограничений.

ГЛАВА 4. ИДЕАЛЬНЫЙ ДИСКРЕТНЫЙ КАНАЛ

В предыдущей главе были ведены и изучены фундаментальные характеристики, касающиеся передачи информации по дискретным каналам. Данная глава посвящена рассмотрению частного, но не менее важного случая — идеального дискретного канала. Фактически речь пойдёт об основных идеях, связанных с методами *сжатия информации*. Вводятся основные понятия, связанные с возможностью использования равномерных и неравномерных кодов. Рассматривается достаточно большое число примеров алгоритмов кодирования, в том числе, оптимальные алгоритмы, позволяющие на определённом классе кодов достичь минимально возможных затрат.

4.1. РАВНОМЕРНОЕ КОДИРОВАНИЕ

Дискретный канал, для которого алфавиты кодовых символов Y на входе и Y' на выходе одинаковы, а вероятности переходов равны

$$p\left(y'_j / y_i\right) = \delta_{ij} = \begin{cases} 1, & i = j; \\ 0, & i \neq j, \end{cases} \qquad (4.1.1)$$

называется *идеальным дискретным каналом* или *каналом без шума.* Иными словами, в идеальном дискретном канале символы на выходе и входе всегда совпадают, и он полностью характеризуется основанием алфавита, а также технической скоростью v, т. е. количеством символов, пропускаемых в среднем в единицу времени.

Идеальные дискретные каналы в природе вряд ли существуют в том смысле, что для них всегда выполняется соотношение (4.1.1), они скорее являются модельными для тех реальных каналов, которые на практике обеспечивают хорошее качество передачи. В литературе (особенно, в технических спецификациях на телекоммуникационные и радиотехнические системы) существует понятие “практически идеальный канал”, или “канал, практически свободный от ошибок”, означающий, что при нормальных условиях работы системы за время, соответствующее передаче законченной порции сообщений (файла данных, теле- или радиовещательного кадра, блока команд управления или телеметрии и др.) ошибок не происходит.

Количественно понятие практически идеального канала описывается значением коэффициента ошибок порядка 10^{-5}, т. е. на 10^5 посланных "элементарных сообщений" (например, двоичных символов) допускаются единицы ошибок.

В идеальном дискретном канале количество информации $I(Y'; Y)$, содержащееся в принятых символах y'_j относительно переданных символов y_i, всегда равно энтропии источника $H(Y)$, поскольку всегда условная энтропия $H(Y / Y') = 0$. Следовательно, пропускная способность идеального канала реализуется входным алфавитом, энтропия которого при фиксированном объёме имеет максимальное значение:

$$C = \frac{1}{T} \max_{P(Y)} H(Y).$$

Таким алфавитом является алфавит с равновероятными элементами, для которого энтропия равна $\log m$ и

$$C = \frac{\log m}{T} = v \log m. \qquad (4.1.2)$$

Из полученного результата следует, что для безызбыточного источника, объём которого l равен объёму m алфавита кодовых символов, задача передачи информации без потерь со скоростью, равной пропускной способности канала, сводится попросту к установлению любым образом взаимно однозначного соответствия каждого из элементов сообщения x_k ($k = 1, \ldots, l$) какому-либо кодовому символу y_i ($i = 1, \ldots, m$).

Более интересным является случай, когда $l = m^a$, где a — натуральное число. Тогда способ достижения пропускной способности канала состоит в следующем.

Построим все возможные последовательности из кодовых символов y_i длиной a (очевидно, их число равно m^a) и установим взаимно однозначное соответствие между каждым элементом x_k и определённой кодовой комбинацией $(y^{(1)}, \ldots, y^{(a)})$. Такой код, в котором все кодовые комбинации имеют одинаковую длину a, называется *равномерным a-разрядным кодом*. Выбрав теперь скорость v_x ввода элементов x_k в кодер, равную $v_x = v / a$, получим скорость передачи информации

$$V \equiv I'(X) = v_x \log l = v_x \log m^a = \frac{v}{a} a \log m = v \log m,$$

равную пропускной способности канала.

Пусть теперь $l = m^b$, но при этом l не обязательно является целой степенью m, например, $l = 10$ (для определённости можно говорить о десятичных цифрах 0, 1, …, 9) и $m = 2$. Применим равномерное кодирование, выбирая четырехразрядный код и следующие кодовые комбинации:

0 — 0000; 1 — 0001; 2 — 0010; 3 — 0011; 4 — 0100;
5 — 0101; 6 — 0110; 7 — 0111; 8 — 1000; 9 — 1001.

Тогда при скорости ввода элементов в кодер $v_x = v / 4$ скорость передачи информации составляет

$$V = (v / 4)\log_2 10 \approx 0{,}831v,$$

что заметно меньше значения пропускной способности $C = v \log_2 2 = v$. Это и не удивительно, поскольку шесть из возможных 16 кодовых комбинаций не используются.

Попробуем приблизить значение скорости передачи информации к пропускной способности канала путём укрупнения источника, рассматривая каждые две буквы исходного источника как одну букву укрупнённого источника. Очевидно, объем l_1 нового источника равен $l_1 = l^2 = 100$. Тогда, выбирая семиразрядный равномерный код с кодовыми комбинациями

0 — 0000000; 1 — 0000001; 2 — 0000010; …
97 — 1100001; 98 — 1100010; 99 — 1100011

и скорость ввода элементов в кодер $v_x = v / 7$, получим значение скорости

$$V = (v / 7)\log_2 100 \approx 0{,}949v,$$

что уже гораздо ближе к искомому значению.

Продолжая дальнейшее укрупнение исходного источника, можно получить ещё большее приближение к C. Так, при укрупнении букв исходного источника в группы по 3, выбрав 10-разрядное кодирование и $v_x = v/10$, получим значение V составляющее $0{,}997v$. Видно, что получаемый от укрупнения источника выигрыш в скорости передачи уменьшается с каждым дальнейшим шагом укрупнения, в то время как сложность кодирования таких групп возрастает, причём, существенно нелинейно.

Кроме того, заметим, что поведение V при каждом очередном шаге укрупнения носит немонотонный характер: достижение новых максимумов сопровождается “откатываниями” вниз. На рис. 4.1 показано поведение нормированной скорости V/v для первых 30 шагов укрупнения, где максимальным является значение 0,997, полученное ещё на третьем шаге

укрупнения. Но уже на следующем, 31-м шаге появляется новый максимум 0,9998, который будет удерживаться ещё большее количество шагов.

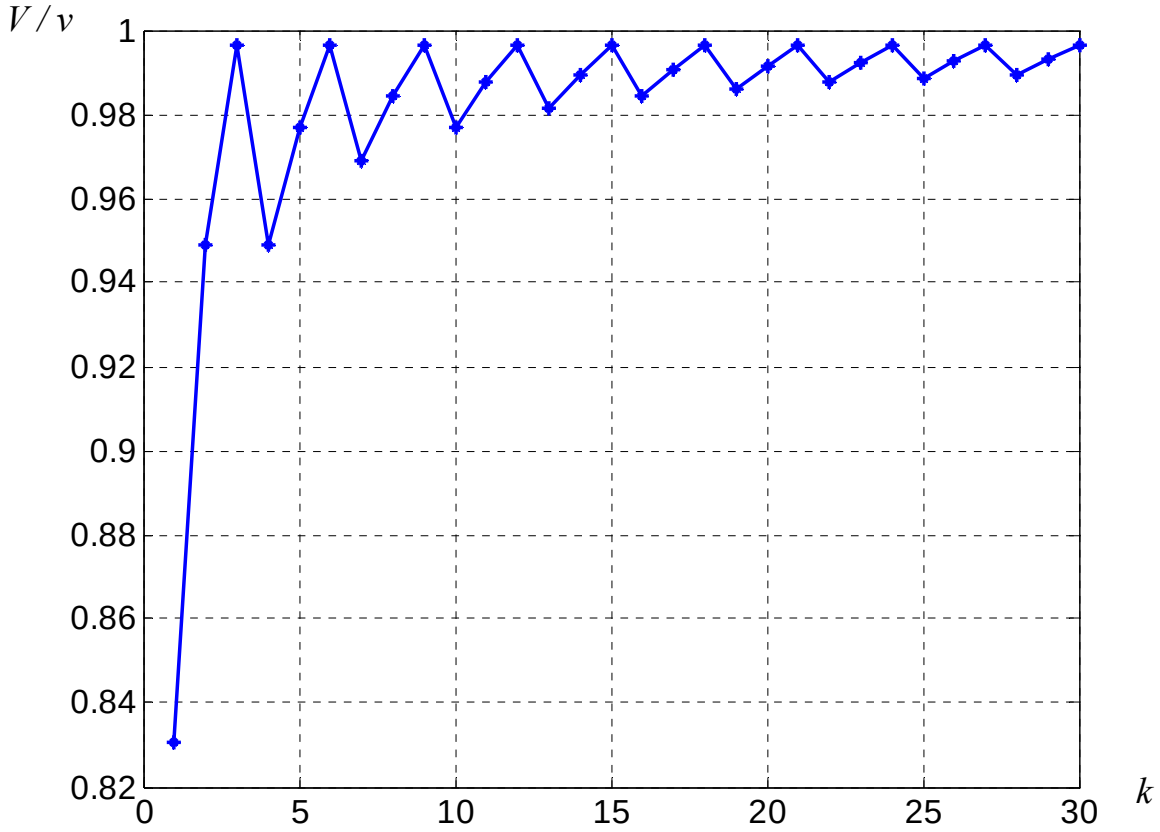


Рис. 4.1. Поведение *V*/*v* при укрупнении алфавита

В данном случае *точное* достижение пропускной способности *C* невозможно. Действительно, если объединить *k* символов, выдаваемых источником, в *k*-разрядное десятичное число, то его можно изобразить *a*-разрядным двоичным числом при условии, что

$$10^k \le 2^a,$$

но знак равенства невозможен, ибо при натуральных значениях *k* и *a* равенство

$$k \log_2 10 = a$$

означает, что число $\log_2 10$ является рациональным, а это не соответствует действительности. Поэтому скорость передачи неизбежно подчиняется соотношению

$$V = \frac{v}{a}\log_2 m^k = \frac{kv}{a} < \frac{v}{\log_2 10}.$$

Тем не менее, несмотря, на медленность и немонотонность поведения V, достижение пропускной способности *со сколь угодной заданной точностью* ε при равномерном кодировании возможно. Покажем это для произвольных значений алфавита источника l и канального алфавита m.

Очевидно, что для любого числа b всегда можно найти такое число a, при котором справедливы неравенства

$$b\frac{\log l}{\log m} < a < b\frac{\log l}{\log m} + 1. \qquad (4.1.3)$$

В соответствии с изложенным методом, будем рассматривать каждые b букв исходного источника как одну букву укрупнённого алфавита объёмом l^b элементов и установим взаимно однозначное соответствие укрупнённых букв кодовым комбинациям равномерного a-разрядного кода, что всегда можно сделать, поскольку $l^b < m^a$ на основании неравенств (4.1.3), соглашаясь с тем, что часть комбинаций неизбежно останется неиспользованной. Таким образом, каждая переданная по каналу комбинация из a канальных символов соответствует b буквам исходного источника, так что скорость передачи составляет

$$V = v\frac{b}{a} = \frac{Cb}{a\log_2 m},$$

где C — пропускная способность канала.

Из (4.1.3) имеем неравенство

$$b\log l < a\log m < \left(b + \frac{\log m}{\log l}\right)\log l,$$

которое можно записать в виде

$$a\log m = (b + c)\log l,\ 0 < c < \frac{\log m}{\log l}.$$

Отсюда

$$V = \frac{C}{\log_2 l}\frac{b}{b+c} = \frac{C}{H}\frac{b}{b+c} > \frac{C}{H}\frac{b}{b + (\log m/\log l)} =$$

$$= \frac{C}{H}\left(1 - \frac{(\log m/\log l)}{b + (\log m/\log l)}\right). \qquad (4.1.4)$$

Если теперь число b выбрать из условия

$$b \geq \left(\frac{C}{H\varepsilon} - 1 \right) \frac{\log m}{\log l},$$

то на основании (4.1.4) получим

$$V = \frac{C}{H} - \varepsilon,$$

что и доказывает требуемое.

Примером практически используемого равномерного кода является хорошо известный код **ASCII** (**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange, американский стандартный код для обмена информацией), устанавливающий соответствие между восьмиразрядными двоичными комбинациями и множеством наиболее часто употребляемых символов (буквы английского и национальных алфавитов, графические символы, различные служебные символы и т. п.).

Таблица 4.1

Таблица ASCII-кодов

	000	016	032	048	064	080	096	112	128	144	160	176	192	208	224	240
00		►		0	@	P	`	p	А	Н	а	▒	└	╨	р	Ё
01	☺	◄	!	1	A	Q	a	q	Б	С	б	▒	┴	╤	с	ё
02	☻	↕	«	2	B	R	b	r	В	Т	в	▓	┬	╥	т	Є
03	♥	‼	#	3	C	S	c	s	Г	У	г	│	├	╙	у	є
04	♦	¶	$	4	D	T	d	t	Д	Ф	д	┤	─	╘	ф	Ї
05	♣	§	%	5	E	U	e	u	Е	Х	е	╡	┼	╒	х	ї
06	♠	▬	&	6	F	V	f	v	Ж	Ц	ж	╢	╞	╓	ц	Ў
07	•	↨	‘	7	G	W	g	w	З	Ч	з	╖	╟	╫	ч	ў
08	◘	↑	(	8	H	X	h	x	И	Ш	и	╕	╚	╪	ш	°
09	○	↓	)	9	I	Y	i	y	Й	Щ	й	╣	╔	┘	щ	·
10	◙	→	*	:	J	Z	j	z	К	Ъ	к	║	╩	┌	ъ	·
11	♂	←	+	;	K	[	k	{	Л	Ы	л	╗	╦	█	ы	√
12	♀	∟	,	<	L	\	l	|	М	Ь	м	╝	╠	▄	ь	№
13	♪	↔	-	=	M	]	m	}	Н	Э	н	╜	═	▌	э	¤
14	♫	▲	.	>	N	^	n	~	О	Ю	о	╛	╬	▬	ю	■
15	☼	▼	/	?	O	_	o	⌂	П	Я	п	┐	╧	▀	я	

Как известно, в таком коде жёстко фиксированы лишь первые 128 комбинаций (с нулевой по 127-ю), а вторая часть таблицы, содержащая коды от 128 до 255, может варьироваться и предназначена для кодирования

символов различных национальных алфавитов, базирующихся не на латинских буквах, в частности кириллицы. К сожалению, единой общепринятой кодировки кириллических шрифтов не существует; наиболее распространены альтернативная кодировка **DOS KOI-8** (табл. 4.1) и кодировка Windows **WIN-1251**. В настоящее время уже внедрён шестнадцатиразрядный стандарт **UNICODE**, включающий **ASCII** как подмножество (с кодами от нуля до 127).

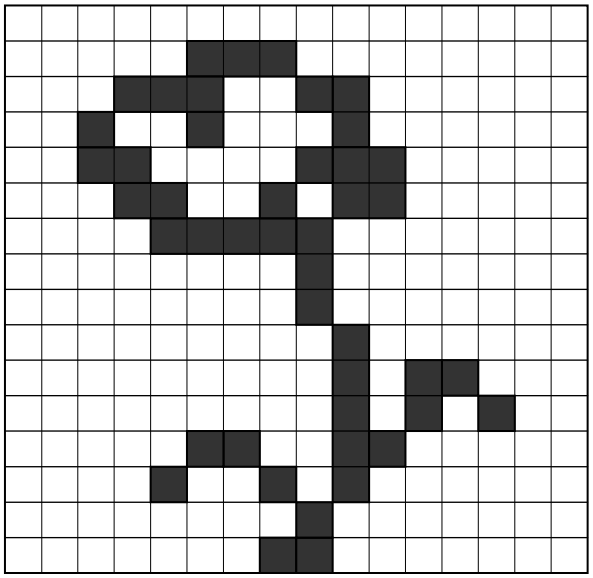


Рис. 4.2. Точечный (растровый) рисунок “цветок”

Другим примером равномерного двоичного кода является представление графических изображений. В простейшем случае двухцветная (традиционно — чёрно-белая) картинка может трактоваться как растр[1], т. е. совокупность отдельных точек или *пикселов*, расставленных на прямоугольной решётке. Сопоставляя чёрным точкам единицы, а белым — нули, можно представить графический объект в виде таблицы из нулей и единиц, которую можно “вытянуть” в одну линию, соединяя её строки или столбцы.

Рассмотрим представленную на рис. 4.2 сетку 16×16 и графический объект в ней. Кодируя столбцы целыми числами в диапазоне 0…65535 ($65535 = 2^{16} - 1$), получим (табл. 4.2) двоичные восьмиразрядные кодовые

[1] Исторически растром (от латинского rastrum — грабли) в полиграфии называли прямоугольную сетку, через которую подлежащее фотографированию тоновое изображение разбивалось на точки.

комбинации (*байты*), которые удобно представить в шестнадцатеричной системе (их обычно называют hex-коды).

Таблица 4.2

Двоичное представление растрового рисунка

№ столбца	Двоичный код	hex-код
1	0000 0000 0000 0000	0000
2	0000 0000 0000 0000	0000
3	0001 1000 0000 0000	1800
4	0010 1100 0000 0000	2C00
5	0010 0110 0000 0100	2604
6	0111 0010 0000 1000	7208
7	0100 0010 0000 1000	4208
8	0100 0110 0000 0101	4605
9	0010 1011 1000 0011	2883
10	0011 1100 0111 1100	3C7C
11	0000 1100 0000 1000	0C08
12	0000 0000 0011 0000	0030
13	0000 0000 0010 0000	0020
14	0000 0000 0001 0000	0010
15	0000 0000 0000 0000	0000
16	0000 0000 0000 0000	0000

Возможность “точечного” рисования изображений широко используется в вычислительной технике и в обработке изображений. Следует заметить, что, символы, представленные в табл. 4.2, также имеют растровое представление при изображении на мониторе или печатающем устройстве (принтере). При этом растр, в зависимости от конкретного устройства, трактуется либо как последовательность строк (монитор), либо как последовательность столбцов (матричный принтер, уже, видимо, ушедший в историю). На основе описанного представления разработано большое число форматов графических файлов.

4.2. ПРИНЦИПЫ НЕРАВНОМЕРНОГО КОДИРОВАНИЯ. НЕРАВЕНСТВО КРАФТА

Как было показано в предыдущем разделе, использование равномерного кодирования позволяет достаточно эффективно сжимать сообщения, имеющие близкие вероятности появления. Однако если вероятности появления отдельных сообщений источника существенно различаются между собой, т. е. источник обладает заметной избыточностью, то применение равномерных кодов оказывается малоэффективным, и возникает идея использовать неравномерные коды, у которых кодовые комбинации содержат различное число кодовых символов.

В общем случае, если для каждого символа источника требуется более одного кодового слова, то по необходимости кодовые символы (символы кодового алфавита) объединяются в *кодовые слова*. Множество допустимых слов образует *кодовый словарь*.

Для того чтобы иметь возможность сравнивать эффективность кодирования какого-либо источника различными кодовыми словарями необходимо, прежде всего, ввести некоторую количественную характеристику, отражающую эффективность кодирования. В качестве такой характеристики обычно рассматривают *среднюю длину кодовых слов*, или, короче, *среднюю длину кода*, определяемую как

$$\overline{n} = \sum_{k=1}^{l} n_k \, p(x_k), \qquad (4.2.1)$$

где $p(x_k)$ ($k = 1, \ldots, l$) — вероятности появления элементов источника, а n_k — это количество символов в кодовом слове w_k которым кодируется элемент x_k. Очевидно, что для равномерного кода средняя длина совпадает с длиной любой из его комбинаций.

Структура $\overline{n}$ согласно (4.2.1) подсказывает главную идею конструирования эффективных кодовых словарей: менее вероятные сообщения источника следует кодировать словами с бо́льшим числом кодовых символов, а более вероятные сообщения — словами с меньшим числом кодовых символов. При этом можно надеяться, что при большом количестве элементов источника $\overline{n}$ окажется ощутимо меньше, чем при кодировании равномерным a-разрядным кодом, и будет получен выигрыш кодирования $\gamma = a / \overline{n}$.

Однако, чтобы словарь имел практическое применение, он должен удовлетворять как минимум одному требованию — условию *однозначной декодируемости*.

Последовательность слов из словаря D называется однозначно декодируемой, если её невозможно интерпретировать как некоторую другую последовательность слов того же словаря.

Данное определение лучше всего проиллюстрировать на примере словаря, который не имеет этого свойства. Пусть словарь D имеет вид

$$D = \{w_1 = 0,\ w_2 = 10,\ w_3 = 1001\}.$$

Невозможно определить, является ли последовательность символов 10010 последовательностью слов w_3w_1 или же слов $w_2w_1w_2$. Такой словарь не является однозначно декодируемым и, следовательно, не может иметь практического значения. Разумеется, между словами неоднозначно декодируемого словаря можно поместить дополнительный элемент — “пробел”. Однако, очевидно, такой приём не является наилучшим с точки зрения построения эффективных кодов. Как будет показано далее, существуют гораздо более эффективные подходы к конструированию кодовых словарей, обладающих свойством однозначной декодируемости и, одновременно, существенно устраняющих избыточность источника.

Хотя любая конечная последовательность слов однозначно декодируемого словаря может быть декодирована лишь одним способом, отсюда не следует, что любая *текущая* последовательность слов обладает этим свойством; может понадобиться некоторая задержка, прежде чем окажется возможным её декодирование.

Рассмотрим, например, словарь

$$D = \{101,\ 00110,\ 10111,\ 11001\}.$$

Последовательность слов 10111, 00110, 11001 нельзя отличить от последовательности 101, 11001, 10110, 00110, пока не будет получен пятнадцатый символ. Исходя из этого, введём следующее понятие.

Число d_c символов в самой длинной последовательности слов, необходимое для однозначного декодирования самого первого слова этой последовательности, называется задержкой декодирования.

Заметим, что исходя из данного определения ещё нельзя однозначно утверждать, что в рассмотренном примере задержка декодирования равна 15, поскольку не осуществлён перебор всех возможных последовательностей кодовых слов.

Задержка декодирования не обязательно является конечной. Можно показать (см. далее), что словарь

$$D_1 = \{101, 00111, 10111, 11001\}$$

является однозначно декодируемым, тем не менее, последовательность

101110011100111…

может быть декодирована как

101, 11001, 11001, …

или как

10111, 00111, 00111, …

Длина самой длинной неоднозначной последовательности в этом случае неограничена. В дальнейшем изложении будут рассматриваться только имеющие практическое применение словари с конечной задержкой декодирования.

По-видимому, первым широко используемым неравномерным кодом, созданным ещё в XIX веке и употреблявшимся в течение всего XX века, был так называемый *двоичный код Морзе* (табл. 2.3), в котором 0 исторически называется *точкой* и обозначается знаком “·”, а 1 — *тире* и обозначается “–”.

Таблица 4.3

Двоичный код Морзе

A	B	C	D	E	F	G	H	I	J	K	L	M
01	1000	1010	100	0	0010	110	0000	00	0111	101	0100	11
N	O	P	Q	R	S	T	U	V	W	X	Y	Z
10	111	0110	1101	010	010	1	001	0001	011	1001	1011	1100

Как бы в силу иронии к вышесказанному, такой код, очевидно, не обладает свойством однозначной декодируемости, поскольку, например, последовательность 1100 может кодировать и букву *Z*, и последовательность букв *GE*. Поэтому для практического использования кода Морзе применялись дополнительные элементы — паузы между буквами и словами, что, конечно же, существенно снижало изначально заложенную эффективность[1].

[1] Небезынтересно, что для составления этого кода не использовались таблицы встречаемости букв (таких данных к тому времени, возможно, просто не существова-

Докажем необходимое условие однозначной декодируемости произвольного кодового словаря, известное в литературе как *неравенство Крафта*[1].

Если однозначно декодируемый m-символьный словарь содержит l слов, длины которых равны $n_1, \ldots, n_l$, то справедливо неравенство

$$\sum_{k=1}^{l} m^{-n_k} \le 1. \qquad (4.2.2)$$

Для доказательства неравенства Крафта рассмотрим конструкцию

$$\left(\sum_{k=1}^{l} m^{-n_k} \right)^L = \sum_{k_1=1}^{l} \cdots \sum_{k_L=1}^{l} m^{-\left(n_{k_1}+\ldots+n_{k_L}\right)},$$

где L — произвольное натуральное число. По отношению к кодовому словарю в этом выражении каждое слагаемое, стоящее в правой части, отражает определённую последовательность, состоящую из L кодовых слов, а сумма $n_{k_1}+\ldots+n_{k_L}$ есть суммарная длина такой последовательности. Обозначим через A_i число последовательностей, состоящих из L кодовых слов и имеющих суммарную длину i. Тогда можно записать:

$$\left(\sum_{k=1}^{l} m^{-n_k} \right)^L = \sum_{i=1}^{Ln_{\max}} A_i m^{-i}, \qquad (4.2.3)$$

где $n_{\max}$ — максимальное из чисел $n_1, \ldots, n_l$.

Заметим, что представленное выше соотношение справедливо для произвольных кодовых словарей, в том числе, не обязательно однозначно декодируемых. Применим теперь (4.2.3) к словарю, обладающему свойством однозначной декодируемости, так что в нем все i-элементные кодовые последовательности различны между собой. Максимальное число различных последовательностей длины i равно m^i. Разумеется, совсем не обязательно, что в словаре используются все такие последовательности, но в любом случае, для них справедлива оценка $A_i \le m^i$, откуда

$$\left(\sum_{k=1}^{l} m^{-n_k} \right)^L \le \sum_{i=1}^{Ln_{\max}} m^i m^{-i} = Ln_{\max},$$

ло), а относительная частота различных букв была оценена простым подсчётом литер в разных ячейках наборной кассы в одной из типографий.

[1] Его иногда называют *неравенством Сцилларда* или *неравенством Макмиллана*.

или

$$\sum_{k=1}^{l} m^{-n_k} \le \left(Ln_{\max}\right)^{1/L}.$$

Последнее неравенство справедливо при произвольном натуральном L, в том числе, и для произвольно большого значения. Поэтому, переходя в неравенстве к пределу при $L \to \infty$ и учтя при этом, что

$$\lim_{L\to\infty}\left(n_{\max}L\right)^{1/L} = \lim_{L\to\infty}\exp\left[\frac{1}{L}\ln\left(n_{\max}L\right)\right] = 1,$$

получим неравенство (4.2.2).

Применяя неравенство Крафта к коду Морзе (табл. 4.3), получим, что

$$\sum_{k=1}^{26} 2^{-n_k} = \frac{15}{4} > 1,$$

т. е. необходимое условие однозначной декодируемости не выполнено.

4.3. ПРЕФИКСНЫЙ КОД. КОДОВОЕ ДЕРЕВО ПРЕФИКСНОГО КОДА

Среди всех однозначно декодируемых кодов существует коды, обладающие свойством *мгновенной декодируемости*, когда каждое кодовое слово может быть однозначно декодировано сразу же после того, как оно полностью принято. Примером таких кодов являются *префиксные коды*.

Последовательность p, состоящая из m-ичных символов называется префиксом слова w, если это слово может быть представлено в виде $w = ps$, *где s — некоторая последовательность, содержащая не менее одного символа, называемая суффиксом слова w.*

Введённое понятие префикса позволяет определить кодовый словарь со свойством префикса как словарь, в котором никакое кодовое слово не является префиксом никакого другого кодового слова.

Наглядным графическим изображением префиксного кодового словаря является *кодовое дерево*, элементами которого являются *корень*, *ветви* и *узлы*. Узел дерева соответствует кодовому слову, так что *порядок узла* есть число кодовых символов. Ветви, идущие от корня дерева к узлам 1-го порядка, определяют выбор первого кодового символа, например, для двоичного кода — между нулём и единицей: левая ветвь соответствует 0, а правая 1. Те же самые правила применяются к узлам более высоких порядков,

и последовательное продвижение от корня дерева к *концевому узлу* определяет выбор конкретного кодового слова.

В качестве примера на рис. 4.3 изображено кодовое дерево, соответствующее двоичному словарю

$$D = \{00, 01, 100, 101, 1100, 1101, 1110, 1111\}.$$

Заметим, что наряду с концевыми узлами, кодовые слова формально могут быть приписаны также и *промежуточным узлам*. Например, двум промежуточным узлам второго порядка можно приписать кодовые слова 10 и 11 (на рисунке представлены в рамках), т. е. первые два символа кодовых слов, соответствующих концевым узлам, порождаемым этими промежуточными узлами. Однако такие кодовые слова не могут быть использованы для представления сообщений источника.

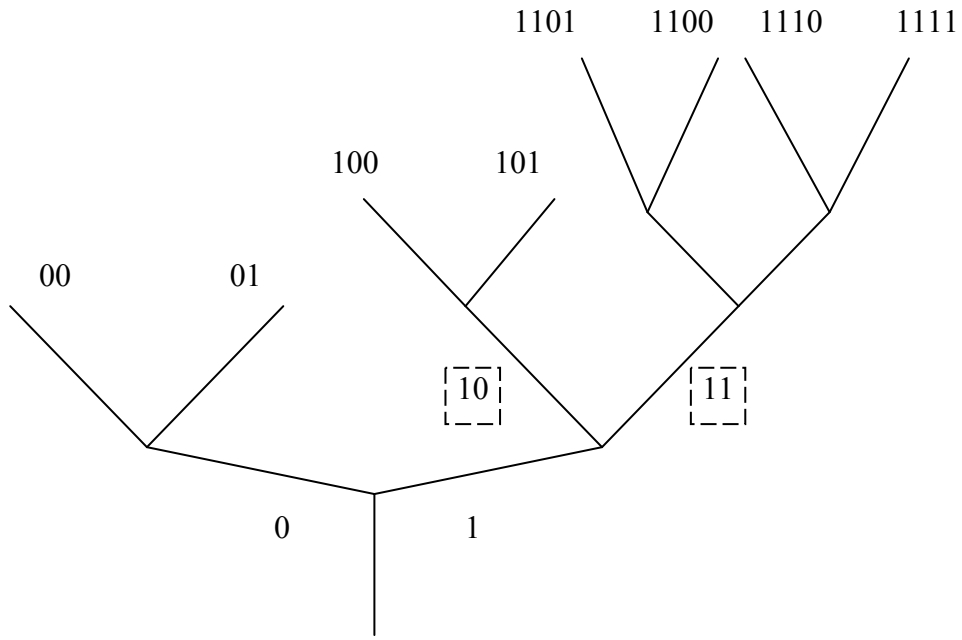


Рис. 4.3. Кодовое дерево двоичного кода

Рассмотрим, пока что в иллюстративных целях, некоторые примеры построения префиксных кодов. Как будет показано далее, эти коды не являются наилучшими по эффективности кодирования, в частности, они не являются оптимальными в смысле минимальной средней длины кодового слова. Однако снижение эффективности оправдывается сравнительной простотой процедуры построения кодовых слов. Кроме того, основные идеи, заложенные в рассматриваемых кодах, в последующем будут ис-

пользоваться для построения практически востребованных методов кодирования.

Код Шеннона

Пусть в произвольном дискретном источнике $X = \{x_k\}$ ($k = 1, \ldots, l$), элементы выбираются с соответствующими вероятностями $p_1, \ldots, p_l$. Расположим элементы источника в порядке возрастания (неубывания) их вероятностей:

$$p_1 \le p_2 \le \ldots \le p_l$$

и сопоставим каждому элементу так называемую *кумулятивную* вероятность $q_k = q(x_k)$, определяемую следующим образом:

$$q_1 = 0;$$

$$q_2 = p_1;$$

$$q_3 = p_1 + p_2;$$

$$\ldots.$$

$$q_l = \sum_{k=1}^{l-1} p_k .$$

Двоичным кодовым словом Шеннона w_k для k-го символа назовём двоичную последовательность, представляющую собой первые $\mu_k = \lceil -\log p_k \rceil$ разрядов после запятой в двоичной записи числа q_k. Здесь и далее символ $\lceil\ \rceil$ означает округление числа “вверх” (т. е. в сторону $+\infty$) до ближайшего целого, а логарифм, если это особо не оговорено, берётся по основанию 2.

Кратко напомним методику перевода десятичных дробей в двоичные дроби.

Если a_{10} — правильная десятичная дробь (нижний индекс в данном случае идентифицирует систему счисления), то её представление в двоичной системе соответствует разложению по отрицательным целым степеням двойки:

$$a_{10} = b_{-1}2^{-1} + b_{-2}2^{-2} + b_{-3}2^{-3} + \ldots,$$

где коэффициенты b_{-k} равны либо 0, либо 1. Тогда, умножив a_{10} на два и взяв целую часть от результата, получим значение старшего коэффициента b_{-1} в разложении:

$$2a_{10} = b_{-1} + b_{-2}2^{-1} + b_{-3}2^{-2} + b_{-4}2^{-2} + \ldots \to b_{-1}.$$

Таким же образом, проводя последовательное умножение на два, вычисляя и вычитая целую часть, можно найти остальные коэффициенты разложения. При этом получающаяся двоичная дробь может оказаться конечной, когда число a_{10} кратно отрицательной целой степени двойки, или бесконечной (в том числе, периодической), если условие кратности не выполняется. В последнем случае разложение следует проводить до тех пор, пока не будет достигнута заданная точность.

Если дробь a_{10} неправильная, то необходимо разделить целую α_{10} и дробную β_{10} части, после чего с каждой из частей отдельно проводится процедура перевода в двоичную форму. Методика перевода целой части аналогична рассмотренной выше с той лишь разницей, что разложение осуществляется по положительным целым степеням двойки:

$$б_{10} = b_0 2^0 + b_1 2^1 + b_2 2^2 + b_3 2^3 + \ldots$$

вследствие чего необходимо не умножение, а последовательное деление на два.

Пусть, к примеру, требуется перевести в двоичную форму число 0,75 — правильную дробь, кратную отрицательной целой степени двойки ($0{,}75 = 3/4 = 3/2^2$). Умножаем 0,75 на 2, получаем число 1,50, целая часть которого равна 1 — это значение b_{-1}. Далее, умножаем 0,50 на 2, получаем число 1,00, целая часть которого равна 1— это значение b_{-2}. Поскольку далее следуют нули, окончательно имеем конечную двоичную дробь:

$$0{,}75_{10} = 0{,}11_2.$$

Пусть теперь необходимо представить в двоичном виде число 0,65. Так как оно не является кратным отрицательной целой степени двойки, следует ожидать, что результирующая двоичная дробь будет бесконечной. Имеем

$$0{,}65 \times 2 = 1{,}30 = 1 + 0{,}30;\ b_{-1} = 1;$$
$$0{,}30 \times 2 = 0{,}60 = 0 + 0{,}60;\ b_{-2} = 0;$$
$$0{,}60 \times 2 = 1{,}20 = 1 + 0{,}20;\ b_{-3} = 1;$$
$$0{,}20 \times 2 = 0{,}40 = 0 + 0{,}40;\ b_{-4} = 0;$$
$$0{,}40 \times 2 = 0{,}80 = 0 + 0{,}80;\ b_{-5} = 0;$$
$$0{,}80 \times 2 = 1{,}60 = 1 + 0{,}60;\ b_{-6} = 1,$$

после чего процесс перевода повторяется, начиная со значения b_{-3}. Это означает, что результирующая двоичная дробь является периодической:

$$0{,}65_{10} = 0{,}10(1001)_2.$$

Рассмотрим построение двоичного кода Шеннона для 10-элементного источника, заданного в табл. 4.4, где символы уже представлены в упорядоченном виде и для определённости обозначены буквами русского алфавита. Энтропия такого источника, как нетрудно видеть, составляет $H(X) = 2{,}712$ бит, в то время как максимальная энтропия источника с $l = 10$ элементами равна $H_{\max} = \log_2 10 = 3{,}322$. Следовательно, избыточность r_x источника равна 0,184, и можно надеяться, что использование неравномерного кодирования приведёт к ощутимому выигрышу.

Таблица 4.4

Пример дискретного источника

x_k	$p(x_k)$	$\log_2 p(x_k)$	x_k	$p(x_k)$	$\log_2 p(x_k)$
А	0,300	– 1,737	Е	0,050	– 4,322

Продолжение табл. 4.4

Б	0,200	– 2,322	Ж	0,040	– 4,645
В	0,200	– 2,322	З	0,025	– 5,322
Г	0,100	– 3,322	И	0,020	– 5,645
Д	0,060	– 4,059	К	0,005	– 7,644

После того, как найдены кумулятивные вероятности $q(x_k)$ и соответствующие длины кодовых слов μ_k (табл. 4.5), нетрудно получить двоичное представление $[q_k]_2$ кумулятивных вероятностей, а затем — кодовые слова w_k. Поскольку максимальное значение μ_k равно семи, достаточно в двоичном представлении $[q_k]_2$ удерживать семь двоичных знаков после запятой.

Таблица 4.5

Построение двоичного кода Шеннона

k	x_k	p_k	q_k	μ_k	$[q_k]_2$	w_k
1	А	0,300	0,000	2	0,0000000	00
2	Б	0,200	0,300	3	0,0100011	010
3	В	0,200	0,500	3	0,1000000	100
4	Г	0,100	0,700	4	0,1011001	1011
5	Д	0,060	0,800	5	0,1100110	11001
6	Е	0,050	0,860	5	0,1101110	11011
7	Ж	0,040	0,910	5	0,1110100	11101
8	З	0,025	0,950	6	0,1111001	111100
9	И	0,020	0,975	6	0,1111100	111110
10	К	0,005	0,995	7	0,1111110	1111110

Как следует из табл. 4.5, средняя длина кодового слова для рассматриваемого источника составляет 3,255 символа.

Докажем однозначную декодируемость кода Шеннона. Рассмотрим произвольные символы источника x_i и x_j ($1 \le i < j \le l$). Поскольку для кумулятивных вероятностей этих символов выполняется соотношение $q_i < q_j$, можно утверждать, что кодовое слово w_i для символа x_i заведомо короче, чем кодовое слово w_j для символа x_j. Поэтому достаточно доказать, что общая часть этих кодовых слов всегда различна, т. е. что они всегда отличаются в одном из первых μ_i двоичных кодовых символов.

Для доказательства этого рассмотрим разность

$$q_j - q_i = \sum_{k=1}^{j-1} p_k - \sum_{k=1}^{i-1} p_k = \sum_{k=i}^{i-1} p_k \geq p_i . \qquad (4.3.1)$$

Поскольку длина кодового слова μ_k и его вероятность p_k связаны соотношением

$$\mu_k = \lceil -\log p_k \rceil \geq -\log p_k ,$$

справедливо неравенство

$$p_k \geq 2^{-\mu_k} . \qquad (4.3.2)$$

С учётом неравенств (4.3.1) и (4.3.2) можно записать, что

$$q_j - q_i \geq 2^{-\mu_k} .$$

Вспомним, что двоичное представление числа $2^{-\mu_k}$ являет собой стоящие после запятой $\mu_k - 1$ нулей и следом за ними — единицу на μ_k-й позиции. Тогда последнее неравенство означает ненулевую разность между числами q_i и q_j, а, следовательно, ненулевое различие между соответствующими кодовыми словами w_i и w_j, по меньшей мере, в одной из μ_k позиций. Иными словами, w_i не является префиксом w_j. Поскольку данное утверждение верно для произвольной пары i и j, это означает префиксность всего кодового словаря для кода Шеннона.

Дадим графическую интерпретацию кода Шеннона. Для этого рассмотрим числовой отрезок [0; 1), на котором будем последовательно, друг за другом, откладывать отрезки, длины которых равны вероятностям p_1, …, p_l появления символов x_1, …, x_l. Начальные точки таких отрезков, соответствующие кумулятивным вероятностям q_1, …, q_l, идентифицируют символы источника (рис. 4.4, *а*).

Предположим, что требуется закодировать, скажем, символ Г. Геометрически он находится в “интервале неопределённости” единичной длины.

На первом шаге кодер, генерируя либо 0, либо 1, указывает, в какой части интервала — левой, если 0 или правой, если 1 — находится соответствующая точка. В данном случае передаётся 1, после чего дальнейшая область возможного положения символа Г сужается вдвое (рис. 4.4, *б*) — в границы интервала [0,5; 1).

Далее, на втором шаге генерируется 0, что приводит к интервалу [0,5; 0,75) неопределённости положения символа, длина которого равна 0,25 (рис. 4.4, *в*).

Третий шаг — при генерировании 1 — состоит в разбиении интервала [0,5; 0,75) надвое и локализации символа Г в правом из получившихся интервалов (рис. 4.4, *г*) — в интервале [0,625; 0,750).

Наконец, на последнем шаге (рис. 4.4, *д*) генерируется символ 1, интервал [0,625; 0,750) разбивается надвое, и в получившейся правой части [0,6875; 0,7500) выбирается одна оставшаяся точка.

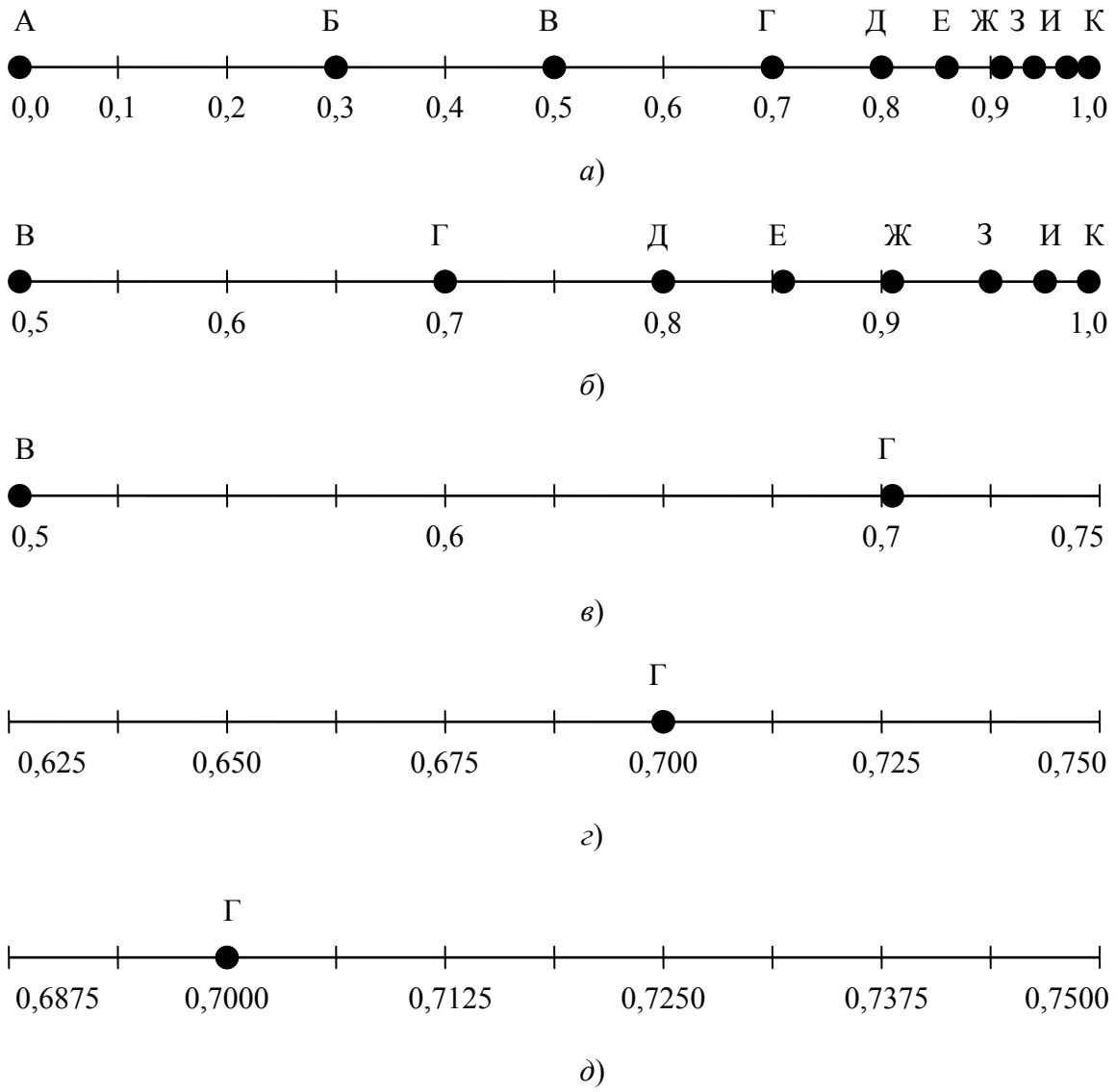


Рис. 4.4. Графическая иллюстрация кода Шеннона

Обратим внимание на тот факт, что 0,0625 — длина последнего интервала разбиения — меньше, чем длины прилегающих к точке символа Г отрезков: 0,1 и 0,2, что и гарантирует единственность оставшейся точки. В общем случае после генерирования μ_k двоичных символов длина интервала

неопределённости равна $2^{-\mu_k}$, и декодирование будет однозначным при выполнении неравенства

$$2^{-\mu_k} \le p_k$$

или, что эквивалентно,

$$\mu_k = -\log p_k .$$

Это условие в точности совпадает с правилом выбора длин кодовых слов в коде Шеннона.

Код Джильберта–Мура

Как нетрудно понять, первым и, пожалуй, наиболее трудоёмким этапом построения рассмотренного выше кода Шеннона является сортировка элементов источника по степени убывания их вероятностей. Попытаемся модифицировать алгоритм построения кодовых слов так, чтобы при этом не требовалась изначальная упорядоченность символов источника. Прямой подход, при котором кумулятивные вероятности формируются прямо из неупорядоченных вероятностей, очевидно, неэффективен, поскольку если перед символом x_k ($k = 1, \ldots, l$) с достаточно большой вероятностью p_k будет находиться символ x_{k-1}, вероятность которого p_{k-1} много меньше p_k, то длина кодового слова w_k неизбежно окажется неоправданно большой, ибо это слово является продолжением кодового слова w_{k-1}, длина которого велика из-за малости вероятности p_{k-1}.

Попытаемся нивелировать влияние символов с малыми вероятностями путём смещения значения кумулятивной вероятности в середины отрезков, длины которых отображают вероятности $p_1, \ldots, p_l$.

Пусть, по-прежнему, $q_1 = 0$, $q_k = \sum_{i=1}^{k-1} p_i$ ($k = 2, \ldots, l$) — кумулятивные вероятности, но значения p_k не обязательно расположены в порядке убывания. Определим величины

$$\pi_1 = \frac{p_1}{2};\ \pi_k = q_k + \frac{p_k}{2} = \sum_{i=1}^{k-1} p_i + \frac{p_k}{2}\ (k = 2, \ldots, l) \qquad (4.3.3)$$

и назовём *двоичным k-м кодовым словом Джильберта–Мура* последовательность, представляющую собой первые $\lambda_k = \left\lceil -\log\left(p_k / 2\right)\right\rceil$ разрядов в двоичном представлении величин π_k.

В табл. 4.6 представлены результаты построения кодовых слов Джильберта–Мура для того же источника, что и в табл. 4.4, но с неупоря-

доченными вероятностями. В последнем столбце табл. 4.6 показаны кодовые слова $w_k^{(Sh)}$ для кода Шеннона, соответствующие неупорядоченному источнику. Видно, что этот код является непрефиксным (является ли он однозначно декодируемым?).

Заметим, что минимальное значение λ_k равно девяти, поэтому точность в двоичном представлении величин π_k составляет девять знаков.

Таблица 4.6

Построение кодовых слов Джильберта–Мура

k	x_k	p_k	q_k	π_k	λ_k	$[q_k]_2$	$[\pi_k]_2$	w_k	$w_k^{(Sh)}$
1	Б	0,200	0,000	0,100	4	0,0000000	0,000110000	0001	000
2	Г	0,100	0,200	0,250	5	0,0011001	0,010000000	01000	0011
3	Е	0,050	0,300	0,325	6	0,0100011	0,010100011	010100	0100
4	З	0,025	0,350	0,363	7	0,0100110	0,010111001	0101110	010011
5	Ж	0,040	0,375	0,395	6	0,0110000	0,011001010	011001	01100
6	К	0,005	0,415	0,418	9	0,0110101	0,011010100	011010100	0110101
7	В	0,200	0,420	0,520	4	0,0110101	0,100001010	1000	011
8	А	0,300	0,620	0,770	3	0,1001111	0,110000101	110	10
9	И	0,020	0,920	0,910	7	0,1110101	0,111010001	1110100	111010
10	Д	0,060	0,940	0,970	6	0,1111000	0,111110000	111110	11111

Подсчёт средней длины кодового слова получившегося кода Джильберта–Мура даёт значение 4,260, что заметно больше, чем код Шеннона для того же источника.

Докажем однозначную декодируемость кода Джильберта–Мура. Для этого, как это делалось и для кода Шеннона, выберем произвольные символы источника x_i и x_j ($1 \le i < j \le l$), при этом, очевидно, $\pi_i < \pi_j$. Необходимо доказать, что кодовые слова w_i и w_j различаются хотя бы в одном из первых λ символов, где $\lambda = \min\{\lambda_i, \lambda_j\}$.

Для доказательства рассмотрим разность

$$\pi_j - \pi_i = \sum_{k=1}^{j-1} p_k - \sum_{k=1}^{i-1} p_k + \frac{p_j}{2} - \frac{p_i}{2} = \sum_{k=i}^{i-1} p_k + \frac{p_j - p_i}{2} \ge$$

$$\ge p_i + \frac{p_j - p_i}{2} = \frac{p_j + p_i}{2} \ge \frac{\max\{p_j; p_i\}}{2}. \qquad (4.3.4)$$

Поскольку длина кодового слова μ_k и его вероятность p_k связаны соотношением

$$\lambda_k = \left\lceil -\log\left(p_k / 2\right)\right\rceil \geq -\log\left(p_k / 2\right),$$

с учётом (4.3.4) справедливо неравенство

$$\pi_j - \pi_i \geq (1/2)\max\{p_j; p_i\} \geq 2^{-\min\{\lambda_j, \lambda_i\}},$$

означающее ненулевую разность между значениями π_i и π_j, а следовательно, ненулевое различие между соответствующими кодовыми словами w_i и w_j, по меньшей мере, в одной из $\lambda = \min\{\lambda_i, \lambda_j\}$ позиций. Поскольку данное утверждение верно для произвольной пары i и j, это означает префиксность всего кодового словаря для кода Джильберта–Мура.

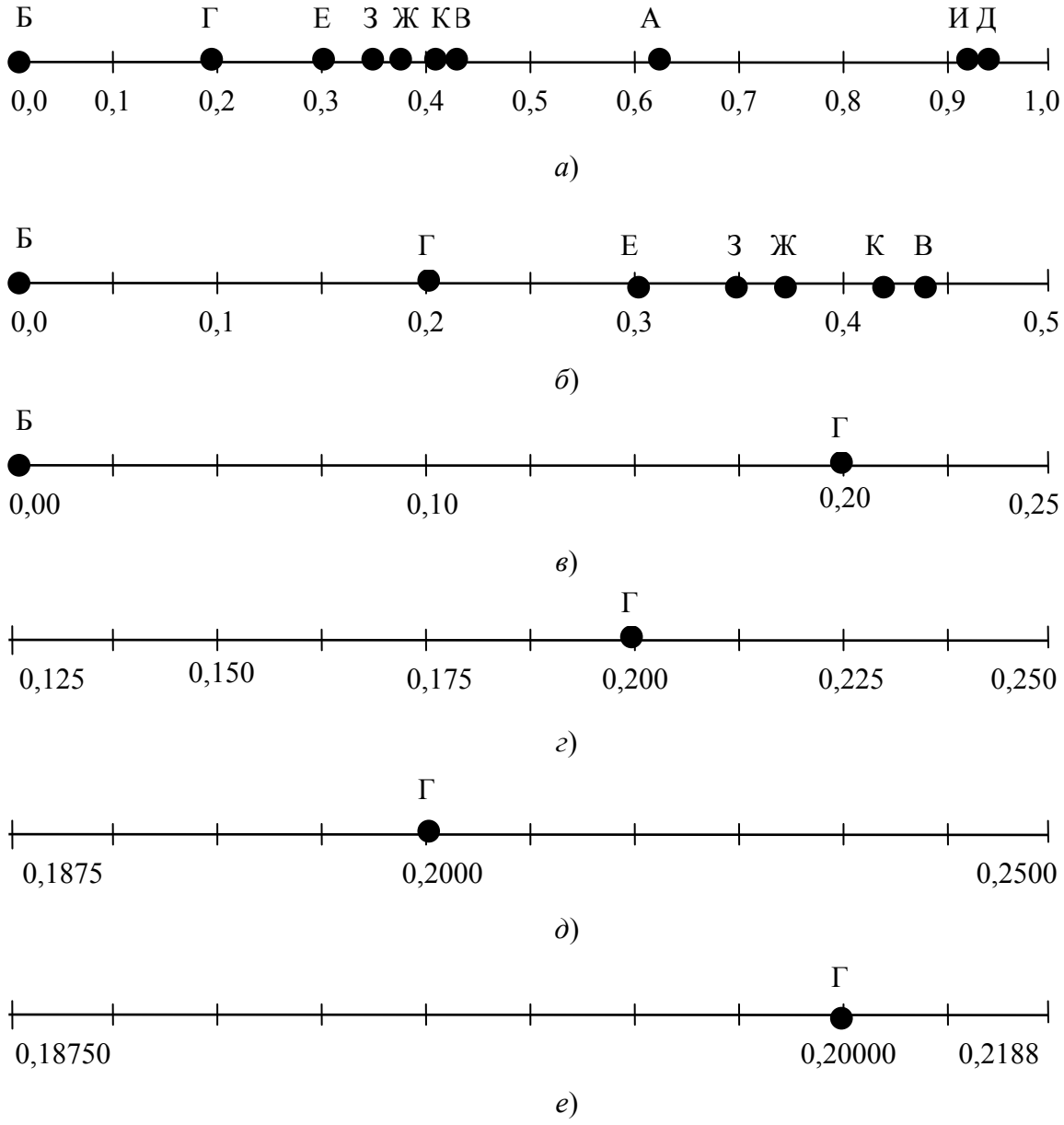


Рис. 4.5. Графическая иллюстрация кода Джильберта–Мура

На рис. 4.5 представлена графическая интерпретация кода Джильберта–Мура при построении кодового слова, соответствующего символу Г источника, заданного в табл. 4.4. По-прежнему, длины отрезков соответствуют вероятностям символов, однако кумулятивные вероятности теперь располагаются в серединах этих отрезков.

Выше было показано, что необходимым условием однозначной декодируемости кодовых словарей является неравенство Крафта. Особенностью префиксных кодов является то, что для них это неравенство является также и достаточным условием, т. е. образует полновесный критерий.

Неравенство (4.2.2) *является необходимым и достаточным условием существования однозначно декодируемого префиксного кодового словаря с l словами, состоящими из m-ичных символов, длины которых равны* $n_1, \ldots, n_l$. Докажем справедливость сформулированного критерия. При этом, несмотря на то, что необходимость условия справедлива как частный случай (неравенство Крафта справедливо для всех однозначно декодируемых словарей, в том числе, префиксных), целесообразно доказать его вновь в терминах кодового дерева, тем более, что доказательство является весьма простым и наглядным.

Итак, пусть существует кодовое дерево, концевые узлы которого имеют порядки $n_1, \ldots, n_l$. При этом из каждого узла исходит либо m ветвей, либо ни одной — в том случае, если этот узел концевой. Следовательно, у дерева существует не более m^l узлов, которые получается, когда ветви разветвляются из каждого узла меньших порядков.

С другой стороны, наличие каждого концевого узла порядка $n_k < n_{\max}$ исключает из кодового дерева m^{l-n_k} возможных узлов максимального порядка. Поэтому длины n_k кодовых слов, совпадающие с порядком концевых узлов, должны удовлетворять неравенству

$$\sum_{k=1}^{l} m^{l-n_k} \le m^l .$$

Деля это неравенство на m^l, получаем (4.2.2), и необходимость критерия доказана.

Доказательство достаточности неравенства Крафта для префиксного словаря состоит в том, что при его выполнении необходимо показать возможность построения кодового дерева с заданными концевыми узлами

(т. е. кодового словаря с заданными длинами кодовых слов). Это можно сделать индуктивно.

Расположим индексы k в порядке возрастания кодовых слов, так что $n_k \le n_{k+1}$ ($k = 1, \ldots, l-1$). Очевидно, всегда можно построить ветви, идущие от корня дерева. Предположим, что необходимо построить v_j концевых узлов порядка j при условии, что удалось построить дерево, содержащее все заданные концевые узлы всех меньших порядков. Число таких узлов равно, скажем i. По условию имеем:

$$1 \ge \sum_{k=1}^{l} m^{-n_k} = \sum_{k=1}^{i} m^{-n_k} + v_j m^{-j} + \sum_{k=i+v_j+1}^{l} m^{-n_k},$$

откуда

$$v_j = m^j - \sum_{k=1}^{i} m^{j-n_k} - \sum_{k=i+v_j+1}^{l} m^{j-n_k}.$$

С другой стороны, число μ_j доступных узлов j-го порядка, как уже было сказано, равно m^j минус число узлов, исключаемых из-за наличия концевых узлов меньшего порядка:

$$\text{м}_j = m^j - \sum_{k=1}^{j} m^{j-n_k}.$$

Таким образом,

$$v_j = \mu_j - \sum_{k=i+v_j+1}^{l} m^{j-n_k} \le \mu_j,$$

т. е. число доступных узлов j-го порядка не меньше заданного числа концевых узлов того же порядка, а поэтому все они могут быть включены в дерево. Поскольку установленное свойство справедливо для произвольного порядка j, то все кодовое дерево всегда может быть построено шаг за шагом, что и доказывает достаточность критерия.

Обратим внимание на то, что согласно доказанному критерию утверждается о существовании соответствующих кодовых словарей для *всех* наборов чисел $\{n_k\}$. Например, при $m = 2$ существует словарь $D_1 = \{1, 01, 001\}$, для которого $n_1 = 1$, $n_2 = 2$, $n_3 = 3$; словарь $D_2 = \{1, 01, 001, 0001\}$, для которого $n_1 = 1$, $n_2 = 2$, $n_3 = 3$, $n_4 = 4$; словарь $D_3 = \{1, 01, 001, 0000, 0001\}$, для которого $n_1 = 1$, $n_2 = 2$, $n_3 = 3$, $n_4 = n_5 = 4$ и т. п. Во всех случаях неравенство Крафта справедливо, следовательно, все такие словари являются

однозначно декодируемыми. Однако, при заданных длинах слов $\{n_k\}$ однозначно декодируемые словари не являются однозначно определёнными и могут быть реализованы многими способами.

Действительно, имеется $C_m^{n_1}$ способов выбрать первые n_1 символов, $C_{m(m-n_1)}^{n_2}$ способов выбрать следующие n_2 символов и т. п., так что общее количество кодовых словарей равно

$$M = C_m^{n_1} C_{m^2-mn_1}^{n_2} C_{m^3-m^2n_1-mn_2}^{n_3} \cdots C_{m^n-m^{n-1}n_1-mn_{n-1}}^{n},$$

где n — наибольшая из длин кодовых слов. Так например, для словаря D_3 существует восемь различных вариантов, представленных в табл. 4.7. Тем не менее, несмотря на такое разнообразие, все представленные варианты кодирования являются эквивалентными с точки зрения средней длины кодовых слов.

Таблица 4.7

Варианты кодовых слов, удовлетворяющих условию Крафта

Кодовое слово	1	2	3	4	5	6	7	8
w_1	1	1	1	1	0	0	0	0
w_2	01	01	00	00	10	10	11	11
w_3	001	000	011	010	110	111	100	101
w_4	0001	0010	0100	0110	1110	1101	1011	1001
w_5	0000	0011	0101	0111	1111	1100	1010	1000

Представленные в табл. 4.7 кодовые словари как будто бы дают возможность осуществить эффективное неравномерное кодирование заданного источника. Для этого надо лишь выбрать подходящий набор длин $\{n_k\}$, построить по ним кодовое дерево, а затем поставить в соответствие концевые узлы и элементы источника таким образом, чтобы менее вероятные элементы источника кодировались б*о*льшим числом кодовых символов. Однако при таком подходе возникает как минимум две существенные проблемы.

Первая из них состоит в отсутствии чёткого ограничения на максимальную длину кодовых слов, и при неудачном выборе набора $\{n_k\}$ можно не только не уменьшить среднюю длину кода $\overline{n}$, но даже увеличить ее.

Вторая проблема заключается в том, что вероятности появления элементов источника могут между собой отличаться незначительно (хотя их общий разброс может быть весьма большим), и для больших объёмов алфавита источника возникает ощутимая неоднозначность выбора различных вариантов соответствия элементов источника и концевых узлов дерева. При этом выбор того или иного варианта, вообще говоря, приводит к изменению значения $\overline{n}$.

Указанные проблемы приводят к необходимости формализации процесса выбора варианта соответствия между символами источника и кодовыми словами в виде какого-то алгоритма.

Конечно же, если распределение элементов источника имело бы вполне определённую регулярность, то обе описанные проблемы могли быть полностью или частично сняты.

Например, для источника $X = \{x_k\}$, у которого вероятности $p(x_k)$ равны целой отрицательной степени двойки (табл. 2.8), построение кода не вызывает сложностей. Выберем длины кодовых слов

$$n_1 = n_2 = 2;\ n_3 = n_4 = 3;\ n_5 = n_6 = n_7 = n_8 = 4,$$

после чего поставим в соответствие элементам источника следующие кодовые слова:

$$w_1 = 00,\ w_2 = 01,\ w_3 = 100,\ w_4 = 101,$$
$$w_5 = 1100,\ w_6 = 1101,\ w_7 = 1110,\ w_8 = 1111.$$

При таком способе кодирования средняя длина кода равна 16/8 = 1/2, что, как будет показано далее, является минимально возможным значением.

Таблица 4.8

Источник с вероятностями, равными целой отрицательной степени двойки

X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$p(x_k)$	1/4	1/4	1/8	1/8	1/16	1/16	1/16	1/16

Однако, если рассматривать реальные источники с произвольной статистикой сообщений то подобные простые рассуждения вряд ли приведут к желаемому результату, и, как уже было сказано, необходим некоторый алгоритм, формализующий метод кодирования и, соответственно, декодирования. Далее вначале будут рассмотрены некоторые эвристические алго-

ритмы, дающие хорошие, но не наилучшие результаты, а затем предложен оптимальный алгоритм префиксного кодирования — алгоритм Д.А. Хаффмана, позволяющий получить код с наименьшей средней длиной кодового слова.

4.4. ОПТИМАЛЬНЫЕ КОДЫ. АЛГОРИТМ ХАФФМАНА ПОСТРОЕНИЯ ОПТИМАЛЬНОГО КОДА

Исторически первым алгоритмом, позволяющим кодировать источники с произвольным распределением элементов, по-видимому, следует считать *алгоритм Шеннона–Фано*[1], формирующий префиксный код с достаточно высокой эффективностью. Опишем его в общем виде.

Алгоритм Шеннона–Фано

На первом шаге исходное множество сообщений $X = \{x_k\}$ ($k = 1, \ldots, l$) подразделяется на m подмножеств Q_{j_1} ($j_1 = 1, \ldots, m$) так, чтобы суммарные вероятности входящих в них сообщений были бы примерно одинаковы. При этом нумерация подмножеств осуществляется произвольным образом, а всем сообщениям, входящим в i-е подмножество, присваивается первый кодовый символ y_1 кодового алфавита $Y = \{y_i\}$ ($i = 1, \ldots, m$). Далее, каждое из подмножеств, полученных на первом шаге, рассматривается в качестве нового множества сообщений, и из них формируются подмножества Q_{j_1, j_2} ($j_2 = 1, \ldots, m$), для которых пошагово выполняются все те же описанные выше операции, до тех пор, пока в каждом подмножестве не окажется ровно по одному исходному сообщению.

Даже беглый анализ описанного алгоритма показывает, что он построен на эвристических идеях и не до конца формализован, в частности, отсутствует чёткое указание на то, каким образом формировать подмножества с “примерно одинаковыми” суммарными вероятностями.

Продолжим рассмотрение примеров алгоритмов кодирования на основе источника, заданного в табл. 4.4. Построим двоичный кодовый словарь, опираясь на алгоритм Шеннона–Фано.

На первом шаге образуем 2 подмножества $Q_1 = \{А, Б\}$ и $Q_2 = \{В, Г, Д, Е, Ж, З, И, К\}$, в которых суммарная вероятность составляет 0,5, и припи-

[1] Не путать с алгоритмом кодирования кодовыми словами Шеннона, рассмотренном в предыдущем разделе.

шем всем входящим в них элементам первые кодовые символы, например, 0 — первому подмножеству и 1 — второму.

Поскольку Q_1 содержит лишь 2 элемента, уже на втором шаге оно будет разделено на финальные подмножества $Q_{1,1} = \{А\}$ и $Q_{1,2} = \{Б\}$, содержащие по одному элементу. Дописывая в эти элементы вторые кодовые символы, получим следующие кодовые слова: $w_1 = w(А) = 00$ и $w_2 = w(Б) = 01$. Напротив, Q_2 содержит гораздо большее число элементов, поэтому разделим его на подмножества $Q_{2,1} = \{В, Е\}$ и $Q_{2,2} = \{Г, Д, Ж, З, И, К\}$ с суммарными вероятностями 0,25 и промежуточными кодовыми символами $w(Q_{2,1}) = 10$ и $w(Q_{2,2}) = 11$.

На третьем шаге из $Q_{2,1}$ будут получены финальные подмножества $Q_{2,1,1} = \{В\}$, $Q_{2,1,2} = \{Е\}$ и, следовательно, кодовые слова $w_3 = w(В) = 100$, $w_4 = w(Е) = 101$. $Q_{2,2}$ разделим на подмножества $Q_{2,2,1} = \{Г, И\}$ (промежуточное кодовое слово $w(Q_{2,2,1}) = 110$) и $Q_{2,2,2} = \{Д, Ж, З, К\}$ (промежуточное кодовое слово $w(Q_{2,2,2}) = 111$) с соответствующими вероятностями 0,12 и 0,13.

Таблица 4.9

Построение кода Шеннона–Фано

1	2	3	4	5
А	А — 00			
Б	Б — 01			
В	В	В — 100		
Г	Е	Е — 101		
Д	Г	Г	Г — 1100	
Е	Д	И	И — 1101	
Ж	Ж	Д	Д	Д — 11100
З	З	Ж	К	К — 11101
И	И	З	Ж	Ж — 11110
К	К	К	З	З — 11111

Четвёртый шаг позволяет получить финальные подмножества $Q_{2,2,1,1} = \{Г\}$ и $Q_{2,2,1,2} = \{И\}$, которым будут соответствовать кодовые слова $w_5 = w(Г) = 1100$, $w_6 = w(И) = 1101$, и очередное разбиение на подмножества $Q_{2,2,2,1} = \{Д, К\}$ (промежуточное кодовое слово

$w(Q_{2,2,2,1}) = 1110$) и $Q_{2,2,2,2} = \{$Ж, З$\}$ (промежуточное кодовое слово $w(Q_{2,2,2,2}) = 1111$) с суммарными вероятностями 0,065.

Наконец, на пятом шаге получим оставшиеся кодовые слова: $w_7 = w$(Д) = 11100, $w_8 = w$(К) = 11101, $w_9 = w$(Ж) = 11110, $w_9 = w$(З) = 11111. Весь процесс получения кодовых слов проиллюстрирован в табл. 4.9.

Средняя длина полученного кода равна $\overline{n} = 2{,}88$, что заметно ниже по сравнению с кодированием двоичным кодом Шеннона и, тем более, кодом Джильберта–Мура. Однако при этом следует заметить, что при других, менее удачных способах разбиения на подмножества эта величина может оказаться ощутимо выше (3 и более). С другой стороны, полученное значение $\overline{n}$, как будет показано далее, не является минимально возможным. Тем не менее, оно позволяет получить выигрыш от кодирования по сравнению с равномерным 4-разрядным кодом, необходимым для кодирования 10-элементного источника, примерно в 1,4 раза, что при большом количестве передаваемых символов уже может привести к существенной экономии.

Итак, при выполнении алгоритма Шеннона–Фано возникает кодовый словарь со свойством префикса, однако он не обязательно минимизирует значение $\overline{n}$ средней длины кодовых слов. Отсюда возникает задача о нахождении такой процедуры кодирования, которая при заданных вероятностях $p(x_k)$ появления исходных сообщений обеспечивала бы минимальную длину кодовых слов — такие коды называются *оптимальными*.

Заметим, что любой префиксный код является оптимальным, если вероятности исходных сообщений являются целыми отрицательными степенями числа m — объёма алфавита источника:

$$p(x_k) = m^{-d_k},\ d_k \in \mathbf{N},\ k = 1, 2, \ldots, m.$$

В частности, для таких источников к оптимальному префиксному коду также приводит и описанная процедура Шеннона — Фано. Однако, как уже было сказано, в общем случае сообщения источника имеют произвольные вероятности, поэтому предметом дальнейшего изучения будет являться поиск процедуры, приводящей к построению оптимального кода для произвольного источника. Такая процедура в классе префиксных кодов была предложена в 1952 г. Д. А. Хаффманом. При её рассмотрении ограничимся вначале двоичными кодами, а затем обобщим полученные результаты на произвольный кодовый алфавит.

Рассмотрим произвольный источник $X = \{x_k\}$ ($k = 1, \ldots, l$; $l \geq 2$) и пронумеруем сообщения в порядке невозрастания вероятностей их появления:

$$p(x_1) \geq p(x_2) \geq \ldots \geq p(x_l).$$

При этом, если для каких-то сообщений вероятности совпадают, то порядок их нумерации не имеет значения. Рассмотрим некоторые свойства, которым должен удовлетворять оптимальный префиксный двоичный код.

Свойство 1. *В оптимальном двоичном префиксном коде кодовое слово, соответствующее наименее вероятному сообщению, имеет наибольшую длину*.

Предположим обратное, т. е. пусть n_k ($k = 1, \ldots, l$) есть длина кодового слова, кодирующего сообщение x_k, и пусть для некоторого $k < l$ имеет место $n_k > n_l$. Построим новый кодовый словарь D', в котором поменяем местами сообщения x_l и x_k, т. е. x_k будет кодироваться n_l кодовыми символами, а x_l — кодироваться n_k кодовыми символами. Тогда в таком словаре средняя длина $\overline{n}'$ кодовых слов равна

$$\begin{aligned} \overline{n}' &= \overline{n} - p(x_k)n_k - p(x_l)n_l + p(x_k)n_l + p(x_l)n_k = \\ &= \overline{n} - \big(n_k - n_l\big)\big(p(x_k) - p(x_l)\big) < \overline{n}, \end{aligned} \qquad (4.4.1)$$

что противоречит предположению об оптимальности исходного кода.

Свойство 2. *В оптимальном двоичном префиксном коде два наименее вероятных сообщения кодируются словами одинаковой длины, различающимися лишь последним символом*.

Пусть слово w_k кодирует сообщение x_k, а имеющее наибольшую длину слово w_l — наименее вероятное сообщение x_l. Прежде всего, заметим, что если бы не существовало некоторого слова w_j ($j = 1, \ldots, l - 1$), имеющего одинаковую с w_l длину, то из слова w_l можно было бы выбросить крайний символ, не нарушая при этом свойства однозначной декодируемости и, одновременно, уменьшая среднюю длину $\overline{n}$. Таким образом, существуют два слова, отличающиеся в последнем символе. Покажем, что эти слова кодируют два наименее вероятных сообщения.

Как и в предыдущем случае, предположим обратное, т. е. что $j < l - 1$, тогда $n_j = n_l > n_{l-1}$, и опять среднюю длину $\overline{n}$ можно было бы уменьшить, поменяв местами слова w_j и w_{l-1}. Следовательно, такое предположение неверно, и наибольшую длину имеют именно слова w_l и w_{l-1}.

На основании установленных свойств можно утверждать, что когда в исходном алфавите сообщения $x_1, \ldots, x_{l-2}$ уже закодированы оптимальным образом, т. е. получена последовательность кодовых слов $w_1, \ldots, w_{l-2}$, оставшиеся кодовые слова w_{l-1} и w_l можно получить посредством добавления к слову w_{l-2} ещё двух кодовых символов.

Если бы алфавит X содержал не более трёх сообщений, то оптимальный код строится очень просто. Например, при $l = 3$ сообщению x_1 ставится в соответствие односимвольное кодовое слово $w_1 = 1$, а сообщения x_2 и x_3 кодируются двухсимвольными словами $w_2 = 01$ и $w_3 = 00$. Реально же количество исходных сообщений гораздо больше, и возникает идея формирования оптимального кода на основе пошаговой процедуры, при которой свойство оптимальности передаётся “в скрытой форме” посредством группирования сообщений на каждом шаге до тех пор, пока не получится алфавит, содержащий всего два сообщения. К такому алфавиту уже можно применить описанные выше действия.

Реализуя такую идею, построим из исходного алфавита X новый, *укрупнённый* алфавит X', содержащий $l - 1$ сообщений $x'_1, \ldots, x'_{l-1}$, вероятности которых равны

$$p(x'_k) = \begin{cases} p(x_k), & k = 1, \ldots, l-2; \\ p(x_{l-1}) + p(x_l), & k = l-1; \end{cases}$$

т. е. алфавит X' получается из алфавита X путём объединения двух наименее вероятных сообщений в одно укрупнённое сообщение с суммарной вероятностью. Очевидно, что префиксный код для X' можно превратить в соответствующий префиксный код для X путём добавления концевых кодовых символов: например, 1 для получения кодового слова w_{l-1} и 0 для получения w_l. Теперь, введя понятие укрупнённого алфавита, обоснуем, что оптимизация на каждом шаге приводит к построению всего оптимального кода.

Свойство 3. *Если двоичный префиксный код оптимален для укрупнённого алфавита X', то этот код также является оптимальным для исходного алфавита X.*

Учитывая, что длины $n'_1, \ldots, n'_{l-1}$ кодовых слов алфавита X' связаны с длинами $n_1, \ldots, n_{l-1}$ кодовых слов алфавита X соотношением

$$\begin{cases} n_k = n'_k, \quad k = 1, \ldots, l-2; \\ n_l = n_{l-1} = n'_{l-1} + 1, \end{cases}$$

представим значение $\overline{n}$ в следующем виде:

$$\overline{n} = \sum_{k=1}^{l-2} p(x_k) n_k + p(x_{l-1}) n_{l-1} + p(x_l) n_l =$$

$$= \sum_{k=1}^{l-1} p(x'_k) n'_k - n'_{l-1}\left[p(x_{l-1}) + p(x_l)\right] + p(x_{l-1}) n_{l-1} + p(x_l) n_l =$$

$$= \overline{n}' + p(x_{l-1})(n_{l-1} - n'_{l-1}) + p(x_l)(n_l - n'_{l-1}).$$

Учитывая, что $n_{l-1} - n'_{l-1} = 1$, а $p(x_{l-1}) + p(x_l) = p(x'_{l-1})$, получаем

$$\overline{n} = \overline{n}' + p(x'_{l-1}),$$

т. е. значения $\overline{n}$ и $\overline{n}'$ отличаются на независящую от выбора кодовых слов величину $p(x'_{l-1})$. Таким образом, при формировании кода для алфавита X' с минимальным значением $\overline{n}'$ такой же код формируется и для алфавита для X.

Распространим теперь полученные результаты на недвоичные коды. Поскольку соотношение (4.4.1) не зависит от величины m, можно утверждать, что первое свойство остаётся справедливым для произвольного объёма кодового алфавита. Такое же утверждение, вообще говоря, нельзя отнести ко второму и третьему свойствам, поскольку при определённых соотношениях между l и m могут возникать *неполные деревья*, т. е. деревья, у которых имеются пустые концевые узлы.

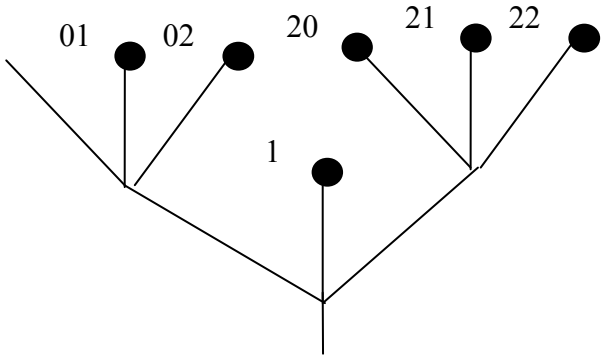


Рис. 4.6. Неполное кодовое дерево

Например, при кодировании шестиэлементного источника троичным алфавитом наилучшим вариантом представляется тот, что изображён на

рис. 4.6, где имеется пустой концевой узел второго порядка[1].

Наличие пустых концевых узлов оставляет открытым вопрос о количестве кодовых слов, отличающихся в последнем символе, что, в свою очередь, делает неопределённым процесс укрупнения. Проблема становится более ясной после введения понятия *полного словаря*.

Словарь, для которого неравенство Крафта превращается в равенство, называется полным словарём.

Полное кодовое дерево, т. е. дерево, соответствующее полному словарю, не может быть дополнено никакими узлами — в таком дереве нет "пустых" ветвей, подобных верхней левой ветви на рис. 4.6. Очевидно, что полные кодовые деревья могут быть построены не для произвольных значений m объёма кодового алфавита. Докажем справедливость следующего утверждения.

Равенство

$$l = \alpha(m - 1) + 1, \qquad (4.4.2)$$

где α *— произвольное натуральное число, является необходимым и достаточным условием существование полного словаря с множеством из* l *концевых узлов.*

Необходимость сформулированного условия можно доказать подсчётом концевых узлов полного кодового дерева, в котором m ветвей разветвляются из каждого промежуточного узла. Будем строить дерево последовательными шагами, обозначив при этом через M_i число имеющихся в дереве свободных узлов, когда дерево построено вплоть до i-го порядка, а через l_i — число промежуточных узлов i-го порядка, когда дерево построено до узлов порядка $j > i$. Очевидно число M_i становится равным l_i после построения узлов очередного j-го порядка. Тогда число свободных узлов первого порядка равно

$$M_1 = m \equiv M_0 + (m - 1),$$

где $M_0 = 1$.

Далее каждый из l_1 промежуточных узлов первого порядка порождает m узлов второго порядка. Следовательно, общее число свободных узлов

[1] В принципе, можно было бы заполнить все узлы второго порядка, сделав пустым второй промежуточный узел, однако при этом увеличивается на единицу длина соответствующего кодового слова и, следовательно, возрастает средняя длина кода.

после построения дерева вплоть до узлов второго порядка равно

$$M_2 = l_1 m = l_1 + l_1(m-1) = M_1 + l_1(m-1);$$

вплоть до узлов третьего порядка:

$$M_3 = l_2 m = l_2 + l_2(m-1) = M_2 + l_2(m-1),$$

и т. д.

В общем случае имеем соотношение

$$M_{i+1} = M_i + l_i(m-1),$$

которое запишем в виде

$$M_{i+1} = M_i + l_i(m-1) = M_{i-1} + l_{i-1}(m-1) + l_i(m-1) = \ldots =$$

$$= M_0 + (m-1) + l_1(m-1) + \ldots + l_{n_N-1}(m-1) = 1 + (m-1)\left[1 + \sum_{k=1}^{n_N-1} l_k\right],$$

где n_N — наивысший порядок концевого узла из заданного множества. Поскольку выражение в квадратных скобках есть сумма натуральных чисел, то равенство (4.4.2) удовлетворяется и необходимость доказана.

Для доказательства достаточности условия (4.4.2) надо показать, что для любого числа N, удовлетворяющего такому условию, существует дерево, имеющее ровно N узлов. В свою очередь это равносильно существованию набора натуральных чисел $\{l_i\}$ ($i < n_N$), для которых

$$\left[1 + \sum_{k=1}^{n_N-1} l_k\right] = \alpha, \alpha \in \mathbb{N}.$$

При этом единственным требованием, налагаемым на них, является условие

$$1 \le l_i \le m l_{i-1},\ i < n_N.$$

Так как на n_N не накладывается никаких ограничений, всегда можно выбрать l_i, чтобы удовлетворить поставленным условиям, что и доказывает достаточность.

Установленное свойство, связывающее значения l и m в полном дереве, позволяет обобщить процедуру оптимального кодирования следующим образом. Будем трактовать произвольное кодовое дерево как полное, в котором, возможно, для выполнения свойства полноты добавлено некоторое число неиспользуемых концевых узлов. Конкретно, необходимо добавить столько неиспользуемых узлов, чтобы выполнялось (4.4.2), т. е. (l – 1) было кратно (m – 1). Ясно, что в оптимальном коде все неиспользуемые кон-

цевые узлы имеют такую же длину, что самое длинное кодовое слово, и можно условно считать, что все m наименее вероятных кодовых слов имеют одинаковую длину. Тем самым обобщается свойство 2, а вслед за этим начинать процесс укрупнения, получая на основании обобщённого свойства 3 оптимальный неравномерный код.

Теперь можно сформулировать в общем виде процедуру построения оптимального неравномерного кода.

Алгоритм Хаффмана

Все буквы исходного алфавита в количестве l выписываются в порядке убывающей вероятности. Если число $l - 1$ не кратно $m - 1$, то в алфавит добавляются дополнительные "буквы", которым приписываются вероятности, равные нулю, так чтобы для объёма полученного алфавита l' соблюдалось условие кратности $l' - 1$ числу $m - 1$. Затем последние m элементов полученного алфавита объединяются в "укрупнённый" элемент, вычисляется его вероятность, которая записывается в соответствующее место алфавита. То же самое проделывается с последними m элементами полученного алфавита (включая укрупнённые), и это продолжается до тех пор, пока не останется алфавит, состоящий из m элементов. Этим элементам приписываются в любом порядке одноразрядные кодовые комбинации 0, 1, …, $m - 1$. Если в числе оставшихся m элементов есть такие, которые принадлежат исходному алфавиту (т. е. не полученные в результате объединения других элементов), то они оказываются закодированными одноразрядными кодовыми комбинациями. Для тех элементов, которые получены в результате объединения m букв исходного алфавита в кодовые комбинации дописываются вторые символы 0, 1, …, $m - 1$, так что эти знаки оказываются закодированными двухразрядными комбинациями. Если среди этих двухразрядных комбинаций имеются такие, которые соответствуют элементам, полученным в результате объединения, то к этим комбинациям приписываются третьи символы и т. д., пока все элементы исходного алфавита не окажутся закодированными.

Построим двоичный код Хаффмана для источника, заданного в табл. 4.4. Процесс построения иллюстрируется на рис. 4.7.

На первом шаге (номера шагов отображаются в верхних строках) объединяются два наименее вероятных символа И и К (соответствующие символы заключены в рамку), образуя укрупнённый символ ИК, имеющий

суммарную вероятность 0,025 (укрупнённые символы, получаемые на каждом шаге, выделены жирным шрифтом). Суммарная вероятность укрупнённого символа оказалась равна вероятности символа З, и, в принципе, символ ИК можно было бы поместить перед символом З. Условимся для определённости считать, что укрупнённый символ помещается в конец списка символов, имеющих одинаковые вероятности, так что в рассматриваемом примере после первого шага символ ИК по-прежнему остаётся на последнем месте.

На втором шаге производится объединение исходного символа источника З и укрупнённого символа ИК с образованием укрупнённого символа ЗИК, имеющего суммарную вероятность 0,050, поэтому он помещается между символами Е и Ж.

1	2	3	4	5
А 0,300	А 0,300	А 0,300	А 0,300	А 0,300
Б 0,200	Б 0,200	Б 0,200	Б 0,200	Б 0,200
В 0,200	В 0,200	В 0,200	В 0,200	В 0,200
Г 0,100	Г 0,100	Г 0,100	Г 0,100	**ДЕ 0,110**
Д 0,060	Д 0,060	Д 0,060	**ЗИКЖ 0,090**	Г 0,100
Е 0,050	Е 0,050	Е 0,050	Д 0,060	ЗИКЖ 0,090
Ж 0,040	Ж 0,040	**ЗИК 0,050**	Е 0,050	
З 0,025	З 0,025	Ж 0,040		
И 0,020	**ИК 0,025**			
К 0,005				

6	7	8	9
А 0,300	А 0,300	**БВ 0,400**	**АГЗИКЖДЕ 0,600**
Б 0,200	**ГЗИКЖДЕ 0,300**	А 0,300	БВ 0,400
В 0,200	Б 0,200	ГЗИКЖДЕ 0,300	
ГЗИКЖ 0,190	В 0,200		
ДЕ 0,110			

Рис.4.7. Процесс построения двоичного кода Хаффмана

Проводя аналогичные действия по укрупнению элементов на протяжении девяти шагов, получим в конечном итоге укрупнённые символы АГЗИКЖДЕ и БВ, которым, согласно алгоритму, необходимо присвоить (в произвольном порядке) кодовые символы 1 и 0. Снова для определённости будем присваивать верхним (по расположению в укрупняемом ансамбле) символам 1, а нижним — 0. Таким образом, имеем

АГЗИКЖДЕ — 1; БВ — 0.

Далее начинается процесс разукрупнения (девять "обратных" шагов) и одновременного получения кодовых слов. Укрупнённый символ БВ распадается на исходные символы Б и В, которым присваиваются искомые кодовые слова

Б — 01; В — 00.

Укрупнённый символ АГЗИКЖДЕ распадается на исходный символ А и укрупнённый символ ГЗИКЖДЕ, так что получается одно искомое и одно "промежуточное" кодовое слово:

А — 11; ГЗИКЖДЕ –10.

Продолжая процесс разукрупнения, окончательно получаем следующие кодовые комбинации:

А — 11; В — 01; В — 00; Г — 1011; Д — 1001; Е — 1000; Ж — 10100;

З — 101011; И — 1010101; К — 1010100.

При этом средняя длина кодового слова составляет $\bar{n} = 2{,}825$, что, конечно, близко к аналогичному значению для алгоритма Шеннона–Фано, но все же меньше его.

Как уже говорилось, количественной характеристикой эффективности сжатия неравномерного кода является выигрыш кодирования $\gamma = a / \bar{n}$, показывающий отличие средней длины кодового слова $\bar{n}$ от количества разрядов, *a*, необходимых для того, чтобы закодировать данный источник равномерным кодом. Как нетрудно подсчитать, для рассматриваемого примера выигрыш кодирования составляет $\gamma = 1{,}416$.

Необходимо, однако, заметить, что γ, вообще говоря, не является "чистым" выигрышем кодирования, поскольку наряду с передаваемой закодированной информацией необходимо также передавать правила соответствия между символами источника и кодовыми словами, что определяется методом практической реализации того или иного кода. В некоторых случаях — когда статистика источника полагается неизменной — правила со-

ответствия фиксированы и известны как на передающей, так и на приёмной стороне. Тогда необходимость в передаче таблицы соответствия отпадает, и γ становится “чистым” выигрышем кодирования. Однако зачастую при работе системы передачи информации допускается значимое изменение статистики сообщений, что, в свою очередь, предполагает возможное изменение и кодовых слов и их соответствия. Такие системы называются адаптивными, следовательно, адаптивными должны быть и методы кодирования. Впрочем, если объёмы передаваемой “полезной” информации достаточно велики, то вклад “служебной” части в виде таблицы соответствия, как правило, мал.

В качестве примера хаффмановского кодирования с фиксированными кодовыми словами в табл. 4.9 представлен двоичный код Хаффмана для русского языка ($l = 32$), составленный, исходя из усреднённых значений частоты встречаемости.

Таблица 4.9

Двоичный код Хаффмана для русского алфавита

x_k	$p(x_k)$	w_k	x_k	$p(x_k)$	w_k
–	0,175	000	Я	0,018	110110
О	0,090	001	Ы	0,016	110111
Е	0,072	0100	З	0,016	111000
А	0,062	0101	Ъ, Ь	0,014	111001
И	0,062	0110	Б	0,014	111010
Т	0,053	0111	Г	0,013	111011
Н	0,053	1000	Ч	0,012	111100
С	0,045	10010	Й	0,010	1111110
Р	0,040	10011	Х	0,009	1111011
В	0,038	1010	Ж	0,007	1111100
Л	0,035	10110	Ю	0,006	1111101
К	0,028	10111	Ш	0,006	11111100
М	0,026	11000	Ц	0,004	11111101
Д	0,025	110010	Щ	0,003	11111110
П	0,023	110011	Э	0,002	111111110
У	0,021	11010	Ф	0,002	111111111

Нетрудно сосчитать, что средняя длина кодового слова составляет $\bar{n} = 4{,}4$ в то время как при равномерном кодировании необходимо использовать пять двоичных разрядов, т. е. выигрыш кодирования составляет 1,136.

Примером кода Хаффмана с возможностью адаптации является кодирование изображений при *факсимильной передаче*.

Факсимильная передача[1] — это метод передачи плоского (двумерного) образа посредством последовательных строчных развёрток. Чаще всего передаче подлежат документы, содержащие буквенно-цифровые последовательности; факсимильная передача графических (тем более, полутоновых) изображений из-за невысокой разрешающей способности нецелесообразна. Кратко рассмотрим основные принципы передачи факсимильных изображений.

Передаваемое изображение, стандартные размеры которого составляют 20,7×29,2 см^2, квантуется в двумерной координатной сетке как последовательность черных (Ч) и белых (Б) элементов — *пикселов*. При этом каждая строка разбивается на 1728 пикселов, а весь документ — на 1188 строк для стандартного разрешения и 2376 строк (в два раза больше) для высокого разрешения (в некоторых факсимильных аппаратах возможен ещё и режим сверхвысокого разрешения с учетверённой — по отношению к стандартной — чёткостью). Таким образом, общее число пикселей составляет 2 052 864 для нормального разрешения и 4 105 728 для высокого разрешения.

Для стандартного документа горизонтальная развёртка страницы обычно представляет собой последовательность, состоящую из достаточно длинных групп чёрных и белых пикселей. Каждая такая группа описывается так называемыми *кодовыми словами деления*, которые определяют двухуровневое разделение групп пикселов.

Первое разделение — это деление групп на отрезки, кратные 64 элементам, т. е. длины отрезков d_k равны $d_k = 64k$, $k = 1, 2, \ldots, 27$. Отрезки белых и черных пикселов кодируются хаффмановским кодом, которые называются *созданными кодовыми словами*. При этом кодовые слова формиру-

[1] Слово “факсимиле” происходит от двух латинских слов: *facere* (делать) *similes* (похожий)

ются на основе фиксированной статистики, собранной посредством типичных документов, пересылаемых посредством факсимильных аппаратов. Конкретно, рассчитывалось количество черных и белых пикселов, в восьми эталонных документах, которые, как считается, содержат типичные пересылаемые тексты и изображения. Описание таких текстов представлено в табл. 2.10 (опубликование самих текстов и изображений невозможно, поскольку на них распространяется авторское право Международного телекоммуникационного союза).

Второе деление определяет количество s однотипных пикселов, оставшихся в группе после того, как в ней уже отсчитано максимально возможное из всех d_k элементов, т. е. s — это разность между общим количеством Q пикселов в группе и максимальным числом $64k$ ($k = 1, 2, \ldots, 27$), не превосходящим значение Q. Понятно, что s принимает значения от 0 до 63, и хаффмановское кодовое слово, соответствующее своему значению s, называется *оконечным кодовым словом*.

Таблица 4.10

Тестовые документы для факсимильной связи

№	Описание документа
1	Типичное деловое письмо на английском языке
2	Рисунок от руки электрической цепи
3	Печатная форма на французском языке, заполненная на машинке
4	Плотно напечатанный отчёт на французском языке
5	Техническая статья с иллюстрациями на французском языке
6	График с напечатанными подписями на французском языке
7	Плотный документ
8	Рукописная записка белым по чёрному на английском языке

Если количество пикселов в группе меньше 64, то первое деление отсутствует, и кодирование осуществляется непосредственно по оконечному кодовому слову.

Кроме того, определён специальный *маркер конца строки* EOL (end of line), появление которого означает невозможность дальнейшего появления черных пикселов в строке, и, следовательно, за ним должна начаться следующая строка.

В табл. 4.11 представлены созданные кодовые слова *W*(Б) и *W*(Ч), соответствующие длинам d_k отрезков белых и черных пикселов, а в табл. 4.12 — оконечные кодовые слова *w*(Б) и *w*(Ч), для соответствующих значений *s* внутри групп.

Таблица 4.11

Созданные кодовые слова

d_k	*W*(Б)	*W* (Ч)	d_k	*W* (Б)	*W* (Ч)
64	11011	0000001111	960	011010100	0000001110011
128	10010	000011001000	1024	011010101	0000001110100
192	010111	000011001001	1088	011010110	0000001110101
256	0110111	000001011011	1152	011010111	0000001110110
320	00110110	000000110011	1216	011011000	0000001110111
384	00110111	000000110100	1280	011011001	0000001010010
448	01100100	000000110101	1344	011011010	0000001010011
512	01100101	0000001101100	1408	011011011	0000001010100
576	01101000	0000001101101	1472	010011000	0000001010101
640	01100111	0000001001010	1536	010011001	0000001011010
704	011001100	0000001001011	1600	010011010	0000001011011
768	011001101	0000001001100	1664	011000	0000001100100
832	011010010	0000001001101	1728	010011011	0000001100101
896	011010011	0000001110010	EOL	000000000001	000000000001

При анализе указанных в табл. 4.10 документов было обнаружено, что чаще всего встречаются серии из двух, трёх и четырёх черных пикселов, поэтому им были присвоены самые короткие коды. Далее по частоте встречаемости идут серии из 2–7 белых пикселов, так что им были присвоены несколько более длинные коды. Большинство остальных серий встречаются гораздо реже, и им были назначены достаточно длинные 12-разрядные кодовые комбинации.

Таблица 4.12

Оконечные кодовые слова

s	*w*(Б)	*w*(Ч)	*s*	*w*(Б)	*w*(Ч)
0	00110101	000110111	32	00011011	000001101010
1	000111	010	33	00010010	000001101011
2	0111	11	34	00010011	000011010010
3	1000	10	35	00010100	000011010011
4	1011	011	36	00010101	000011010100

Окончание таблицы 4.12

5	1100	0011	37	00010110	000011010101
6	1110	0010	38	00010111	000011010110
7	1111	00011	39	00101000	000011010111
8	10011	000101	40	00101001	000001101100
9	10100	000100	41	00101010	000001101101
10	00111	0000100	42	00101011	000011011010
11	01000	0000101	43	00101100	000011011011
12	001000	0000111	44	00101101	000001010100
13	000011	00000100	45	00000100	000001010101
14	110100	00000111	46	00000101	000001010110
15	110101	000011000	47	00001010	000001010111
16	101010	0000010111	48	00001011	000001100100
17	101011	0000011000	49	01010010	000001100101
18	0100111	0000001000	50	01010011	000001010010
19	0001100	00001100111	51	01010100	000001010011
20	0001000	00001101000	52	01010101	000001000100
21	0010111	00001101100	53	00100100	000000110111
22	0000011	00000110111	54	00100101	000000111000
23	0000100	00000101000	55	01011000	000000100111
24	0101000	00000010111	56	01011001	000000101000
25	0101011	00000011000	57	01011010	00000101100
26	0010011	000011001010	58	01011011	000001011001
27	0100100	000011001011	59	01001010	000000101011
28	0011000	000011001100	60	01001011	000000101100
29	00000010	000011001101	61	00110010	000001011010
30	0000011	000001101000	62	00110011	000001100110
31	00011010	000001101001	63	00110100	000001100111

Пусть, например, квантованная строка изображения со стандартным разрешением имеет следующий вид (группы пикселов для удобства разделены запятыми):

219 Б, 14 Ч, 15 Б, 14 Ч, 97 Б, 114 Ч, 1255 Б.

Чтобы закодировать первую группу белых пикселов необходимо выбрать отрезок длиной $d_3 = 192$ и созданное кодовое слово $W(\text{Б}) = 010111$, после чего в группе останется ещё $s = 27$ пикселов, что соответствует оконечному кодовому слову $w(\text{Б}) = 0100100$. Рассуждая аналогично, для последующих групп имеем:

14 Ч: $s = 14$, w(Ч) = 00000111;

15 Б: $s = 15$, w(Б) = 110101;

14 Ч: $s = 14$, w(Ч) = 00000111;

97 Б: $d_1 = 64$, W(Б) = 11011, $s = 33$, w(Б) = 00010010;

114 Ч: $d_1 = 64$, W(Ч) = 0000001111, $s = 50$, w(Ч) = 000001010010;

1255 Б: $d_{19} = 1216$, W(Б) = 011011000, $s = 39$; w(Б) = 00101000.

Таким образом, последовательность кодовых слов, соответствующая рассматриваемой строке (содержащей 1728 пикселов), имеет вид

0101110100100000001111101010000011100000011110000 0…

10100100110110000010100 0.

Отношение 1728/73 = 23,671 является оценкой выигрыша кодирования.

4.5. ПОТЕНЦИАЛЬНЫЕ ВОЗМОЖНОСТИ НЕРАВНОМЕРНОГО КОДИРОВАНИЯ

Алгоритм Хаффмана приводит к построению префиксного кода с наименьшим значением средней длины кодовых слов. Однако, несмотря на то, что такой код является наилучшим в классе префиксных кодов, все же остаётся открытым вопрос о потенциальных возможностях кодирования для произвольных неравномерных кодов, в частности, каково при этом минимально возможное значение средней длины кода. Ответ на этот вопрос содержится в утверждении, известном как *теорема Шеннона для идеального дискретного канала*.

Пусть $H(X)$ — энтропия первого приближения для источника X, т. е. когда его элементы x_k ($k = 1, \ldots, l$) выбираются независимым образом. Тогда существует способ кодирования элементов x_k символами m-ичного алфавита, при котором средняя длина кода $\overline{n}$ удовлетворяет неравенству

$$\frac{H(X)}{\log m} \le \overline{n} < \frac{H(X)}{\log m} + 1. \qquad (4.5.1)$$

Для получения нижней границы в (2.5.1) рассмотрим разность

$$H(X) - \overline{n}\log m = -\sum_{k=1}^{l} p(x_k)\log p(x_k) + \sum_{k=1}^{l} p(x_k)\log m^{-n_k} =$$

$$= \sum_{k=1}^{l} p(x_k)\log\frac{m^{-n_k}}{p(x_k)} \le \log e\sum_{k=1}^{l} p(x_k)\left[\frac{m^{-n_k}}{p(x_k)} - 1\right] = \log e\left(\sum_{k=1}^{l} m^{-n_k} - 1\right),$$

где использовано неравенство $\ln z \le z - 1$. Далее учтём, что для произвольного однозначно декодируемого кода необходимо выполнение неравенства Крафта, так что

$$H(X) - \overline{n} \log m \le 0,$$

откуда следует искомая нижняя граница. Заметим, что знак равенства имеет место тогда и только тогда, когда $p(x_k) = m^{-n_k}$, т. е. когда вероятности сообщений равны целой неотрицательной степени объёма кодового алфавита. На этот факт было указано ещё в предыдущем разделе.

Получим теперь верхнюю границу в (4.5.1). Для этого просто покажем, каким образом следует выбрать длины кодовых слов $\{n_k\}$ ($k = 1, \ldots, l$), чтобы они удовлетворяли поставленному условию. Разумеется, невозможно произвольным образом подобрать набор значений n_k, удовлетворяющих неравенству

$$\sum_{k=1}^{l} n_k p(x_k) < \frac{H(X)}{\log m} + 1,$$

поскольку они должны быть натуральными числами, поэтому поступим следующим образом. Выберем значения n_k так, чтобы удовлетворить системе двойных неравенств

$$m^{-n_k} \le p(x_k) < m^{-n_k+1}, \; k = 1, \ldots, l,$$

после чего просуммируем левое неравенство по всем k. Результирующее неравенство является неравенством Крафта, что гарантирует существование префиксного кода с выбранными длинами.

С другой стороны, логарифмируя правые неравенства, имеем

$$n_k < \frac{-\log p(x_k)}{\log m} + 1.$$

Умножая эти неравенство на $p(x_k)$ и суммируя по всем k, получаем требуемую верхнюю границу.

Итак, одним из способов снижения значения $\overline{n}$ для заданного источника является использование многопозиционных кодов. К сожалению, в большинстве практических случаев переход от двоичных к многопозиционным кодам нереализуем, так что получаемый при этом выигрыш представляет в значительной степени академический интерес. Возникает вопрос, нельзя ли уменьшить $\overline{n}$, оставаясь в рамках двоичных кодов?

Оказывается, можно получить более сильные результаты, если кодовые слова приписывать не отдельным элементам источника, а целыми L-элементным последовательностям. Действительно, энтропия такого источника есть $H_L(X)$, а средняя длина кодовых слов равна $\overline{n}_L = \overline{n}L$. Тогда, применяя теорему Шеннона, можно записать:

$$\frac{H_L(X)}{\log m} \le \overline{n}_L < \frac{H_L(X)}{\log m} + 1,$$

откуда

$$\frac{H(X)}{L\log m} \le \overline{n} < \frac{H_L(X)}{L\log m} + \frac{1}{L}.$$

Для иллюстрации кодирования по алгоритму Хаффмана L-элементных последовательностей обратимся к следующему примеру.

Пусть источник сгенерировал сообщение

КОЛОКОЛ_ОКОЛО_КОЛОКОЛА

Рассмотрим, прежде всего, кодирование однобуквенных элементов.

В разд. 1.3 исследовались информационные характеристики данного сообщения; из статистики однобуквенных элементов следует, что энтропия первого приближения составляет 2,03 бит. Следовательно, существует возможность кодирования такого источника двоичным кодом, для которого значение средней длины заключено в границах $2{,}03 \le \overline{n} < 3{,}03$. Попробуем приблизиться к нижней границе, используя двоичный код Хаффмана.

Как нетрудно убедиться, кодовые слова с точностью до симметричных перестановок кодовых символов имеют следующий вид:

$$w_1 = w(\text{‘О’}) = 1,\ w_2 = w(\text{‘Л’}) = 01,\ w_3 = w(\text{‘К’}) = 000,$$
$$w_4 = w(\text{‘\_’}) = 0010,\ w_5 = w(\text{‘А’}) = 0011.$$

Тогда средняя длина кода составляет 2,16, что очень близко к нижней границы. При этом выигрыш от кодирования равномерным трехразрядным кодом составляет $\gamma = 1{,}39$.

Таблица 4.13

Статистика двухбуквенных сочетаний

x_k	КО	ЛО	Л_	ОК	ОЛ	О_	ЛА
$p(x_k)$	4/11	2/11	1/11	1/11	1/11	1/11	1/11

Будем теперь кодировать двоичным кодом Хаффмана двухбуквенные элементы источника. Соответствующая статистика представлена в табл. 4.13, откуда видно, что энтропия равна $H_L(X) = 2{,}20$, и, следовательно, средняя длина кода заключена в границах $1{,}10 \le \overline{n} < 1{,}60$, т. е. верхняя граница для данного случая меньше нижней границы, получающейся в предыдущем варианте, когда кодируются однобуквенные элементы.

Кодируя данный источник по Хаффману, имеем следующие слова:

$$w_1 = w(\text{‘КО’}) = 11,\ w_2 = w(\text{‘ЛО’}) = 01,\ w_3 = w(\text{‘Л\_’}) = 100,$$
$$w_4 = w(\text{‘О\_’}) = 001,\ w_5 = w(\text{‘ЛА’}) = 000,\ w_6 = w(\text{‘ОК’}) = 1011,$$
$$w_7 = w(\text{‘ОЛ’}) = 1010.$$

Средняя длина полученных кодовых комбинаций составляет $\overline{n}_L = 2{,}64$ на каждый двухбуквенный элемент, или $\overline{n} = 2{,}64 / 2 = 1{,}32$ на каждую букву, а выигрыш кодирования по сравнению с равномерным трехразрядным кодом равен $\gamma = 2{,}27$. Заметим, если выигрыш кодирования определять относительно равномерного 4-разрядого кода (по числу двухбуквенных сочетаний), то значение γ было бы ещё больше, однако с практической точки зрения такой подход вряд ли целесообразен.

Итак, использование алгоритма Хаффмана для кодирования многобуквенных сочетаний позволяет существенно повысить эффективность сжатия информации, однако, такой подход представляет скорее исторический и академический интерес.

Сама идея оптимального кодирования префиксными кодами по-прежнему активно используется, и реализация в той или иной форме хаффмановского алгоритма присутствует в подавляющем числе современных телекоммуникационных систем. Однако применение хаффмановского кодирования (даже в динамической адаптивной форме) для сжатия буквенной информации в настоящее время представляется малоэффективным: простые способы не позволяют достичь нужного коэффициента сжатия; сложные реализации, оперирующие с многобуквенными сочетаниями, требуют больших аппаратно-временных ресурсов.

Альтернативой при решении задач сжатия буквенной информации являются *методы словарного сжатия*, разработанные в 70–80 гг. XX века Дж. Зивом и А. Лемпелем, различные модификации которых в настоящее время являются неотъемлемой частью всех компьютерных операционных систем. К сожалению, рассмотрение принципов работы таких методов выходит за рамки данного учебного пособия.

ВОПРОСЫ И ЗАДАНИЯ К ГЛАВЕ 4

2.1. Что такое свойство однозначной декодируемости неравномерного кодового словаря?

2.2. Дайте определение задержки декодирования.

2.3. Какой код называется префиксным?

2.4. Что называется порядком узла кодового дерева?

2.5. Является ли неравенство Крафта необходимым и достаточным условием однозначной декодируемости произвольного кодового словаря?

2.6. Какой код называется оптимальным?

2.7. Могут ли быть средние дины кодовых слов в алгоритмах кодирования Шеннона–Фано и Хаффмана быть одинаковыми?

2.8. Сформулируйте теорему Шеннона для идеального дискретного канала.

2.9. Для следующих источников рассчитать энтропию первого и второго приближения, провести кодирование по алгоритмам Хаффмана и Шеннона–Фано для одно- и двухбуквенных сочетаний, определить в каждом случае среднюю длину кодового слова и сравнить её с нижней границей Шеннона:

а) КОЛОКОЛ_ОКОЛО_КОЛОКОЛА;

б) КОТ_ЛОМОМ_КОЛОЛ_ДРОВА;

в) ШЛА_МАША_ПО_ШОССЕ_И_СОСАЛА_СУШКУ;

г) ЛЕСА_ЛЫСЫ_ЛЕСА_ОБЕЗЛИСИЛИ_ЛЕСА_ОБЕЗЛОСИЛИ;

д) СОРОК_СОРОК_ЕЛИ_СЫРОЙ_СЫРОК.

ГЛАВА 5. ПРИНЦИПЫ ПОМЕХОУСТОЙЧИВОГО КОДИРОВАНИЯ

Использование методов сжатия информации, рассмотренных в предыдущей главе, возможно лишь в идеальном канале, когда можно пренебречь возникающими в канале ошибками, что, например, имеет место при передаче информации по внутренним магистралям какого-либо процессора. Однако большинство реальных каналов передачи информации по своей природе являются открытыми и подвержены заметному влиянию внешних помех. В этой связи становится актуальной задача помехоустойчивого кодирования, когда передаваемая информация подвергается дополнительным преобразованиям, с помощью которых возможна корректировка ошибок при приёме передаваемых данных.

В данной главе будут рассмотрены основные понятия и характеристики, связанные с передачей информации по неидеальным дискретным каналам, т. е. каналам в которых имеет место искажение информации, а также сформулирована и доказана одна из фундаментальных теорем всей теории информации — теорема Шеннона о передаче информации по дискретному каналу с шумами.

5.1. ПОСТАНОВКА ЗАДАЧИ КОДИРОВАНИЯ В ДИСКРЕТНОМ КАНАЛЕ

Вероятности $p(y'_j / y_i)$, являющиеся элементами канальной матрицы **P**, описывают то, насколько достоверно могут быть переданы по такому каналу различные символы его входного алфавита, и вместе с технической скоростью v следования символов полностью задают дискретный канал.

В действительности при организации системы передачи информации дискретный канал сам по себе, как правило, не существует, а является *дискретным отображением непрерывного* канала, который, в свою очередь, представляет собой составную часть системы передачи информации (рис. 5.1).

Исходными объектами системы, подлежащими передаче к удалённому получателю, являются дискретные сообщения x_r ($r = 1, \ldots, l$), которые

поступают с выхода источника с произвольной или фиксированной скоростью. Природа сообщений x_r может быть различна: это могут быть изначально дискретные символы (текстовые сообщения, управляющие команды и др.) или же отсчётные (дискретизированные) значения непрерывных сообщений, как это, например, имеет место при факсимильной передаче или в современных системах телевизионного и радиовещания.

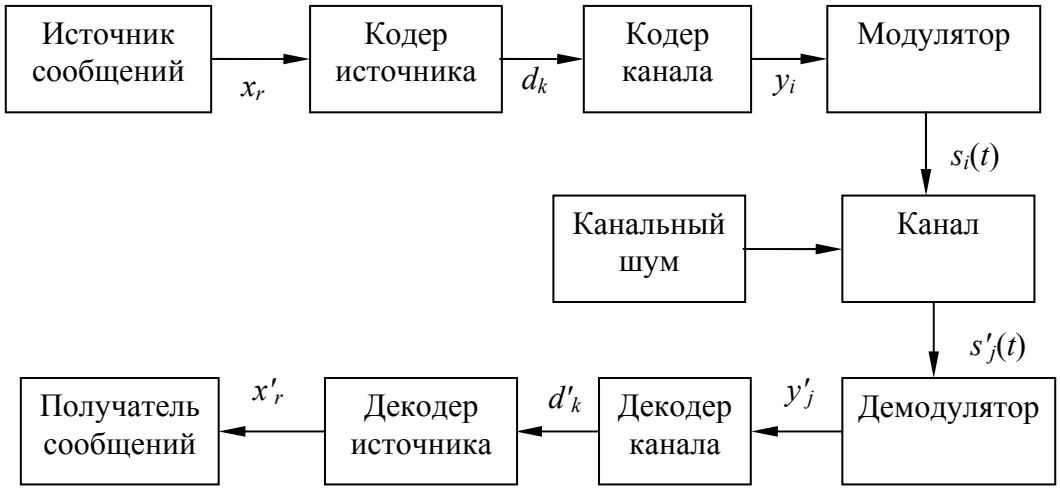


Рис. 5.1. Упрощённая структурная схема системы передачи информации

Сообщения x_r в конечном итоге должны быть доставлены получателю по какой-то физической среде, в соответствии с чем необходим *физический переносчик информации*. Таким переносчиком может быть электромагнитное колебание, ток (напряжение), акустические (или даже гравитационные) волны, распространяющиеся в воздухе или в более плотной среде. В любом случае это есть *физический сигнал* $s(t)$, для которого должны быть известны способы формирования, приёма и обработки, а также свойства распространения по каналу. Таким образом, одним из основных элементов системы является тот, где исходному сообщению x_r (или, что практически имеет место, группе таких сообщений) ставится в соответствие одна из форм сигнала:

$$x_r \ (r = 1, ..., l) \ \rightarrow \ s_i(t) \ (i = 1, ..., m) . \qquad (5.1.1a)$$

Технически это означает изменение каких-либо физических параметров переносчика информации, например, амплитуды и/или начальной фазы электромагнитной волны, в соответствии с заранее установленными пра-

вилами. Такая операция, как известно, называется *модуляцией*, а соответствующий функциональный блок — *модулятором*.

Принципиальным положением в теории построения систем передачи информации является *наличие шума в канале передачи*, делающего невозможным абсолютно точную передачу исходных сообщений. Это означает, что сигнал $s'_j(t)$ на выходе канала будет, вообще говоря, отличаться от входного сигнала $s_i(t)$, и если попытаться в *демодуляторе* (т. е. устройстве, выполняющем действия, обратные тем, что проводились в модуляторе) совершить обратное преобразование

$$s'_j(t)\ (j = 1, ..., m') \ \rightarrow\ x'_r\ (r = 1, ..., l). \qquad (5.1.1б)$$

то получателю будет доставлено сообщение x'_r, вообще говоря, отличающееся от исходного x_r.

В некоторых случаях подобная ситуация может оказаться приемлемой, например, когда речь идет о передаче значений какого-нибудь датчика, и ошибка в младших разрядах не является принципиальной. Однако в общем случае такая работа системы является недопустимой, и необходим поиск методов, способных в должной степени устранять негативное воздействие шумов. Такие методы, основанные на введении в передаваемую информацию дополнительных служебных символов, называются *методами помехоустойчивого кодирования*, а соответствующие функциональные блоки — *канальным кодером* (передающая часть) и *канальным декодером* (приемная часть).

Итак, упрощенная процедура передачи информации выглядит следующим образом. Последовательность исходных сообщений x_r (r = 1, …, l) преобразуется в последовательность кодовых символов y_i (i = 1, …, m). Далее, каждому кодовому символу y_i соответствует определённой формы сигнал $s_i(t)$, который при прохождении по каналу с шумами претерпевает искажения, превращаясь в сигнал $s'_j(t)$ (j = 1, …, m'). Задачей демодулятора является установление соответствия между $s'_j(t)$ и y'_j на основе анализа принятого сигнала с учетом всех сведений, которые могут быть известны о характеристиках шума, прежде всего, вероятностных (статистических). Наконец, имея совокупность кодовых символов y'_j, можно выносить решения x'_r (r = 1, …, l) об исходных переданных символах x_r. При этом количественной характеристикой качества работы системы является вероятность ошибочного приема символов.

Сравнивая действия, фигурирующие в описанной процедуре с тем, что изображено на рис. 5.1, можно заметить наличие еще двух функциональных узлов: *кодера источника* и *декодера источника*. Несмотря на внешнюю схожесть названий, их назначение совершенно отличается от канальных кодера и декодера. Основной задачей кодера источника является устранение (или, по крайней мере, существенное уменьшение) избыточности, содержащейся в исходных сообщениях x_r. Фактически действия кодера–декодера источника представляют собой передачу и обработку информации по идеальному дискретному каналу, основные особенности которых были рассмотрены в предыдущей главе.

Отметим, что в научно-технической литературе совокупность модулятора и демодулятора называется *модемом*, а совокупность кодера и декодера — *кодеком*.

Несмотря на то, что сколь-нибудь серьезное рассмотрение методов помехоустойчивого кодирования выходит далеко за рамки данного учебного пособия, тем не менее, представляется целесообразным рассмотреть хотя бы основные понятия.

Определим *код длины n и объемом M* как состоящее из M пар множество G, вида

$$G = \left\{\mathbf{u}_1, A_1; \mathbf{u}_2, A_2; \ldots; \mathbf{u}_M, A_M\right\}, \qquad (5.1.2)$$

где $\mathbf{u}_q$ ($q = 1, \ldots, M$) — последовательности входных символов длины n ($\mathbf{u}_q \in Y^n$), а подмножества $A_q \subseteq Y'^n$ ($q = 1, \ldots, M$), каждое из которых соответствует определённой входной последовательности $\mathbf{u}_q$, разбивают выходное множество значений на непересекающиеся подмножества. При этом последовательности $\mathbf{u}_q$ называются *кодовыми словами*, а подмножества A_q — *решающими областями*. Таким образом, задание кода означает задание как множества кодовых слов, так и правила, по которым приёмник принимает решение о переданном кодовом слове, а именно: если на выходе канала появляется последовательность $\mathbf{y}' \in A_q$, то приёмник выносит решение о том, что по каналу было передано кодовое слово $\mathbf{u}_q$.

Следует заметить, что множество решений $W = \{w_1, w_2, \ldots, w_{M+1}\}$ в общем случае содержит $M + 1$ элемент: подмножество элементов[1] w_q

[1] Не путать с обозначениями кодовых слов в идеальном дискретном канале, рассмотренном в предыдущей главе.

($q = 1, \ldots, M$), если $\mathbf{y}' \in A_q$, и дополнительно элемент w_{M+1} как решение о том, что последовательность на выходе канала не принадлежит ни одной из решающих областей.

Определим *скорость кода* R как отношение

$$R = \frac{\log M}{n}. \tag{5.1.3}$$

Из этого определения следует, что скорость кода представляет собой максимальное количество информации, которое может быть передано с помощью одного символа, поскольку $\log M$ есть максимальное количество информации, могущее быть переданным посредством кодового слова, и оно действительно передается в том случае, когда кодовые слова появляются равновероятно. Если скорость кода равна R бит/символ, то с помощью такого кода, имеющего объем $M = a^{nR}$ (a — основание логарифма, по которому определяется единица измерения количества информации), за время передачи одного кодового слова можно передать nR двоичных единиц информации. Исходя из этого, в дальнейшем выбранный код часто будем обозначать как $G(n, R)$, подчеркивая значения его параметров.

Понятно, что общее число M кодовых слов не может превышать числа m^n всех n-элементных последовательностей, образованных входными символами y_i ($i = 1, \ldots, m$), поэтому $R \le \log m$.

Пусть для заданного кода G по каналу передаётся кодовое слово $\mathbf{u}_q$. В том случае, если последовательность $\mathbf{y}'$ на выходе канала не принадлежит решающей области A_q, то возникает *ошибка декодирования кодового слова* $\mathbf{u}_q$, вероятность которой p_q вычисляется как

$$p_q = \sum_{\mathbf{y}' \in \overline{A_q}} p\left(\mathbf{y}'/\mathbf{u}_q\right), \tag{5.1.4}$$

где $\overline{A_q}$ — дополнение к A_q.

Количественной мерой, с помощью которой можно в целом охарактеризовать ненадежность передачи, может служить средняя вероятность ошибки

$$p_e = \sum_{q=1}^{M} p_q p(\mathbf{u}_q), \tag{5.1.5}$$

где $p(\mathbf{u}_q)$ — априорная вероятность передачи q-го кодового слова, и в случае, когда все кодовые слова равновероятны, что, как правило, имеет место, средняя вероятность ошибки равна

$$p_e = \frac{1}{M}\sum_{q=1}^{M} p_q \,. \tag{5.1.6}$$

Отметим, что иногда (особенно, в теоретических рассмотрениях) используется максимальная вероятность ошибки, определяемая как максимальное значение из всех вероятностей (5.1.4):

$$p_{\max} = \max\left\{p_1,\ldots,p_M\right\}. \tag{5.1.7}$$

Пусть задан некоторый канал[1]

$$\left\{Y^n, Y'^n, v, p\left(\mathbf{y}'/\mathbf{y}\right)\right\},$$

для которого выбраны кодовые слова $\mathbf{u}_1$, …, $\mathbf{u}_M$, и требуется указать такой выбор решающих областей A_1, …, A_M, который был бы в каком-то смысле наилучшим, например, минимизировал бы среднюю или максимальную вероятности ошибки декодирования.

Если $p(\mathbf{u}_1)$, …, $p(\mathbf{u}_M)$ — распределение вероятностей на множестве кодовых слов, то апостериорные (после получения сообщений $\mathbf{y}'$ на выходе канала) вероятности $p(\mathbf{u}_q / \mathbf{y}')$ равны

$$p\left(\mathbf{u}_q/\mathbf{y}'\right) = \frac{p\left(\mathbf{y}'/\mathbf{u}_q\right)p\left(\mathbf{u}_q\right)}{p\left(\mathbf{y}'\right)} = \frac{p\left(\mathbf{y}'/\mathbf{u}_q\right)p\left(\mathbf{u}_q\right)}{\sum\limits_{q=1}^{M} p\left(\mathbf{y}'/\mathbf{u}_q\right)p\left(\mathbf{u}_q\right)},\ q = 1,\ \ldots,\ M, \tag{5.1.8}$$

и среднюю вероятность ошибки p можно представить в виде

$$p_e = \sum_{Y'^n} p_e(\mathbf{y}')p(\mathbf{y}') = \sum_{q=1}^{M}\sum_{A_q} p_e(\mathbf{y}')p(\mathbf{y}'),$$

где $p_e(\mathbf{y}')$ — условная вероятность ошибки, когда на выходе канала появляется последовательность $\mathbf{y}'$.

---

[1] В разд. 1.1 дискретный канал задавался одномерными переходными вероятностями $p(y'_j / y_i)$, но указывалось, что они в общем случае зависят от ранее переданных символов. В этом смысле задание канала векторными переходными вероятностями $p(\mathbf{y}' / \mathbf{y})$ для n-элементных последовательностей $\mathbf{y}$ и $\mathbf{y}'$ является эквивалентным.

Для всех выходных последовательностей $\mathbf{y}'$, принадлежащих решающей области A_q, ошибка декодирования происходит в том случае, когда переданное кодовое слово отличается от $\mathbf{u}_q$, следовательно,

$$p_e(\mathbf{y}') = 1 - p\left(\mathbf{u}_q / \mathbf{y}'\right),\ \mathbf{y}' \in A_q,$$

откуда

$$p_e = 1 - \sum_{q=1}^{M} \sum_{A_q} p\left(\mathbf{y}'\right) p\left(\mathbf{u}_q / \mathbf{y}'\right). \qquad (5.1.9)$$

Одной из возможностей минимизации средней вероятности ошибки (5.1.9) является максимизация апостериорных вероятностей (5.1.8), когда в каждую решающую область A_q будут входить те выходные последовательности $\mathbf{y}'$, для которых условная вероятность $p(\mathbf{u}_q / \mathbf{y}')$ больше, чем для какого-нибудь другого кодового слова:

$$A_q = \{\mathbf{y}' : p(\mathbf{u}_q / \mathbf{y}') \geq p(\mathbf{u}_s / \mathbf{y}'),\ s = 1, 2, \ldots, M\}. \qquad (5.1.10)$$

Такая стратегия декодирования, т. е. стратегия разбиения решающих множеств, называется *декодированием по критерию максимума апостериорной вероятности* (МАВ-декодированием).

Из (5.1.8) следует, что апостериорные вероятности зависят как от канальных вероятностей $p(\mathbf{y}' / \mathbf{u}_q)$, так и от априорных распределений $p(\mathbf{u}_q)$ кодовых слов. При произвольных вероятностях $p(\mathbf{u}_q)$ стратегия, когда решающие области A_q состоят из выходных последовательностей $\mathbf{y}'$, для которых условная вероятность $p(\mathbf{y}' / \mathbf{u}_q)$ больше, чем для какого-нибудь другого кодового слова:

$$A_q = \{\mathbf{y}' : p(\mathbf{y}' / \mathbf{u}_q) \geq p(\mathbf{y}' / \mathbf{u}_s),\ s = 1, 2, \ldots, M\}. \qquad (5.1.11)$$

называется[1] *декодированием по критерию максимума правдоподобия* (МП-декодированием).

Очевидным преимуществом МП-декодирования является то, что оно может быть применено тогда, когда априорные вероятности кодовых слов неизвестны. В случае равновероятных кодовых слов стратегии декодирования (5.1.10) и (5.1.11) совпадают, ибо если $\mathbf{u}_q$ максимизирует апостериорную вероятность $p(\mathbf{u}_q / \mathbf{y}')$, то оно также максимизирует вероятность $p(\mathbf{y}' / \mathbf{u}_q)$.

Заметим, что, поскольку взаимная информация $I(\mathbf{y}'; \mathbf{y})$ между входной $\mathbf{y}$ и выходной $\mathbf{y}'$ последовательностями равна

[1] Название обусловлено тем, что условная вероятность $p(\mathbf{y}' / \mathbf{u}_q)$ в математической статистике называется *функцией правдоподобия*.

$$I(\mathbf{y}';\mathbf{y}) = \log \frac{p(\mathbf{y}'/\mathbf{y})}{p(\mathbf{y}')},$$

МП-декодирование можно рассматривать и как *декодирование по критерию максимума взаимной информации*.

Рассмотрим некоторые примеры.

Пусть в двоичном симметричном канале (3.3.3) на входном алфавите $Y = \{0, 1\}$ используется код с повторением, при котором выбраны два равновероятных кодовых слова $\mathbf{u}_1 = (0, 0, 0)$ и $\mathbf{u}_2 = (1, 1, 1)$. При МП-декодировании множество выходных сообщений

$$\{(0, 0, 0)\ (0, 0, 1)\ (0, 1, 0)\ (0, 1, 1)\ (1, 0, 0)\ (1, 0, 1)\ (1, 1, 0)\ (1, 1, 1)\}$$

требуется разделить на два непересекающихся подмножества так, чтобы максимизировать условные вероятности

$$p\left(y'^{(1)}, y'^{(2)}, y'^{(3)} \middle/ y^{(1)}, y^{(2)}, y^{(3)}\right) = \prod_{k=1}^{3} p\left(y'^{(k)} / y^{(k)}\right).$$

Для кодового слова $\mathbf{u}_1$ имеем:

$$P\{(0, 0, 0)\ /(0, 0, 0)\} = (1-p)^3;$$
$$P\{(0, 0, 1)\ /(0, 0, 0)\} = (1-p)^2 p;$$
$$P\{(0, 1, 0)\ /(0, 0, 0)\} = (1-p)^2 p;$$
$$P\{(0, 1, 1)\ /(0, 0, 0)\} = (1-p) p^2;$$
$$P\{(1, 0, 0)\ /(0, 0, 0)\} = (1-p)^2 p;$$
$$P\{(1, 0, 1)\ /(0, 0, 0)\} = (1-p) p^2;$$
$$P\{(1, 1, 0)\ /(0, 0, 0)\} = (1-p) p^2;$$
$$P\{(1, 1, 1)\ /(0, 0, 0)\} = p^3.$$

Для кодового слова $\mathbf{u}_2$ имеем:

$$P\{(0, 0, 0)\ /(1, 1, 1)\} = p^3;$$
$$P\{(0, 0, 1)\ /(1, 1, 1)\} = (1-p) p^2;$$
$$P\{(0, 1, 0)\ /(1, 1, 1)\} = (1-p) p^2;$$
$$P\{(0, 1, 1)\ /(1, 1, 1)\} = (1-p)^2 p;$$
$$P\{(1, 0, 0)\ /(1, 1, 1)\} = (1-p) p^2;$$
$$P\{(1, 0, 1)\ /(1, 1, 1)\} = (1-p)^2 p;$$
$$P\{(1, 1, 0)\ /(1, 1, 1)\} = (1-p)^2 p;$$
$$P\{(1, 1, 1)\ /(1, 1, 1)\} = (1-p)^3.$$

Считая, что $p < 1/2$, видим, что, например,

$$P\{(0, 0, 0)\ /(0, 0, 0)\} = (1-p)^3 > P\{(0, 0, 0)\ /(1, 1, 1)\} = p^3,$$

и выходное сообщение (0, 0, 0) следует отнести в решающую область A_1, связанную с кодовым словом $\mathbf{u}_1$. Полное разбиение выходного пространства на решающие области имеет следующий вид:

A_1: (0, 0, 0) (0, 0, 1) (0, 1, 0) (1, 0, 0)

A_2: (0, 1, 1) (1, 0, 1) (1, 1, 0) (1, 1, 1)

В общем случае для n-элементных последовательностей, как нетрудно видеть,

$$p(\mathbf{y}'/\mathbf{y}) = \prod_{k=1}^{n} p\left(y'^{(k)}/y^{(k)}\right) = p^t(1-p)^{n-t},$$

где t — количество позиций, в которых различаются $\mathbf{y}'$ и $\mathbf{y}$, которое называется *расстоянием Хемминга* (*хеммингoвым расстоянием*), и при МП-декодировании $\mathbf{y}'$ отображается в кодовое слово, соответствующее минимальному значению t.

В качестве другого примера рассмотрим МП-декодирование в двоичном стирающем канале (3.3.5), с вероятностью ошибки p и вероятностью стирания q, причем $q + 2p < 1$, для которого

$$p(\mathbf{y}'/\mathbf{y}) = \prod_{k=1}^{n} p\left(y'^{(k)}/y^{(k)}\right) = p^t q^s (1-p)^{n-s-t},$$

где s — число стираний в выходной последовательности, а t — количество нестертых позиций, в которых различаются последовательности $\mathbf{y}'$ и $\mathbf{y}$.

Число s определяется только выходом канала и не зависит от того, какое кодовое слово передавалось, поэтому при фиксированном s и условии $q + 2p < 1$ величина $p^t q^s (1 - p - q)^{n-s-t}$ монотонно убывает с ростом t. Таким образом, МП-декодирование в двоичном стирающем канале отображает каждую выходную последовательность в такое кодовое слово, которому соответствует минимальное значение t. Иными словами, декодирование осуществляется по минимуму хеммингова расстояния на нестёртых позициях.

Рассмотрим произвольный дискретный канал, на входном двоичном алфавите которого выбраны два n-элементных кодовых слова $\mathbf{u}_1$ и $\mathbf{u}_2$, а на выходе производится МП-декодирование, при котором декодируется сообщение 1, если $p(\mathbf{y}'/\mathbf{u}_1) > p(\mathbf{y}'/\mathbf{u}_2)$, и сообщение 2 — в противном случае. При этом, согласно (5.7.4), вероятность ошибки при посылке сообщения 1 равна

$$p_1 = \sum_{\mathbf{y}' \in \overline{A_1}} p\left(\mathbf{y}'/\mathbf{u}_1\right).$$

Поскольку в решающей области $A_2 = \overline{A_1}$ выполняется неравенство

$$p(\mathbf{y}' / \mathbf{u}_2) \geq p(\mathbf{y}' / \mathbf{u}_1),$$

для любого[1] значения s, $0 < s < 1$, также справедливо неравенство

$$p(\mathbf{y}' / \mathbf{u}_2)^s \geq p(\mathbf{y}' / \mathbf{u}_1)^s.$$

Отсюда следует, что

$$p\left(\mathbf{y}'/\mathbf{u}_1\right) = p\left(\mathbf{y}'/\mathbf{u}_1\right)\frac{p\left(\mathbf{y}'/\mathbf{u}_1\right)^s}{p\left(\mathbf{y}'/\mathbf{u}_1\right)^s} \leq p\left(\mathbf{y}'/\mathbf{u}_1\right)^{1-s} p\left(\mathbf{y}'/\mathbf{u}_2\right)^s,$$

так что существует возможность оценки сверху вероятности p_1 при помощи суммирования по всем выходным последовательностям:

$$p_1 = \sum_{\mathbf{y}' \in \overline{A_1}} p\left(\mathbf{y}'/\mathbf{u}_1\right) \leq \sum_{\mathbf{y}'} p\left(\mathbf{y}' / \mathbf{u}_1\right)^{1-s} p\left(\mathbf{y}' / \mathbf{u}_2\right)^s \qquad (5.1.12a)$$

и аналогично для вероятности p_2:

$$p_2 = \sum_{\mathbf{y}' \in \overline{A_2}} p\left(\mathbf{y}'/\mathbf{u}_2\right) \leq \sum_{\mathbf{y}'} p\left(\mathbf{y}' / \mathbf{u}_2\right)^{1-r} p\left(\mathbf{y}' / \mathbf{u}_1\right)^r, \; 0 < r < 1. \quad (5.1.12б)$$

Для постоянных каналов соотношения (5.1.12) упрощаются:

$$p_1 \leq \sum_{Y_1'} \cdots \sum_{Y_n'} \prod_{k=1}^{n} p\left(y'^{(k)}/y^{(k)}(\mathbf{u}_1)\right)^{1-s} p\left(y'^{(k)}/y^{(k)}(\mathbf{u}_2)\right)^s =$$

$$= \sum_{Y_1'} p\left(y'^{(1)}/y^{(1)}(\mathbf{u}_1)\right)^{1-s} p\left(y'^{(1)}/y^{(1)}(\mathbf{u}_2)\right)^s \times \cdots \times$$

$$\times \sum_{Y_n'} p\left(y'^{(n)}/y^{(n)}(\mathbf{u}_1)\right)^{1-s} p\left(y'^{(n)}/y^{(n)}(\mathbf{u}_2)\right)^s =$$

$$= \prod_{k=1}^{n} \sum_{Y_k'} p\left(y'^{(k)}/y^{(k)}(\mathbf{u}_1)\right)^{1-s} p\left(y'^{(k)}/y^{(k)}(\mathbf{u}_2)\right)^s,$$

где $y^{(k)}(\mathbf{u}_i)$ обозначает произвольный символ кодового слова $\mathbf{u}_i$ ($i = 1, 2$), находящийся на k-й позиции. (Заметим, что запись $y_i^{(k)}$ в данном случае была бы некорректной, поскольку означает, что на k-й позиции находится конкретно символ y_i.)

[1] Неравенство справедливо и для $s \geq 1$, но здесь это не будет использоваться.

В последнем равенстве сумма по Y'_k имеет важное значение для всей теории помехоустойчивого кодирования; обозначим ее как

$$g_k(s) = \sum_{Y'_k} p\left(y'^{(k)} / y_1^{(k)}\right)^{1-s} p\left(y'^{(k)} / y_2^{(k)}\right)^s, \qquad (5.1.13)$$

тогда

$$p_1 \le \prod_{k=1}^{n} g_k(s).$$

Заметим, что если подставить $1 - s$ вместо r, то получаются одинаковые границы для p_1 и p_2. Поэтому выпишем универсальную оценку для вероятности ошибки:

$$p_m \le \prod_{k=1}^{n} g_k(s),\ m = 1, 2;\ 0 < s < 1. \qquad (5.1.14)$$

Может оказаться так, что для некоторых типов каналов граничные значения $g_k(0)$ и $g_k(1)$ могут быть формально не определены. Тогда их можно доопределить односторонними пределами:

$$g_k(0) = \lim_{s \to 0^+} g_k(s) = \sum_{Y'_k} P\left(y'^{(k)} / y_1^{(k)}\right),$$

где суммирование производится по всем $y'^{(k)}$, для которых $p(y'^{(k)}/y_2^{(k)}) \neq 0^+$, и

$$g_k(1) = \lim_{s \to 1^-} g_k(s) = \sum_{Y'_k} P\left(y'^{(k)} / y_2^{(k)}\right),$$

где суммирование производится по всем $y'^{(k)}$, для которых $p(y'^{(k)}/y_1^{(k)}) \neq 0$. В любом случае и $g_k(0)$, и $g_k(1)$ всегда меньше или равны единицы. Отсюда (если взять вторую производную) функция $g_k(s)$ вогнута на интервале $[0; s]$ и на этом интервале не превосходит единицу.

Рассмотрим двоичный симметричный канал (3.3.3) с параметром $p < 1/2$, в который подаются n-элементные последовательности нулей и единиц:

$$\mathbf{u}_1 = (0\ 0\ \ldots\ 0),\ \mathbf{u}_2 = (1\ 1\ \ldots\ 1).$$

Тогда для всех $k = 1, \ldots, n$ функция $g_k(s)$ равна

$$g_k(s) = p^{1-s}(1-p)^s + p^s(1-p)^{1-s}. \qquad (5.1.15)$$

Беря производную, нетрудно убедиться, что стоящее в правой части выражение минимизируется при $s = 1/2$, при этом

$$\min_{[0;1]} g_k(s) = p_k(1/2) = 2\sqrt{p(1-p)},$$

т. е. имеет место оценка

$$p_m \le 2^n \left(p(1-p) \right)^{n/2}, \; m = 1, 2. \qquad (5.1.16)$$

На рис. 5.2 изображён вид функции $g_k(s)$ для двоичного симметричного канала.

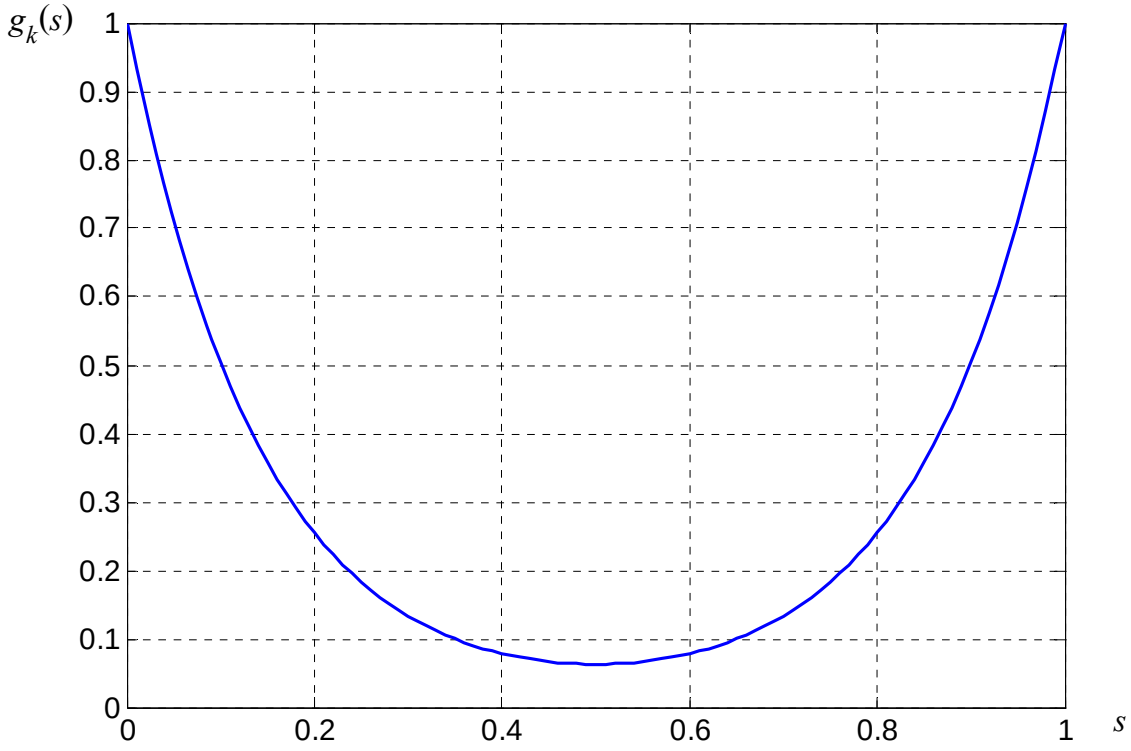


Рис. 5.2. Функция $g_k(s)$ для двоичного симметричного канала при $p = 10^{-3}$

С другой стороны, для этого простого примера вероятности p_1 и p_2 можно подсчитать в явном виде. Действительно, вероятность $p(\mathbf{y}' / \mathbf{u}_2)$ будет больше или равна вероятности $p(\mathbf{y}' / \mathbf{u}_2)$, если принятая последовательность содержит $n/2$ и более единиц (для простоты, будем полагать, что n чётно). Поскольку вероятность того, что в принятой последовательности *какие-либо*[1] i символы из n равны единице, подчиняется биномиальному распределению

$$p_n(i) = C_n^i p^i (1-p)^{n-i} = \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i}$$

[1] Вероятность того, что *определённые* i символы в последовательности равны единице, есть $p^i(1-p)^{n-i}$.

(что является отличительной чертой постоянного симметричного канала), вероятность ошибки p_1 равна

$$p_1 = \sum_{i=n/2}^{n} C_n^i p^i (1-p)^{n-i} . \qquad (5.1.17)$$

Биномиальное распределение как функция целочисленного аргумента имеет максимум при $i_0 \approx np$, и слагаемые в (5.1.17) также концентрируются вблизи этого значения. При этом наибольшим слагаемым в сумме является первое, в котором $i = n/2$. Применяя формулу Стирлинга для факториала:

$$n! \approx \sqrt{2\pi n} n^n \exp(-n) ,$$

получим выражение для биномиального коэффициента

$$C_n^{n/2} \approx \sqrt{2/(\pi n)} 2^n ,$$

отсюда

$$p_1 \approx \sqrt{2/(\pi n)} 2^n \left(p(1-p) \right)^{n/2} . \qquad (5.1.18)$$

Сравнение (5.1.16) и (5.1.18) показывает, что в обоих случаях имеет место экспоненциальная зависимость вероятности ошибки от длины кодового слова. Однако во втором случае наряду с этим вероятность ошибки дополнительно убывает как $n^{-1/2}$, так что верхняя граница (5.1.16) является более грубой оценкой. К сожалению, получить относительно точные формулы вычисления вероятности ошибки, кроме как в двоичном симметричном канале, не удается.

5.2. ТЕОРЕМЫ КОДИРОВАНИЯ ШЕННОНА ДЛЯ ДИСКРЕТНОГО КАНАЛА С ПОМЕХАМИ

В данном разделе будет сформулировано и доказано фундаментальное свойство, касающееся возможности обеспечить передачу информации со сколь угодно малой вероятностью ошибки. Это свойство известно в литературе как *теорема кодирования Шеннона для дискретного канала*.

Обычно различают *прямую* и *обратную* теоремы кодирования. Прямая теорема кодирования утверждает о существовании такого способа кодирования информации n-элементными последовательностями входных символов, что при любом значении кодовой скорости R, меньшей пропускной способности канала C, вероятность ошибочного декодирования может быть сделана сколь угодно малой. Обратная теорема кодирования

говорит о невозможности обеспечить сколь угодно малую вероятность ошибки, если кодовая скорость превысит пропускную способность канала.

Сразу заметим, что несмотря на фундаментальность результатов, содержащихся в формулировке теоремы, её доказательство *неконструктивно*, т. е. способ доказательства не даёт никаких конкретных предложений по методу конструирования такого замечательного кода, а также при этом ничего нельзя сказать об особенностях поведении вероятности ошибочного декодирования в зависимости от параметров кода и канала.

Рассматриваемое в данном разделе доказательство теоремы Шеннона основано на использовании двух фундаментальных неравенств — *неравенства Фано* и *неравенства Файнстейна*, полезных не только применительно к данному случаю, но также при решении большого числа других теоретико-информационных задач.

Рассмотрим произвольный дискретный канал, в котором для повышения верности передачи информации используется помехоустойчивое кодирование.

Пусть $U = \{\mathbf{u}_1, \ldots, \mathbf{u}_M\}$ — множество кодовых слов, соответствующих решающим областям $A_1, \ldots, A_M$, и $(p(\mathbf{u}_1), \ldots, p(\mathbf{u}_M))$ — вероятностный вектор их появления. Тем самым, определён ансамбль $\{UW, p(\mathbf{u}, w)\}$, элементами которого являются п*а*ры “кодовое слово — решение” $(\mathbf{u}_q, w_s)$ $(q = 1, \ldots, M;\ s = 1, \ldots, M + 1)$, а совместные вероятности равны

$$p\left(\mathbf{u}_q, w_s\right) = p\left(w_s / \mathbf{u}_q\right) p\left(\mathbf{u}_q\right),$$

где вероятность принятия решения в пользу s-го решения

$$p\left(w_s / \mathbf{u}_q\right) = \sum_{\mathbf{y}' \in A_s} p\left(\mathbf{y}' / \mathbf{u}_q\right),\ s = 1, \ldots, M,$$

а вероятность отказа от принятия решения в пользу какого-нибудь из кодовых слов есть

$$p\left(w_{M+1} / \mathbf{u}_q\right) = 1 - \sum_{s=1}^{M} p\left(w_s / \mathbf{u}_q\right).$$

Поскольку вероятность ошибочного декодирования зависит как от кодовых слов, так и от вида решающих областей, попытаемся уйти от ансамбля пар “кодовое слово — решение”, рассматривая вместо этого ансамбль пар “решение — ошибка”.

Обозначим через ε_s *ошибку s-го решения*, т. е. событие, состоящее в появлении пары $(\mathbf{u}_q, w_s)$, $q \neq s$, тогда условная (при условии принятия s-го решения) вероятность $p(\varepsilon_s)$ ошибочного декодирования q-го кодового слова может быть представлена как

$$p(\varepsilon_s) = \sum_{q \neq s} p\left(\mathbf{u}_q / w_s\right) = 1 - p\left(\mathbf{u}_s / w_s\right). \qquad (5.2.1)$$

Отметим, что в (5.2.1) фигурируют "обратные" вероятности $p(\mathbf{u}_q / w_s)$, поскольку "прямыми" являются вероятности $p(w_s / \mathbf{u}_q)$ вычисляемые по заданным канальным вероятностям $p(\mathbf{y}' / \mathbf{u}_q)$. При необходимости "обратные" вероятности можно вычислить, используя формулу Байеса:

$$p(\mathbf{u}_q / w_s) = \frac{p(\mathbf{u}_q, w_s)}{p(w_s)} = \frac{p(w_s / \mathbf{u}_q)p(\mathbf{u}_q)}{p(w_s)} = \frac{1}{p(w_s)} \sum_{\mathbf{y}' \in A_s} p\left(\mathbf{y}' / \mathbf{u}_q\right) p(\mathbf{u}_q),$$

где безусловная вероятность s-го решения получается путём суммирования совместной вероятности $p(\mathbf{u}_q, w_s)$ по всем кодовым словам:

$$p(w_s) = \sum_{q=1}^{M} p(\mathbf{u}_q, w_s).$$

Рассмотрим множество $E = \left\{\varepsilon_s, \overline{\varepsilon}_s\right\}$, где $\overline{\varepsilon}_s$ — событие, обратное к ε_s, т. е. *правильное решение*, наступающее при появлении любой пары $(\mathbf{u}_s, w_s)$ ($s = 1, \ldots, M$). Несмотря на то, что ошибка ε_s может быть вызвана любым s-м решением, множество E всегда состоит только из двух элементов: "ошибка" и "отсутствие ошибки", вероятности которых $p(\varepsilon_s)$ и $1 - p(\varepsilon_s)$, тем не менее, зависят от принимаемых решений. Тогда энтропия такого множества — условная энтропия ошибки $H(E / w_s)$ — равна

$$H(E / w_s) = -p(\varepsilon_s)\log p(\varepsilon_s) - \left(1 - p(\varepsilon_s)\right)\log\left(1 - p(\varepsilon_s)\right),$$

а средняя (безусловная) энтропия ошибки есть

$$H(E) = \sum_{s=1}^{M} H(E / w_s)p(w_s).$$

С другой стороны, на множестве E можно сосчитать безусловное распределение

$$\mu = \sum_{s=1}^{M} p(\varepsilon_s)p(w_s) \qquad (5.2.2)$$

и найти "стандартную" энтропию двоичного алфавита с независимыми элементами:

$$H(\mu) = -\mu \log \mu - (1-\mu)\log(1-\mu).$$

Установим связь между энтропиями $H(U / W)$, $H(E)$ и вероятностью ошибки μ.

Для всех $s < M + 1$ имеем

$$H\left(U/w_s\right) = -\sum_{q=1}^{M} p\left(\mathbf{u}_q/w_s\right)\log p\left(\mathbf{u}_q/w_s\right) =$$

$$= -p\left(\mathbf{u}_s/w_s\right)\log p\left(\mathbf{u}_s/w_s\right) - \left[1-p\left(\mathbf{u}_s/w_s\right)\right]\log\left[1-p\left(\mathbf{u}_s/w_s\right)\right] -$$

$$-\sum_{q\neq s} p\left(\mathbf{u}_q/w_s\right)\log p\left(\mathbf{u}_q/w_s\right) + \left[1-p\left(\mathbf{u}_s/w_s\right)\right]\log\left[1-p\left(\mathbf{u}_s/w_s\right)\right].$$

Из (5.2.1) следует, что $p(\mathbf{u}_s / w_s) = 1 - p(\varepsilon_s)$, поэтому первые два слагаемые представляют собой условную энтропию $H(E / w_s)$, тогда, учитывая, что

$$p(\varepsilon_s) = \sum_{q\neq s} p\left(\mathbf{u}_q/w_s\right),$$

имеем:

$$H\left(U / w_s\right) = H\left(E / w_s\right) - \sum_{q\neq s} p\left(\mathbf{u}_q/w_s\right)\log p\left(\mathbf{u}_q/w_s\right) + p(\varepsilon_s)\log p(\varepsilon_s) =$$

$$= H\left(E / w_s\right) - \sum_{q\neq s} p\left(\mathbf{u}_q/w_s\right)\log p\left(\mathbf{u}_q/w_s\right) + \sum_{q\neq s} p\left(\mathbf{u}_q/w_s\right)\log p(\varepsilon_s) =$$

$$= H\left(E / w_s\right) - \sum_{q\neq s} p\left(\mathbf{u}_q/w_s\right)\frac{\log p\left(\mathbf{u}_q/w_s\right)}{\log p(\varepsilon_s)} =$$

$$= H\left(E / w_s\right) - p(\varepsilon_s)\sum_{q\neq s}\frac{p\left(\mathbf{u}_q/w_s\right)}{p(\varepsilon_s)}\frac{\log p\left(\mathbf{u}_q/w_s\right)}{\log p(\varepsilon_s)}. \qquad (5.2.3)$$

Из соотношения (5.2.1) также следует, что

$$\sum_{q\neq s}\frac{p\left(\mathbf{u}_q/w_s\right)}{p_s} = 1,$$

поэтому второе слагаемое в (5.2.3) представляет собой умноженную на $p(\varepsilon_s)$ энтропию некоторого ансамбля, состоящего из $M - 1$ сообщения, вероятности которых равны $p(\mathbf{u}_q / w_s) / p(\varepsilon_s)$. Оценим эту энтропию сверху величиной $\log M$, тогда

$$H(U / w_s) \le H(E / w_s) + p(\varepsilon_s)\log(M-1) < H(E / w_s) + p(\varepsilon_s)\log M =$$

$$= -p(\varepsilon_s)\log p(\varepsilon_s) - \left(1-p(\varepsilon_s)\right)\log\left(1-p(\varepsilon_s)\right) + p(\varepsilon_s)\log M. \qquad (5.2.4)$$

Если же $s = M + 1$, что соответствует случаю отказа от принятия решения в пользу какого-либо кодового слова, то, как следует из (5.2.1), вероятность ошибочного декодирования $p(\varepsilon_{M+1})$ равна единице, и, следовательно, соответствующая условная энтропия $H(E / w_{M+1})$ равна нулю. Тогда, вводя в неравенство (5.2.4) формально нуль и единицу, можно записать:

$$H(U / w_{M+1}) = -\sum_{q=1}^{M} p(\mathbf{u}_q / w_{M+1}) \log p(\mathbf{u}_q / w_{M+1}) \leq$$

$$\leq \log M = H(E / w_{M+1}) + p(\varepsilon_{M+1}) \log M . \qquad (5.2.5)$$

Усредним оба неравенства (5.2.4) и (5.2.5), т. е. умножим их на $p(w_s)$ и просуммируем по всем s ($s = 1, \ldots, M + 1$). Тогда получим искомую связь между энтропиями $H(U/W)$, $H(E)$ и вероятностью ошибки μ в виде неравенства

$$H(U / W) \leq H(E) + \mu \log M . \qquad (5.2.6)$$

Более того, учитывая, что энтропия $H(E)$, полученная путём усреднения условных энтропий $H(E / w_s)$, не превосходит энтропию $H(\mu)$ алфавита с независимыми элементами, можно записать усугубляющее соотношение

$$H(U / W) \leq H(\mu) + \mu \log M$$

или

$$H(U / W) \leq -\mu \log \mu - (1 - \mu) \log(1 - \mu) + \mu \log M , \qquad (5.2.7)$$

которое обычно и называется неравенством Фано.

Величину $H(U / W)$, представляющую собой среднюю условную информацию совокупности кодовых слов при фиксированном множестве решений, часто называют *ненадежностью передачи с помощью кода G*, поскольку она характеризует количество информации, потерянное при передаче. Следовательно, неравенство Фано устанавливает связь между ненадежностью передачи и средней вероятностью ошибки декодирования для выбранного кода.

Неравенство Фано можно трактовать следующим образом. Для того, чтобы получатель мог точно интерпретировать сообщение, он, прежде всего, должен знать, совершает или нет кодер ошибку, и необходимое для этого среднее количество информации равно $H(\varepsilon)$. Далее, если получатель уверен, что произошла ошибка декодирования, то он должен дополнительно установить, какое из оставшихся $M - 1$ кодовых слов в действительности было передано, и необходимое для этого среднее количество информа-

ции не превосходит величины $\log M$. Такая необходимость возникает с вероятностью μ, поэтому среднее количество дополнительной информации не превосходит величины $\mu \log M$.

Так как $H(\varepsilon)$ полностью определяется значением μ, вся правая часть неравенства Фано зависит только от μ. Рассматривая функцию

$$f(\mu) = H\left[\varepsilon(\mu)\right] + \mu \log M,$$

можно заметить (рис. 5.3), что $f(\mu) \geq 0$, причем неравенство нулю имеет место только для $\mu = 0$. Функция возрастает на интервале $[0; M/(M+1))$ и, соответственно, убывает на интервале $(M/(M+1); 1]$. В точке $\mu = M/(M+1)$ функция имеет максимум, равный $f_{\max} = \log(M+1)$. При этом, поскольку реальные значения M много больше единицы, то максимум достигается практически на правой границе интервала $[0; 1]$.

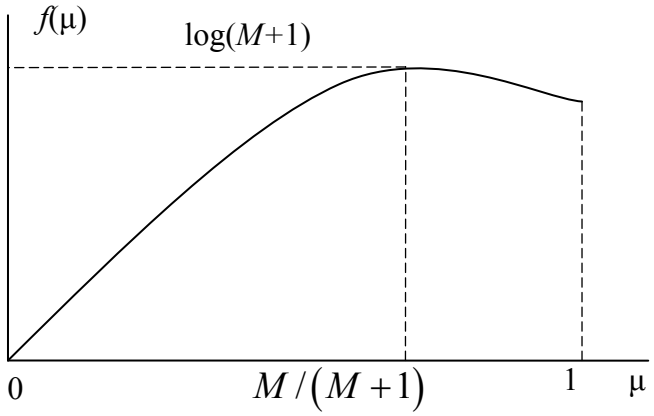


Рис. 5.3. График функции $f(p)$

Используем неравенство Фано для доказательства обратной теоремы кодирования.

Рассмотрим произвольный дискретный канал, для входных последовательностей которого задано распределение $p(\mathbf{y})$. Такое распределение наряду с условными вероятностями $p(y'_j / y_i)$ ($i = 1, \ldots, m$; $j = 1, \ldots, m'$), задающими канал, определяет ансамбль $\{Y^n Y'^n, (\mathbf{y}, \mathbf{y}')\}$ n-элементных сообщений на входе и выходе. Среднее количество взаимной информации, передаваемое по каналу с помощью таких последовательностей, равно

$$I(Y'^n; Y^n) = \sum_{Y^n} \sum_{Y'^n} p(\mathbf{y}'/\mathbf{y}) p(\mathbf{y}) \log \frac{p(\mathbf{y}'/\mathbf{y})}{p(\mathbf{y}')},$$

где

$$p(\mathbf{y}') = \sum_{Y^n} p(\mathbf{y}'/\mathbf{y}) p(\mathbf{y}),$$

а C_n — пропускная способность канала, достигаемая на n-элементных входных последовательностях:

$$C_n = \max_{P(Y^n)} \frac{v}{n} I(Y'^n; Y^n).$$

Предположим теперь, что для повышения верности передачи выбран некоторый код G с длиной n, скоростью R и набором кодовых слов $U = \{\mathbf{u}_1, \ldots, \mathbf{u}_M\}$. Зададим на множестве Y^n следующее распределение вероятностей входных последовательностей:

$$p(\mathbf{y}) = \begin{cases} 1/M, & \mathbf{y} \in U; \\ 0, & \mathbf{y} \notin U; \end{cases}$$

тогда, очевидно, $I(Y^n, Y'^n) = I(U, Y'^n)$. При этом, по определению пропускной способности канала, $I(U, Y'^n) \le nC_n / v$.

Пусть W — множество решений, являющихся результатом отображения множества U кодовых слов в соответствующие решающие области. Поскольку любое преобразование сообщения не может увеличить содержащуюся в нем информацию, для взаимной информации справедлива оценка $I(U; W) \le I(U; Y'^n)$. Запишем

$$H(U/W) = H(U) - I(U;W),$$

тогда, учитывая, что $H(U) = \log M$, $I(U; Y'^n) \le nC_n / v$ и $I(U; W) \le I(U; Y'^n)$, получим

$$H(U/W) = \log M - I(U;W) \ge \log M - nC_n / v = n\left(R - C_n / v\right).$$

Воспользуемся неравенством Фано, справедливым для любого кода и любого распределения вероятностей кодовых слов. С учетом последнего неравенства имеем:

$$H(\mu) + \mu^{(n)} \log M \ge H(U/W) \ge n\left(R - C_n / v\right), \qquad (5.2.8)$$

где $\mu^{(n)}$ — средняя вероятность ошибки для выбранного кода с n-элементными кодовыми словами. Пусть $\mu_0^{(n)}$ — наименьший корень уравнения

$$-\mu \log \mu - (1-\mu)\log(1-\mu) + \mu \log M = n\left(R - C_n / v\right),$$

тогда, если $R > C_n / v$, то неравенство (5.2.8) означает (рис. 5.3), что $\mu^{(n)} \geq \mu_0^{(n)}$, т. е. средняя вероятность ошибки декодирования ограничена снизу. При этом, в силу ограниченности самой энтропии $H(\mu)$

$$\mu_0^{(n)} = \frac{n\left(R - C_n / v\right) - H(\mu)}{\log M} = 1 - \frac{C_n}{Rv} - \frac{H(\mu)}{\log M} \underset{n\to\infty}{\to} 1 - \frac{C}{Rv},$$

т. е. даже для сколь угодно большой длины кода вероятность ошибки не может быть приближена к нулю.

Еще раз сформулируем обратную теорему кодирования для дискретного канала с шумом.

Пусть C / v — нормированная пропускная способность дискретного канала и задана кодовая скорость $R > C / v$. Тогда существует положительное число δ, зависящее от R, такое, что для любого кода, характеризуемого длиной n и кодовой скоростью R, вероятность ошибки декодирования μ удовлетворяет соотношению $\mu \geq \delta$.

Будем по-прежнему рассматривать дискретный канал, задаваемый переходными вероятностями $p(\mathbf{y}' / \mathbf{y})$, $\mathbf{y} \in Y^n$, $\mathbf{y}' \in Y'^n$. Пусть $I(\mathbf{y}'; \mathbf{y})$ — взаимная информация[1] между двумя n-элементными последовательностями $\mathbf{y}'$ и $\mathbf{y}$:

$$I\left(\mathbf{y}';\mathbf{y}\right) = \log \frac{p\left(\mathbf{y}'/\mathbf{y}\right)}{p\left(\mathbf{y}'\right)}.$$

Построим код длины n, выбирая кодовые слова из множества U, являющегося подмножеством алфавита Y^n на входе канала. Для того чтобы выбрать решающие области, каждой входной последовательности $\mathbf{y}$ поставим в соответствие множество D выходных последовательностей $\mathbf{y}'$, таких, что количество взаимной информации между ними всеми и входной последовательностью $\mathbf{y}$ превышает заданную величину $n\gamma$:

$$D \equiv D(\mathbf{y}) = \left\{\mathbf{y}' \in Y'^n : I\left(\mathbf{y}';\mathbf{y}\right) > n\gamma\right\}. \qquad (5.2.9)$$

Тем самым определён ансамбль пар $Q_\gamma = \left\{\left(\mathbf{y}',\mathbf{y}\right) : \mathbf{y} \in Y^n, \mathbf{y}' \in D\right\}$.

Пусть $\mathbf{u}_1, \ldots, \mathbf{u}_M$ — кодовые слова кода $G(n, R)$, выбранного для заданного дискретного канала. Разобьём пространство выходных последовательностей на решающие области $\{A_q\}$ ($q = 1, \ldots, M$) следующим образом[1]:

[1] Не путать со средней взаимной информацией входного и выходного алфавитов.

$$A_1 = D(\mathbf{u}_1)\,;\; A_q = D(\mathbf{u}_q) \setminus \bigcup_{p=1}^{q-1} A_p\,,\; q = 1,\, \ldots,\, M, \qquad (3.2.10)$$

т. е. первую решающую область A_1 составляют все те выходные последовательности $\mathbf{y}'$, взаимная информация которых с первым кодовым словом $\mathbf{u}_1$ превышает число $n\gamma$; во вторую решающую область A_2 входят все те выходные последовательности $\mathbf{y}'$, взаимная информация которых со вторым кодовым словом $\mathbf{u}_2$ превышает число $n\gamma$, но за вычетом всех последовательностей, вошедших в область A_1 и т. д. При этом соответствующие условные вероятности ошибки декодирования p_q ($q = 1,\, \ldots,\, M$) удовлетворяют неравенствам

$$p_q \equiv P(\mathbf{y}' \notin A_q \,/\, \mathbf{u}_q) \le \alpha\,,\; q = 1,\, \ldots,\, M, \qquad (5.2.11)$$

где α — произвольное заданное число в интервале (0; 1]. Обозначим через $p_{\max}$ максимальное из всех значений p_q.

Подчеркнем, что построенный код $G(n, R)$ является максимальным, т. е. к нему нельзя более добавить ни одного кодового слова из множества U, чтобы при этом не нарушить условий (5.2.9)–(5.2.11).

Неравенство Файнстейна утверждает существование такого кода $G(n, R)$, что описанные выше действия по формированию множества D согласно (5.2.9)–(5.2.11) могут быть осуществлены таким образом, что для максимальной вероятности ошибочного декодирования будет выполняться соотношение

$$p_{\max} \le \frac{1}{P(U)}\left[2^{-n(\gamma - R)} + 1 - P(Q_\gamma)\right], \qquad (5.2.12)$$

или, что эквивалентно,

$$2^{nR} \ge 2^{n\gamma}\left[\, p_{\max} P(U) - 1 + P(Q_\gamma)\right], \qquad (5.2.12a)$$

где

$$P(Q_\gamma) = \sum_{(\mathbf{y},\mathbf{y}') \in Q_\gamma} \left(\mathbf{y}' \in D/\mathbf{y}\right) p\left(\mathbf{y}\right)$$

---

[1] Напомним, что $B_1 \setminus B_2$ есть множество всех элементов, принадлежащих B_1, но не принадлежащих B_2.

есть мера[1] множества Q_γ, т. е. вероятность появления входящих в него пар последовательностей (**y**, **y**′), а

$$P(U) = \sum_{\mathbf{y} \in U} p(\mathbf{y})$$

есть мера всего множества U.

При рассмотрении множества Q_γ будем предполагать, что для него выполняется неравенство

$$P(Q_\gamma) > 1 - p_{\max} P(U). \qquad (5.2.13)$$

В противном случае, когда $P(Q_\gamma) \le 1 - p_{\max} P(U)$, правая часть в (5.2.12*a*) обнуляется, и неравенство оказывается справедливым для любых значений nR, в частности, при $M = 2^{nR} = 1$, т. е. для кода, содержащего лишь одно кодовое слово и имеющего, следовательно, нулевую вероятность ошибочного декодирования.

Покажем, прежде всего, что $M \ge 1$, т. е. существует, по крайней мере, одно кодовое слово из множества U, удовлетворяющее наложенным выше условиям. Для этого предположим обратное, а именно, что для всех последовательностей **y** из множества U справедливо неравенство

$$\alpha < P(\mathbf{y}' \in D/\mathbf{y}) \le 1$$

или, что эквивалентно,

$$P(\mathbf{y}' \in D/\mathbf{y}) \le 1 - \alpha. \qquad (5.2.14)$$

Тогда, усредняя (5.2.14) по всему множеству U, имеем:

$$\sum_{\mathbf{y} \in U} P(\mathbf{y}' \in D/\mathbf{y})\, p(\mathbf{y}) \le (1 - \alpha) P(U).$$

С другой стороны, мера множества Q_γ равна

$$P(Q_\gamma) = \sum_{Y^n} P(\mathbf{y}' \in D/\mathbf{y})\, p(\mathbf{y}).$$

Представляя входной алфавит Y^n в виде

$$Y^n = \left(Y^n \setminus U\right) \cup U,$$

запишем меру Q_γ в виде

---

[1] В данном случае, когда рассматриваются дискретные алфавиты с конечным числом элементов, мерой множества является отношение числа последовательностей, удовлетворяющих свойству, определяющему данное множество, к общему числу возможных последовательностей в алфавите.

$$P(Q_\gamma)=\sum_{U}P\left(\mathbf{y}'\in D/\mathbf{y}\right)p\left(\mathbf{y}\right)+\sum_{Y^n\setminus U}P\left(\mathbf{y}'\in D/\mathbf{y}\right)p(\mathbf{y})\le\sum_{U}\left(1-\alpha\right)p(\mathbf{y})+$$
$$+\sum_{Y^n\setminus U}P\left(\mathbf{y}'\in D/\mathbf{y}\right)p(\mathbf{y})=\left(1-\alpha\right)P(U)+\sum_{Y^n\setminus U}P\left(\mathbf{y}'\in D/\mathbf{y}\right)p(\mathbf{y}).$$

(5.2.15)

Теперь, используя грубую оценку $p(\mathbf{y}'\in D/\mathbf{y})\le 1$, второе слагаемое можно оценить как

$$\sum_{Y^n\setminus U}P\left(\mathbf{y}'\in D/\mathbf{y}\right)p(\mathbf{y})\le\sum_{Y^n\setminus U}p(\mathbf{y})=1-P(U).\qquad(5.2.16)$$

Таким образом, для меры Q_γ оказывается справедлива оценка

$$P(Q_\gamma)\le\left(1-\alpha\right)P(U)+1-P(U)=1-\alpha P(U)\le 1-p_{\max}P(U),$$

противоречащая исходному предположению (5.2.12). Отсюда следует вывод о том, что $M\ge 1$, и существует хотя бы одно кодовое слово, удовлетворяющее условиям (5.2.9)–(5.2.11).

Далее, поскольку код максимальный, не существует никаких других последовательностей из U, которые можно было бы добавить в код, не нарушив при этом ни одно из вышеупомянутых условий. Иными словами, использование любой последовательности $\mathbf{y}'$ из множества D кроме тех, что уже разнесены по решающим областям, приводит к ошибке декодирования, превышающей заданное значение α, т. е.

$$P\left(\mathbf{y}'\in\left[D\setminus\bigcup_{q=1}^{M}A_q\right]/\mathbf{y}\right)>\alpha$$

или, что эквивалентно,

$$P\left(\mathbf{y}'\in\left[D\setminus\bigcup_{q=1}^{M}A_q\right]/\mathbf{y}\right)\le 1-\alpha.\qquad(5.2.17)$$

Так как для произвольных событий V_1 и V_2 справедливо неравенство

$$P(V_1\setminus V_2)\ge P(V_1)-P(V_2),$$

оценим левую часть (5.2.17) как

$$P\left(\mathbf{y}'\in\left[D\setminus\bigcup_{q=1}^{M}A_q\right]/\mathbf{y}\right)\le P\left(\mathbf{y}'\in D/\mathbf{y}\right)-P\left(\mathbf{y}'\in\bigcup_{q=1}^{M}A_q/\mathbf{y}\right).$$

Тогда имеем неравенство

$$P\left(\mathbf{y}' \in D / \mathbf{y}\right) \le 1 - \alpha + P\left(\mathbf{y}' \in \bigcup_{q=1}^{M} A_q / \mathbf{y} \right), \qquad (5.2.18)$$

справедливое для всех входных последовательностей из U, поэтому его можно усреднить по этому множеству. Умножая обе части (5.2.18) на $p(\mathbf{y})$ и суммируя по всем $\mathbf{y}$ из U, имеем:

$$\sum_{U} P\left(\mathbf{y}' \in D / \mathbf{y}\right) p(\mathbf{y}) \le \left(1 - \alpha\right) P(U) + \sum_{S} P\left(\mathbf{y}' \in \bigcup_{q=1}^{M} A_q / \mathbf{y} \right) p(\mathbf{y}) =$$

$$= \left(1 - \alpha\right) P(U) + P\left(\mathbf{y}' \in \bigcup_{q=1}^{M} A_q, \mathbf{y} \in U \right).$$

Сумму, стоящую в левой части последнего неравенства, оценим, пользуясь соотношениями (5.2.15) и (5.2.16):

$$\sum_{U} P\left(\mathbf{y}' \in D / \mathbf{y}\right) p(\mathbf{y}) = P(Q_\gamma) - \sum_{Y^n \setminus U} P\left(\mathbf{y}' \in D / \mathbf{y}\right) p(\mathbf{y}) \ge$$

$$\ge P(Q_\gamma) - 1 + P(U). \qquad (5.2.19)$$

Теперь, используя неравенства (5.2.19) и (5.2.18), получим оценку вероятности совместного события, заключающегося в том, что будет выбрана произвольная входная последовательность $\mathbf{y}$, а последовательность $\mathbf{y}'$ на выходе окажется в одной из решающих областей[1]:

$$P\left(\mathbf{y}' \in \bigcup_{q=1}^{M} A_q, \mathbf{y} \in U \right) \ge \alpha P(U) - 1 + P(O_\gamma). \qquad (5.2.20)$$

Стоящее в правой части (5.2.20) выражение уже фигурирует в утверждении неравенства Файнстейна (5.2.12*а*), и для завершения доказательства необходимо связать левую часть (5.2.20) с объёмом кода M. Имеем:

$$P\left(\mathbf{y}' \in \bigcup_{q=1}^{M} A_q, \mathbf{y} \in U \right) \le P(\mathbf{y} \in U) + P\left(\mathbf{y}' \in \bigcup_{q=1}^{M} A_q \right) \le P\left(\mathbf{y}' \in \bigcup_{q=1}^{M} A_q \right),$$

а поскольку решающие области — это непересекающиеся множества,

$$P\left(\mathbf{y}' \in \bigcup_{q=1}^{M} A_q \right) = \sum_{q=1}^{M} P\left(\mathbf{y}' \in A_q\right).$$

[1] Объединение решающих областей не обязательно должно совпадать со всем выходным алфавитом Y^n.

Оценим вероятность того, что выходная последовательность $\mathbf{y}'$ попадёт в решающую область A_q. Вспомним, что решающие области формируются, исходя из того, что взаимная информация между входной и выходной последовательностями больше заданного числа γ:

$$\log \frac{p(\mathbf{y}'/\mathbf{y})}{p(\mathbf{y}')} > \gamma,$$

или, когда используются двоичные логарифмы,

$$p(\mathbf{y}'/\mathbf{y}) > p(\mathbf{y}')2^{n\gamma}.$$

Если рассматривать решающую область A_q и множество $D(\mathbf{y}_q)$ — совокупность всех тех выходных последовательностей $\mathbf{y}'$, что имеют с входной последовательностью $\mathbf{y}_q$ взаимную информацию, бо́льшую, чем $n\gamma$, то нетрудно понять, что A_q является подмножеством множества $D(\mathbf{y}_q)$. Тогда, суммируя последнее неравенство по множеству $D(\mathbf{y}_q)$, получим следующую цепочку неравенств:

$$1 \geq P\left(\mathbf{y}' \in D\left(\mathbf{y}_q\right)/\mathbf{y}\right) > P\left(\mathbf{y}' \in D(\mathbf{y}_q)\right)2^{n\gamma} \geq P\left(\mathbf{y}' \in A_q\right)2^{n\gamma},$$

откуда

$$P\left(\mathbf{y}' \in A_q\right) \leq 2^{-n\gamma}.$$

Возвращаясь к (5.2.20), запишем

$$P\left(\mathbf{y}' \in \bigcup_{q=1}^{M} A_q, \mathbf{y} \in U\right) \leq M2^{-n\gamma} = 2^{nR}2^{-n\gamma},$$

поэтому,

$$2^{n\gamma}\left[\alpha P(U) - 1 + P(Q_\gamma)\right] \leq 2^{nR},$$

а это, с учётом того, что $p_{\max} \leq \alpha$, и доказывает неравенство Файнстейна.

Итак, существует код, для которого максимальная ошибка декодирования удовлетворяет неравенству (5.2.12). Заметим, что значение правой части этого неравенства можно варьировать за счёт выбора γ и распределения входных вероятностей $P(Q_\gamma)$. Для доказательства прямой теоремы кодирования необходимо так подобрать параметры, чтобы при возрастании n оба слагаемых, и $2^{-n(\gamma - R)}$, и $1 - P(Q_\gamma)$ стремились к нулю.

Будем полагать, что в качестве множества U, из которого формируются кодовые слова, рассматривается весь алфавит Y^n входных n-элементных сообщений, тогда независимо от входного распределения мера множества $P(U)$ всегда равна единице.

Согласно формулировке прямой теоремы кодирования, будем считать, что кодовая скорость R всегда меньше пропускной способности канала C, т. е. существует произвольное положительное число δ, такое, что $R = C - \delta$. Тогда выбрав, например, значение $\gamma = C - \delta / 2$, можно обеспечить в неравенстве Файнстейна стремление к нулю слагаемого

$$2^{-n(\gamma - R)} = 2^{-n\delta/2}$$

при неограниченном возрастании n.

Второе слагаемое определяется мерой множества $P(Q_\gamma)$, формирующего п*а*ры ($\mathbf{y}'$, $\mathbf{y}$), для которых взаимная информация больше $n\gamma$. С учетом выбранного значения γ эта вероятность равна

$$P(Q_\gamma) = P\left(I(\mathbf{y}';\mathbf{y}) > n\left(C - \delta / 2\right)\right)$$

или, что эквивалентно,

$$1 - P(Q_\gamma) \equiv P(\overline{Q_\gamma}) = P\left(I(\mathbf{y}';\mathbf{y}) / n \le C - \delta / 2\right) =$$

$$= P\left(I(\mathbf{y}';\mathbf{y}) / n \le I(Y'^n;Y^n) / n - \delta / 2 + C - I(Y'^n;Y^n) / n\right).$$

Обратим ещё раз внимание, что в последнем соотношении $I(\mathbf{y}'; \mathbf{y})$ — взаимная информация между отдельными последовательностями $\mathbf{y}'$ и $\mathbf{y}$, в то время как $I(Y'^n; Y^n)$ — средняя взаимная информация между алфавитами Y'^n и Y^n.

Теперь, если считать, что для каждого n выбирается такое входное распределение $p(\mathbf{y})$, которое максимизирует среднюю взаимную информацию $I(Y'^n; Y^n) / n$ между n-элементными последовательностями, превращая её в пропускную способность канала C, то последнее неравенство принимает следующий вид:

$$1 - P(Q_\gamma) \le P\left(I(\mathbf{y}';\mathbf{y}) / n - I(Y'^n;Y^n) / n \le -\delta / 2\right) =$$

$$= P\left(\left|I(\mathbf{y}';\mathbf{y}) / n - I(Y'^n;Y^n) / n\right| \ge \delta / 2\right). \qquad (5.2.21)$$

Таким образом, вопрос о поведении вероятности ошибки сводится к исследованию правой части неравенства (5.2.21): если вероятность того, что $I(\mathbf{y}'; \mathbf{y}) / n$ отличается от $I(Y'^n; Y^n) / n$ на величину, большую, чем $\delta / 2$, убывает с ростом n, то также убывает с ростом n и максимальная ошибка декодирования.

Наиболее просто эта задача решается для постоянных каналов, для которых $I(Y'^n; Y^n) = nI(Y'; Y)$ и

$$\frac{1}{n} I(\mathbf{y}';\mathbf{y}) = \frac{1}{n} \sum_{k=1}^{n} I\left(y'^{(k)}; y^{(k)}\right),$$

поэтому

$$1 - P(Q_\gamma) \le P\left\{\left|\frac{1}{n}\sum_{k=1}^{n} I\left(y'^{(k)}; y^{(k)}\right) - I(Y';Y)\right| \ge \frac{\delta}{2}\right\}. \qquad (5.2.22)$$

При этом случайные величины $I(y'^{(k)}; y^{(k)})$ независимы, одинаково распределены, имеют математическое ожидание $I(Y'; Y)$ и ограниченную дисперсию, так что в силу закона больших чисел правая часть (5.2.22) стремится к нулю, когда n стремится к бесконечности.

Таким образом, доказана прямая теорема кодирования для постоянного дискретного канала.

Пусть C / v — нормированная пропускная способность дискретного канала и задана кодовая скорость $R < C / v$. Тогда существует код $G(n, R)$, для которого максимальная ошибка декодирования меньше, чем любое наперед заданное (сколь угодно малое) число.

5.3. ВЕРХНЯЯ ГРАНИЦА ВЕРОЯТНОСТИ ОШИБКИ ДЛЯ ДИСКРЕТНЫХ КАНАЛОВ

В разд. 5.1 было показано, что при использовании кода с повторением вероятность ошибочного декодирования двух кодовых слов стремится к нулю экспоненциально с ростом длины кодового слова. Однако в этом случае с передачей одного кодового слова передаётся только один информационный символ, так что вероятность ошибки уменьшается только за счёт скорости передачи информации. В связи с этим правомерна постановка задачи о возможности снижения вероятности ошибки без уменьшения скорости передачи. Основной идеей, позволяющей успешно решить такую задачу, является рассмотрение не какого-то конкретного кода G, а сразу множества кодов $\Xi = \{G_s\}$, содержащего всевозможные n-элементные кодовые слова (допуская, между прочим, и заведомо плохие коды, которые, например, могут состоять из совпадающих кодовых слов). Каждому коду G_s, входящему в множество Ξ, припишем априорную вероятность q_s его выбора, такую, что

$$\sum_{(s)} q_s = 1.$$

Пусть при использовании конкретного кода G_s обеспечивается вероятность ошибки λ_s. Тогда, если удастся доказать, что средняя по всему ансамблю кодов вероятность ошибки λ ограничена заданным значением ε, т. е., если

$$\lambda = \mathbf{E}\left[\lambda_s\right] = \sum_{(s)} q_s \lambda_s \le \varepsilon, \tag{5.3.1}$$

то, тем самым, можно будет утверждать, что среди всех возможных кодов найдётся хотя бы один, для которого вероятность ошибки не превосходит ε. Как это ни парадоксально, но построение оценки для средней по ансамблю вероятности ошибки оказывается существенно проще, чем для какого-нибудь отдельного кода.

Итак, рассмотрим в заданном канале $\{Y^n,\ Y'^n,\ v,\ p(\mathbf{y}' / \mathbf{y})\}$ множество всех кодов длины n и объёма $M = 2^{nR}$, где R — скорость кода. Согласно определению, каждый код G_s представляет собой совокупность M пар кодовых слов, которые, как будем считать, выбираются равновероятно, и соответствующих им решающих областей:

$$G_s = \left\{\mathbf{u}_1^{(s)}, A_1^{(s)};\ \ldots;\ \mathbf{u}_M^{(s)}, A_M^{(s)}\right\},$$

причём для каждого кода обеспечивается вероятность ошибочного декодирования λ_s, равная, согласно (5.1.6),

$$\lambda_s = \frac{1}{M} \sum_{q=1}^{M} \sum_{\mathbf{y}' \in \overline{A_q^{(s)}}} p\left(\mathbf{y}' / \mathbf{u}_q^{(s)}\right),$$

где $\overline{A_q^{(s)}}$ – дополнение множества $A_q^{(s)}$.

Для дальнейших вычислений введём *индикатор* $\chi_q(\mathbf{y}')$ *ошибки декодирования* при передаче кодового слова $\mathbf{u}_q$:

$$\chi_q(\mathbf{y}') = \begin{cases} 1, & \mathbf{y}' \in \overline{A_q} \\ 0, & \mathbf{y}' \in A_q \end{cases}, \ q = 1, \ldots, M. \tag{5.3.2}$$

Функция $\chi_q(\mathbf{y}')$ оценивается сверху следующим образом[1]:

[1] Объяснить, почему именно используется такая оценка, довольно трудно. Тем не менее, она фигурирует в различной литературе по теории информации, в частности, при рассмотрении границ вероятности ошибки декодирования, и даёт хорошие результаты.

$$\chi_q(\mathbf{y}') \le \left(\frac{\sum\limits_{q \ne i} p\left(\mathbf{y}' / \mathbf{u}_i\right)^{\frac{1}{1+\beta}}}{p\left(\mathbf{y}' / \mathbf{u}_q\right)^{\frac{1}{1+\beta}}} \right)^{\beta}, \ \beta \ge 0. \qquad (5.3.3)$$

Справедливость (5.3.3) устанавливается достаточно просто: если $\chi_q(\mathbf{y}') = 0$, то неравенство тривиально (правая часть и так всегда неотрицательна); если же $\chi_q(\mathbf{y}') = 1$, то для одного из слагаемых в числителе, а именно, для которого индекс q соответствует области A_q, содержащей принятую последовательность $\mathbf{y}'$, выполняется неравенство $p(\mathbf{y}' / \mathbf{u}_q) \ge p(\mathbf{y}' / \mathbf{u}_i)$, следовательно, правая часть не меньше единицы для любого $\beta \ge 0$.

Используем индикатор (5.3.2) и его оценку (5.3.3) для получения верхней границы вероятности ошибки λ_s. Имеем:

$$\lambda_s = \frac{1}{M} \sum_{q=1}^{M} \sum_{Y'^n} p\left(\mathbf{y}' / \mathbf{u}_q^{(s)}\right) \chi_q(\mathbf{y}') \le$$

$$\le \frac{1}{M} \sum_{q=1}^{M} \sum_{Y'^n} p\left(\mathbf{y}' / \mathbf{u}_q^{(s)}\right)^{\frac{1}{1+\beta}} \left[\sum_{q \ne i} p\left(\mathbf{y}' / \mathbf{u}_i^{(s)}\right)^{\frac{1}{1+\beta}} \right]^{\beta}, \qquad (5.3.4)$$

и это неравенство выполняется для любого кода, состоящего из кодовых слов $\mathbf{u}_1, \ldots, \mathbf{u}_M$ и декодируемого по максимуму правдоподобия.

Теперь, оперируя с заданным ансамблем кодов $\Xi = \{G_s\}$, применим *метод случайного кодирования*, означающий следующее. Пусть $\pi(\mathbf{y})$ – некоторое распределение вероятностей на n-элементных входных последовательностях алфавита Y^n, и будем считать, что распределение вероятностей q_s на ансамбле кодов Ξ задаётся посредством распределения $\pi(\mathbf{y})$, т. е.

$$q_s = \prod_{q=1}^{M} \pi\left(\mathbf{u}_q^{(s)}\right).$$

Иными словами, кодовые слова выбираются из алфавита Y^n независимо и в соответствии с распределением $\pi(\mathbf{y})$. Тогда средняя по ансамблю вероятность ошибки λ может быть оценена с применением (5.3.4):

$$\lambda = \frac{1}{M} \sum_{q=1}^{M} \sum_{Y'^n} \mathbf{E} \left[p\left(\mathbf{y}' / \mathbf{u}_q^{(s)}\right)^{1/(1+\beta)} \left(\sum_{q \ne i} p\left(\mathbf{y}' / \mathbf{u}_i^{(s)}\right)^{1/(1+\beta)} \right)^{\beta} \right],$$

где использована перестановочность операций суммирования и взятия математического ожидания.

В последнем соотношении выражение, стоящее под знаком математического ожидания, представляет собой произведение двух случайных величин, первая из которых зависит только от кодового слова $\mathbf{u}_q^{(s)}$, а вторая — от всех остальных кодовых слов. Но при методе случайного кодирования выбор кодовых слов осуществляется независимо, следовательно, сомножители статистически независимы, и

$$\lambda = \frac{1}{M}\sum_{q=1}^{M}\sum_{Y^n}\mathbf{E}\left[p\left(\mathbf{y}'/\mathbf{u}_q^{(s)}\right)^{1/(1+\beta)}\right]\mathbf{E}\left[\left(\sum_{q\neq i}p\left(\mathbf{y}'/\mathbf{u}_i^{(s)}\right)^{1/(1+\beta)}\right)^{\beta}\right]. \quad (5.3.5)$$

Дальнейшая оценка λ основана на свойствах выпуклых функций. Напомним (см. разд. 1.8), что функция $f(\mathbf{v})$ называется выпуклой в некоторой выпуклой области, если для любых неотрицательных чисел $b_1, \ldots, b_S$, сумма которых равна единице, и любых векторов $\mathbf{v}_1, \ldots, \mathbf{v}_S$ из этой области выполняется неравенство

$$\sum_{i=1}^{S} b_i f\left(\mathbf{v}_i\right) \le f\left(\sum_{i=1}^{S} b_i \mathbf{v}_i\right).$$

В решаемой задаче свойство выпуклости будет использовано применительно к скалярным элементам, рассматриваемым на интервалах вещественной числовой оси. Пусть $v_1, \ldots, v_S$ — набор из S элементов, на котором задано распределение вероятностей $b_1, \ldots, b_S$. Тогда левая часть последнего неравенства — это математическое ожидание $\mathbf{E}[f(v)]$ случайной величины $f(v)$, а правая — значение функции f от аргумента, равного математическому ожиданию случайной величины v, т. е. $f(\mathbf{E}[v])$. Таким образом, применение свойства выпуклости к случайным величинам на вещественной оси позволяет записать последнее неравенство в виде[1]

$$\mathbf{E}[f(v)] \le f(\mathbf{E}[v]).$$

В частности, функция $f(v) = v^{\beta}$ — выпуклая на положительной полуоси $v \ge 0$ при $0 \le \beta \le 1$, поэтому можно записать

$$\mathbf{E}[v^{\beta}] \le (\mathbf{E}[v])^{\beta},\ 0 \le \beta \le 1. \quad (5.3.6)$$

Применим неравенство (5.3.6) к дальнейшей оценке средней вероятности ошибки λ, полагая, что

$$v = \sum_{q\neq i} p\left(\mathbf{y}'/\mathbf{u}_i\right)^{1/(1+\beta)}.$$

[1] Это неравенство обычно называют *неравенством Йенсена*.

Имеем:

$$\lambda = \frac{1}{M}\sum_{q=1}^{M}\sum_{Y'^n}\mathbf{E}\left[\, p\left(\mathbf{y}'/\mathbf{u}_q^{(s)}\right)^{1/(1+\beta)}\right]\left(\mathbf{E}\left[\sum_{q\neq i} p\left(\mathbf{y}'/\mathbf{u}_i^{(s)}\right)^{1/(1+\beta)}\right]\right)^{\beta}.$$

Далее заметим, что математическое ожидание

$$\mathbf{E}\left[\, p\left(\mathbf{y}'/\mathbf{u}_q^{(s)}\right)^{1/(1+\beta)}\right] = \sum_{\mathbf{u}_q^{(s)}\in Y^n} p\left(\mathbf{y}'/\mathbf{u}_q^{(s)}\right)^{1/(1+\beta)}\pi\left(\mathbf{u}_q^{(s)}\right)$$

от $\mathbf{u}_q^{(s)}$ не зависит (в данном случае $\mathbf{u}_q^{(s)}$ — это просто "сложный" индекс суммирования) и, следовательно,

$$\mathbf{E}\left[\, p\left(\mathbf{y}'/\mathbf{u}_q^{(s)}\right)^{1/(1+\beta)}\right] = \mathbf{E}\left[\, p\left(\mathbf{y}'/\mathbf{u}_i^{(s)}\right)^{1/(1+\beta)}\right] = \mathbf{E}\left[\, p\left(\mathbf{y}'/\mathbf{y}\right)^{1/(1+\beta)}\right] =$$

$$= \sum_{Y^n} p\left(\mathbf{y}'/\mathbf{y}\right)^{1/(1+\beta)}\pi\left(\mathbf{y}\right).$$

Это позволяет записать

$$\lambda \le \frac{1}{M}\sum_{Y'^n} M\mathbf{E}\left[\, p\left(\mathbf{y}'/\mathbf{y}\right)^{1/(1+\beta)}\right]\left((M-1)\mathbf{E}\left[\, p\left(\mathbf{y}'/\mathbf{y}\right)^{1/(1+\beta)}\right]\right)^{\beta} =$$

$$= (M-1)^{\beta}\sum_{Y'^n}\mathbf{E}\left[\, p\left(\mathbf{y}'/\mathbf{y}\right)^{1/(1+\beta)}\right]^{\beta+1} \le M^{\beta}\sum_{Y'^n}\mathbf{E}\left[\, p\left(\mathbf{y}'/\mathbf{y}\right)^{1/(1+\beta)}\right]^{\beta+1} =$$

$$= M^{\beta}\sum_{Y'^n}\left[\sum_{Y^n} p\left(\mathbf{y}'/\mathbf{y}\right)^{1/(1+\beta)}\pi\left(\mathbf{y}\right)\right]^{\beta+1}. \qquad (5.3.7)$$

Итак, неравенство (5.3.7) представляет собой среднюю по ансамблю всех возможных кодов оценку вероятности ошибочного декодирования, справедливую для произвольного дискретного канала. Посмотрим, каким образом эта оценка может быть конкретизирована для постоянных каналов.

Пусть дискретный постоянный канал задаётся переходными вероятностями

$$p\left(\mathbf{y}'/\mathbf{y}\right) = \prod_{k=1}^{n} p\left(y'^{(k)}/y^{(k)}\right).$$

Кроме того, будем предполагать, что также факторизуется распределение вероятностей $\pi(\mathbf{y})$, задающее ансамбль кодов:

$$\pi(\mathbf{y}) = \prod_{k=1}^{n}\pi\left(y^{(k)}\right), \qquad (5.3.8)$$

т. е. независимыми являются не только кодовые слова в ансамбле Ξ, но также сами кодовые символы. Тогда, подставляя (5.3.8) в (5.3.7), получаем:

$$\lambda \le M^{\beta} \sum_{Y'^n} \left[\sum_{Y^n} \prod_{k=1}^{n} p\left(y'^{(k)} / y^{(k)}\right)^{1/(1+\beta)} \pi\left(y^{(k)}\right) \right]^{\beta+1} =$$

$$= M^{\beta} \sum_{Y'^n} \left[\sum_{Y_1} p\left(y'^{(1)} / y^{(1)}\right)^{1/(1+\beta)} \pi\left(y^{(1)}\right) \cdots \sum_{Y_n} p\left(y'^{(n)} / y^{(n)}\right)^{1/(1+\beta)} \pi\left(y^{(n)}\right) \right]^{\beta+1}.$$

В силу постоянства канала распределения $\pi(y^{(k)})$ от k не зависят (несмотря на формальное наличие верхнего индекса, необходимого для представления вероятности $\pi(\mathbf{y})$ в виде последовательных множителей). Кроме того, вследствие факторизации многомерных распределений суммирование по многомерному выходному алфавиту Y'^n сводится к суммированию по одномерному выходному источнику Y. Поэтому

$$\lambda \le M^{\beta} \left\{ \sum_{Y'} \left[\sum_{Y} p\left(y' / y\right)^{1/(1+\beta)} \pi\left(y\right) \right]^{\beta+1} \right\}^{n}.$$

Введём обозначение

$$E_0\left[\pi(\mathbf{y}), \beta\right] = -\log \sum_{Y'} \left[\sum_{Y} p\left(y' / y\right)^{1/(1+\beta)} \pi\left(y\right) \right]^{\beta+1}. \qquad (5.3.9)$$

Функция $E_0[\pi(\mathbf{y}), \beta]$ называется *функция Галлагера*. Тогда, вспоминая, что $M = 2^{nR}$, оценку вероятности ошибки можно записать в окончательном виде как

$$\lambda \le \exp_2 \left\{ -n\left(-\beta R + E_0\left[\pi(\mathbf{y}), \beta\right] \right) \right\}. \qquad (5.3.10)$$

В литературе оценку (5.3.10) обычно записывают в другом виде. А именно, вводится функция

$$E\left(R, \pi(\mathbf{y}), \beta\right) = \max_{\pi(\mathbf{y}), \beta} \left\{ -\beta R + E_0\left[\pi(\mathbf{y}), \beta\right] \right\}. \qquad (3.3.11)$$

где максимум в правой части разыскивается по всем вероятностным распределениям $\pi(\mathbf{y})$ и по всем значениям β, $0 \le \beta \le 1$. Функция $E[R, \pi(\mathbf{y}), \beta]$ называется *экспонентой случайного кодирования*. В этих обозначениях

$$\lambda \le \exp_2 \left\{ -nE\left[R, \pi(\mathbf{y}), \beta\right] \right\}. \qquad (5.3.12)$$

Итак, рассмотрев множество кодов, имеющих скорость R и содержащих n-элементные кодовые слова, удалось показать, что средняя по всему

ансамблю вероятность ошибочного декодирования λ удовлетворяет соотношению (5.3.12). А на основании (5.3.1) это означает, что в ансамбле кодов $\Xi = \{G_s\}$ существует, по меньшей мере, один код, для которого вероятность ошибочного декодирования λ_s также удовлетворяет (5.3.12).

Покажем теперь, что $E(R) > 0$ для всех скоростей R, меньших пропускной способности канала C. Для этого вычислим частную производную по β функции Галлагера $E_0[\pi(\mathbf{y}), \beta]$ в точке $\beta = 0$. Имеем:

$$\frac{\partial E_0\left[\pi(\mathbf{y}),\beta\right]}{\partial\beta} = -\frac{\log e}{\sum\limits_{Y'}\left[\sum\limits_{Y} p\left(y'/y\right)^{1/(1+\beta)}\pi\left(y\right)\right]^{\beta+1}}\times$$

$$\times\sum_{Y'}\left\{\left[\sum_{Y} p\left(y'/y\right)^{1/(1+\beta)}\pi\left(y\right)\right]^{\beta+1}\ln\left[\sum_{Y} p\left(y'/y\right)^{1/(1+\beta)}\pi\left(y\right)\right]-\right.$$

$$\left.-\frac{\beta+1}{\left(1+\beta\right)^2}\left[\sum_{Y} p\left(y'/y\right)^{1/(1+\beta)}\pi\left(y\right)\right]^{\beta}\sum_{Y} p\left(y'/y\right)^{1/(1+\beta)}\ln p\left(y'/y\right)\pi\left(y\right)\right\},$$

и при $\beta = 0$ производная равна

$$-\log e\sum_{Y'}\left\{\sum_{Y} p\left(y'/y\right)\pi\left(y\right)\ln\sum_{Y} p\left(y'/y\right)\pi\left(y\right)-\right.$$

$$\left.-\sum_{Y} p\left(y'/y\right)\ln p\left(y'/y\right)\pi\left(y\right)\right\}=$$

$$=-\sum_{Y'}\left[p(y')\log p(y')-\sum_{Y} p\left(y'/y\right)\log p\left(y'/y\right)\pi\left(y\right)\right]=$$

$$=\sum_{Y'}\sum_{Y} p\left(y'/y\right)\pi\left(y\right)\log\frac{p\left(y'/y\right)}{p(y')}=I(Y';Y).$$

На основании проведённого анализа можно утверждать, что производная функции Галлагера при $\beta = 0$ равна средней взаимной информации между входным и выходным алфавитами. Тогда для производной экспоненты случайного кодирования $E[R, \pi(\mathbf{y}), \beta]$ можно написать[1]

[1] Предполагается, что в (5.3.11) операции дифференцирования и нахождения максимума переставимы

$$\left.\frac{\partial E\left[R,\pi(\mathbf{y}),\beta\right]}{\partial\beta}\right|_{\beta=0} = I(Y';Y) - R,$$

т. е. производная $E[R, \pi(\mathbf{y}), \beta]$ в точке $\beta = 0$ положительна, если скорость кода удовлетворяет неравенству $R < I(Y'; Y)$. Поскольку функция $E[R, \pi(\mathbf{y}), \beta]$ непрерывна по β, а в точке $\beta = 0$ она обращается в ноль, существуют значения β, $0 < \beta \le 1$, для которых E строго положительна для любых R и $\pi(\mathbf{y})$. В частности, это будет справедливо для такого распределения $\pi(\mathbf{y})$, при котором максимизируется взаимная информация $I(Y'; Y)$, т. е. достигается пропускная способность канала C.

Таким образом, можно сформулировать следующее утверждение, представляющее собой количественное уточнение прямой теоремы кодирования.

При использовании произвольного постоянного дискретного канала существует код $G(n, R)$ с длиной n и скоростью R, для которого средняя вероятность ошибки декодирования удовлетворяет неравенству (5.3.12), *где функция $E[R, \pi(\mathbf{y}), \beta]$ — экспонента случайного кодирования — зависит только от канальной матрицы и скорости кода. При этом $E[R, \pi(\mathbf{y}), \beta] > 0$ для всех скоростей, меньших пропускной способности канала C.*

Далее будут исследоваться свойства функции Галлагера, дающие возможность вычисления или хотя бы приближенной оценки экспоненты случайного кодирования.

Рассмотрим произвольный дискретный постоянный канал с матрицей переходных вероятностей

$$\mathbf{P} = \left\{ p\left(y'_j / y_i\right)\right\},\ y_j \in Y = \left\{y_1, \ldots y_m\right\},\ y'_j \in Y' = \left\{y'_1, \ldots y'_{m'}\right\}.$$

Введём в рассмотрение условные распределения $\mathbf{H} = \{h(y_i / y'_j)\}$, обладающие тем свойством, что $h(y_i / y'_j) = 0$ в том и только в том случае, когда для тех же значений i и j в канальной матрице $\mathbf{P}$ имеет место равенство $p(y'_j / y_i) = 0$. Построим на распределениях $\pi(\mathbf{y})$ и $\mathbf{H}$ функцию

$$D_0\left[\pi(\mathbf{y}), \mathbf{H}, \beta\right] = -\log \sum_{Y'} \sum_{Y} p\left(y' / y\right) h\left(y / y'\right)^{-\beta} \pi\left(y\right)^{1+\beta}, \qquad (5.3.13)$$

похожую по виду на функцию Галлагера (некий аналог “смешанной” функции Галлагера). Попробуем установить аналитическую связь между $E_0[\pi(\mathbf{y}), \beta]$ и $D_0[\pi(\mathbf{y}), \mathbf{H}, \beta]$.

Будем предполагать, что выбор того или иного распределения **H** даёт б*о*льшие или меньшие значения функции $D_0[\pi(\mathbf{y}), \mathbf{H}, \beta]$, и возможно существование некоторого распределения $\mathbf{H}_0$, на котором в (5.3.13) достигается экстремальное (максимальное) значение.

В соответствии с методом неопределённых множителей Лагранжа необходимые условия достижения максимума состоят в решении системы уравнений

$$\frac{\partial D_0\left[\pi(\mathbf{y}), \mathbf{H}, \beta\right]}{\partial h\left(y_i / y'_j\right)} - \mu = 0\,,\ i = 1,\ \ldots,\ m, \tag{5.3.14}$$

где μ — неизвестный множитель, обусловленный наличием ограничения

$$\sum_{i=1}^{m} h\left(y_i / y'_j\right) = 1, \tag{5.3.15}$$

которое и должно быть использовано для нахождения μ.

Частная производная $D_0[\pi(\mathbf{y}), \mathbf{H}, \beta]$ по компонентам **H** равна

$$\frac{\partial D_0\left[\pi(\mathbf{y}), \mathbf{H}, \beta\right]}{\partial h\left(y_i / y'_j\right)} = \log e \frac{\beta p\left(y'_j / y_i\right) h\left(y_i / y'_j\right)^{-(\beta+1)} \pi\left(y_i\right)^{1+\beta}}{\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m'} p\left(y'_j / y_i\right) h\left(y_i / y'_j\right)^{-\beta} \pi\left(y_i\right)^{1+\beta}}.$$

Введём

$$q(y_i, y'_j) = \frac{p\left(y'_j / y_i\right) h\left(y_i / y'_j\right)^{-\beta} \pi\left(y_i\right)^{1+\beta}}{\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{m'} p\left(y'_j / y_i\right) h\left(y_i / y'_j\right)^{-\beta} \pi\left(y_i\right)^{1+\beta}}. \tag{5.3.16}$$

Поскольку каждое из $q(y_i, y'_j) \geq 0$ и сумма

$$\sum_{Y}\sum_{Y'} q(y_i, y'_j) = 1,$$

их можно трактовать как вероятности, причём суммирование по входному или выходному алфавитам даёт соответствующие маргинальные одномерные вероятности

$$q(y) = \sum_{Y'} q(y', y)\,,\ q(y') = \sum_{Y} q(y', y)\,.$$

Тогда уравнения (5.3.14) принимают вид

$$q\left(y_i, y'_j\right) - \frac{\mu h\left(y_i / y'_j\right)}{\beta \log e} = 0\,,\ i = 1,\ \ldots,\ m,$$

а суммирование всех уравнений по входному алфавиту Y с учётом условия нормировки (5.3.15) даёт неизвестный множитель:

$$q(y') = \frac{\mu}{\beta \log e}.$$

Итак, если существует распределение $\mathbf{H}_0$, реализующее максимум функции $D_0[\pi(\mathbf{y}), \mathbf{H}, \beta]$, то его компоненты $h_0(y / y')$ удовлетворяют системе уравнений

$$q(y_i, y'_j) - q(y'_j) h(y_i / y'_j) = 0,\ i = 1, \ldots, m. \qquad (5.3.17)$$

Далее заметим, что фигурирующая в (5.3.13) функция $[h(y / y')]^{-\beta}$ является вогнутой при $\beta \geq 0$, следовательно, стоящее под знаком логарифма выражение является комбинацией вогнутых функций с неотрицательными коэффициентами. Таким образом, в силу монотонности логарифма (и с учётом минуса!) функция $D_0[\pi(\mathbf{y}), \mathbf{H}, \beta]$ есть выпуклая по $h(y / y')$ функция, так что решение (5.3.17) единственно.

Покажем, что таким решением является набор вероятностей

$$h(y_i / y'_j) = \frac{p(y'_j / y_i)^{1/(1+\beta)} \pi(y_i)}{\sum_{i=1}^{m} p(y'_j / y_i)^{1/(1+\beta)} \pi(y_i)},\ i = 1, \ldots, m; j = 1, \ldots, m'. \qquad (5.3.18)$$

Действительно, в этом случае для всех $i = 1, \ldots, m$ и $j = 1, \ldots, m'$ имеем

$$q(y_i, y'_j) = \frac{p(y'_j / y_i)^{1/(1+\beta)} \pi(y_i) \left[\sum_{i=1}^{m} p(y'_j / y_i)^{1/(1+\beta)} \pi(y_i) \right]^{\beta}}{\sum_{j=1}^{m'} \left[\sum_{i=1}^{m} p(y'_j / y_i)^{1/(1+\beta)} \pi(y_i) \right]^{1+\beta}}, \qquad (5.3.19)$$

$$q(y'_j) = \frac{\left[\sum_{i=1}^{m} p(y'_j / y_i)^{1/(1+\beta)} \pi(y_i) \right]^{\beta}}{\sum_{j=1}^{m'} \left[\sum_{i=1}^{m} p(y'_j / y_i)^{1/(1+\beta)} \pi(y_i) \right]^{1+\beta}}, \qquad (5.3.20)$$

и после подстановки этих значений в (5.3.17) имеем тождество.

Кроме того, подставив (5.3.18) в (5.3.13), получим, что функция $D_0[\pi(\mathbf{y}), \mathbf{H}, \beta]$ имеет следующий вид:

$$D_0\left[\pi(\mathbf{y}),\mathbf{H},\beta\right]=-\log\sum_{Y'}\sum_{Y}p\left(y'/y\right)\pi\left(y\right)^{1+\beta}\frac{\pi\left(y\right)^{-\beta}\,p\left(y'/y\right)^{-\beta/(1+\beta)}}{\sum\limits_{Y}p\left(y'/y\right)^{1/(1+\beta)}\pi\left(y\right)}=$$

$$=-\log\sum_{Y'}\left[\sum_{Y}p\left(y'/y\right)^{1/(1+\beta)}\pi\left(y\right)\right]^{1+\beta}.$$

Видно, что выражение, стоящее в правой части последнего неравенства, есть ни что иное, как функция Галлагера (5.3.9). Таким образом, можно утверждать, что максимизация $D_0[\pi(\mathbf{y}),\ \mathbf{H},\ \beta]$ по всем возможным распределениям $\mathbf{H}$ даёт в итоге функцию Галлагера:

$$\max_{\mathbf{H}}D_0\left[\pi(\mathbf{y}),\mathbf{H},\beta\right]=E_0\left[\pi(\mathbf{y}),\beta\right]. \qquad (5.3.21)$$

Обратимся теперь к исследованию наличия возможного экстремума $D_0[\pi(\mathbf{y}),\ \mathbf{H},\ \beta]$ как функции β. Для этого зафиксируем распределение $\mathbf{H}$ и найдём частную производную (5.3.13) по параметру β. На основании (5.3.16) и (5.3.17) имеем:

$$\frac{\partial D_0\left[\pi(\mathbf{y}),\mathbf{H},\beta\right]}{\partial\beta}=\log e\frac{\sum\limits_{Y}\sum\limits_{Y'}p\left(y'/y\right)h\left(y/y'\right)^{-\beta}\pi\left(y\right)^{1+\beta}\ln\dfrac{h\left(y/y'\right)}{\pi\left(y\right)}}{\sum\limits_{Y}\sum\limits_{Y'}p\left(y'/y\right)h\left(y/y'\right)^{-\beta}\pi\left(y\right)^{1+\beta}}=$$

$$=\sum_{Y}\sum_{Y'}q\left(y',y\right)\log\frac{h\left(y/y'\right)}{\pi\left(y\right)}=\sum_{Y}\sum_{Y'}q\left(y',y\right)\log\frac{h\left(y/y'\right)q\left(y\right)}{\pi\left(y\right)q\left(y\right)}=$$

$$=\sum_{Y}\sum_{Y'}q\left(y',y\right)\log\frac{h\left(y/y'\right)}{q\left(y\right)}+\sum_{Y}\sum_{Y'}q\left(y',y\right)\log\frac{q\left(y\right)}{\pi\left(y\right)}=$$

$$=\sum_{Y}\sum_{Y'}q\left(y',y\right)\log\frac{h\left(y/y'\right)}{q\left(y\right)}+\sum_{Y}q\left(y\right)\log\frac{q\left(y\right)}{\pi\left(y\right)}.$$

В последнем выражении

$$I(Y;Y')=\sum_{Y}\sum_{Y'}q\left(y',y\right)\log\frac{h\left(y/y'\right)}{q\left(y\right)}$$

является средней взаимной информацией между входным и выходным алфавитами, когда входное распределение задаётся соотношениями (5.3.16), а величина

$$H\left(q\left(y\right),\pi\left(y\right)\right)=\sum_{Y}q(y)\log\frac{q\left(y\right)}{\pi\left(y\right)}$$

есть просто некоторая взаимная энтропия двух входных распределений $q(y)$ и $\pi(y)$.

Для произвольных распределений p_1 и p_2 функция $H(p_1, p_2)$ неотрицательна и равна нулю только в том случае, когда распределения совпадают (это легко показывается с использованием традиционного неравенства для логарифма $\ln z \le z - 1$).

Поскольку неотрицательными являются и $I(Y';Y)$, и $H(p_1, p_2)$, можно утверждать, что для произвольного $\beta \ge 0$ первая частная производная $D_0[\pi(\mathbf{y}), \mathbf{H}, \beta]$ по β является неотрицательной функцией.

При $\beta = 0$ функция Галлагера равна нулю, кроме того, как это видно из (5.3.16), $q(y', y) = p(y' / y)\pi(y)$, и суммирование по выходному алфавиту Y' даёт, что

$$q(y) \equiv \sum_{Y'} q(y', y) = \sum_{Y'} p(y' / y)\pi(y) = \pi(y).$$

Это, в свою очередь, обнуляет взаимную энтропию $H(p_1, p_2)$.

Итак, первая частная производная $D_0[\pi(\mathbf{y}), \mathbf{H}, \beta]$ по параметру β, равная

$$\frac{\partial D_0\left[\pi(\mathbf{y}), \mathbf{H}, \beta\right]}{\partial \beta} = I\left[Y, Y'\right] + H\left[q(y), \pi(y)\right], \qquad (5.3.22)$$

является неотрицательной функцией, причём

$$\left.\frac{\partial D_0\left[\pi(\mathbf{y}), \mathbf{H}, \beta\right]}{\partial \beta}\right|_{\beta=0} = I(Y; Y'),$$

т. е. касательная к кривой, выходящая из начала координат, численно равна среднему значению взаимной информации.

Найдём теперь вторую производную. Имеем:

$$\frac{\partial^2 D_0\left[\pi(\mathbf{y}), \mathbf{H}, \beta\right]}{\partial \beta^2} = -\log e \left[\frac{\sum_Y \sum_{Y'} p(y' / y) h(y / y')^{-\beta} \pi(y)^{1+\beta} \ln^2 \dfrac{h(y / y')}{\pi(y)}}{\sum_Y \sum_{Y'} p(y' / y) h(y / y')^{-\beta} \pi(y)^{1+\beta}} - \right.$$

$$\left. - \left(\frac{\sum_Y \sum_{Y'} p(y' / y) h(y / y')^{-\beta} \pi(y)^{1+\beta} \ln \dfrac{h(y / y')}{\pi(y)}}{\sum_Y \sum_{Y'} p(y' / y) h(y / y')^{-\beta} \pi(y)^{1+\beta}} \right)^2 \right] =$$

$$= -\frac{1}{\log e}\left[\sum_{Y}\sum_{Y'} q(y',y)\log^2\frac{h(y/y')}{\pi(y)} - \left(\sum_{Y}\sum_{Y'} q(y',y)\log\frac{h(y/y')}{\pi(y)}\right)^2\right] =$$

$$= -\frac{1}{\log e}\sum_{Y}\sum_{Y'} q(y',y)\left[\log\frac{h(y/y')}{\pi(y)} - \sum_{Y}\sum_{Y'} q(y',y)\log\frac{h(y/y')}{\pi(y)}\right]^2 =$$

$$= -\frac{1}{\log e}\sum_{Y}\sum_{Y'} q(y',y)\left[\log\frac{h(y/y')}{\pi(y)} - \frac{\partial D_0[\pi(\mathbf{y}),\mathbf{H},\beta]}{\partial\beta}\right]^2,$$

откуда видно, что для всех $\beta \geq 0$ вторая производная неположительна:

$$\frac{\partial^2 D_0[\pi(\mathbf{y}),\mathbf{H},\beta]}{\partial\beta^2} \leq 0. \qquad (5.3.23)$$

Это, в свою очередь, означает, что первая производная $D_0[\pi(\mathbf{y}), \mathbf{H}, \beta]$ по параметру β является непрерывной невозрастающей функцией β, ограниченной своим максимальным значением, достигаемым при $\beta = 0$. Таким образом, можно записать следующую оценку:

$$0 \leq \frac{\partial D_0[\pi(\mathbf{y}),\mathbf{H},\beta]}{\partial\beta} \leq I(Y;Y').$$

Установленные свойства функций $E_0[\pi(\mathbf{y}), \beta]$ и $D_0[\pi(\mathbf{y}), \mathbf{H}, \beta]$ позволяют глубже понять механизм поведения экспоненты случайного кодирования (5.3.11) и провести её максимизацию по β при заданном входном распределении $\pi(\mathbf{y})$:

$$E(R,\pi(\mathbf{y})) = \max_{0\leq\beta\leq 1}\left\{-\beta R + E_0[\pi(\mathbf{y}),\beta]\right\}.$$

Дифференцируя выражение, стоящее под знаком максимума в (5.3.11), и приравнивая нулю производную, получим уравнение для максимизирующего значения β:

$$\frac{\partial E_0[\pi(\mathbf{y}),\beta]}{\partial\beta} - R = 0. \qquad (5.3.24)$$

На рис. 5.4 показан вид функции Галлагера при заданном распределении $\pi(\mathbf{y})$. При увеличении β от нуля к единице $E_0[\pi(\mathbf{y}), \beta]$ монотонно и выпукло возрастает от нуля. Если в произвольной точке кривой провести касательную с наклоном, равным R, то абсцисса точки касания $\beta = \beta_1$ будет являться корнем уравнения (5.3.24). При этом сама касательная отсекает по оси ординат отрезок, равный $E[R, \pi(\mathbf{y})]$, который будет положительным для

всех значений скорости R, превышающих среднюю взаимную информацию $I(Y'; Y)$. Как только значение R сравняется с $I(Y'; Y)$, $E[R, \pi(\mathbf{y})]$ обнулится, что повлечёт за собой прекращение экспоненциального убывания вероятности ошибки.

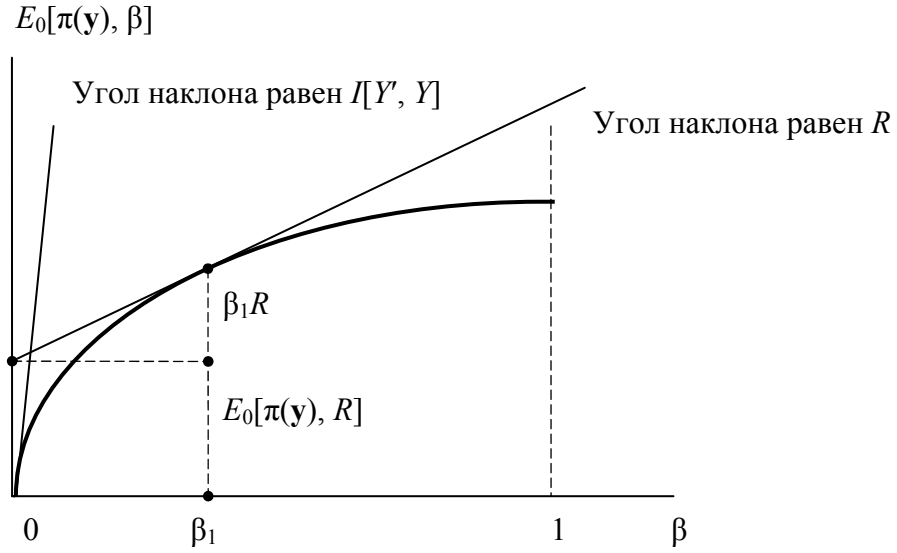


Рис. 5.4. Типичный вид функции Галлагера

Значение

$$R = R_{\text{кр}} = \frac{\partial E_0\left[\pi(\mathbf{y}),\beta\right]}{\partial\beta}\Big|_{\beta=1},$$

т. е. предельное значение угла наклона касательной в точке $\beta = 1$ называется *критической скоростью* кода при заданном распределении $\pi(\mathbf{y})$. Таким образом, соотношения

$$R = \frac{\partial E_0\left[\pi(\mathbf{y}),\beta\right]}{\partial\beta},\ 0 \le \beta \le 1,$$

$$E\left[\pi(\mathbf{y}), R\right] = E_0\left[\pi(\mathbf{y}),\beta\right] - \frac{\partial E_0\left[\pi(\mathbf{y}),\beta\right]}{\partial\beta}\beta$$

задают значения функции $E[R, \pi(\mathbf{y})]$ при скоростях $R_{\text{кр}} \le R \le I(Y';Y)$.

Если теперь выбрать входное распределение $\pi(\mathbf{y})$ таким, чтобы взаимная информация $I(Y',Y)$ оказалась равной значению C пропускной способности канала, то, в полном соответствии с теоремой Шеннона, вероятность ошибочного декодирования будет экспоненциально убывает с ростом n

для всех значений кодовой скорости R, меньших пропускной способности канала.

Итак, показано, что для любого дискретного постоянного канала существует код, для которого вероятность ошибочного декодирования λ удовлетворяет неравенству

$$\lambda \le 2^{-nE(R,\mathbf{P})},$$

где показатель экспоненты зависит только от кодовой скорости R и матрицы $\mathbf{P}$ канальных вероятностей.

Это соотношение было получено посредством рассмотрения ансамбля кодов Ξ, среди которых могут быть такие коды для которых, вероятность ошибки может отличаться от λ как в худшую, так и в лучшую сторону. Например, в ансамбле Ξ могут быть заведомо плохие коды, содержащие совпадающие кодовые слова. С другой стороны, может показаться, что в ансамбле Ξ существуют очень хорошие коды, для которых вероятность ошибки существенно меньше, чем λ. Так, полагая, по-прежнему, что зависимость вероятности ошибки от n экспоненциальная, можно предположить, что существует код $G_s(n, R)$, для которого

$$\lambda_s \le 2^{-nE_1(R,\mathbf{P})},$$

и если E_1 заметно превышает E, то может оказаться, что $\lambda_s \ll \lambda$.

Однако оказывается, что такое оптимистичное предположение не оправдывается, и для достаточно широкого диапазона скоростей функции E_1 и E асимптотически (для достаточно больших n) совпадают, т. е. в ансамбле Ξ не существует кодов, для которых вероятность ошибочного декодирования была бы заметно меньше, чем λ. Одновременно это означает, что при указанных выше условиях все коды в ансамбле Ξ имеют приблизительно одинаковую вероятность ошибки, определяемую экспонентой случайного кодирования (5.3.12). Формально это означает следующее.

Пусть $\lambda_{\min}$ — наименьшая по всем возможным кодам $G(n, R)$ средняя вероятность ошибки декодирования. Тогда справедлива оценка

$$\lambda_{\min} \ge 2^{-n[E^*(R,\mathbf{P})+\varepsilon]}, \qquad (5.3.25)$$

где $E^*(R, \mathbf{P})$ — некоторая функция, зависящая от скорости и переходных канальных вероятностей, а $\varepsilon > 0$ может быть выбрана сколь угодно малой при достаточно больших n. Более того, для всех скоростей $R \ge R_{\text{кр}}$ функции

$E(R, \mathbf{P})$ и $E^*(R, \mathbf{P})$ практически совпадают, т. е. показатели экспонент в (5,3,12) и (5.3.25) сколь угодно близки друг к другу.

Указанная оценка получается на основании изучения свойств нижней границы вероятности ошибки для дискретных каналов, которое оказывается заметно более сложной задачей по сравнению с аналогичной задачей для верхней границы и выходит за рамки данного пособия.

ВОПРОСЫ И ЗАДАНИЯ К ГЛАВЕ 5

5.1. Что представляют собой решающие области при помехоустойчивом кодировании?

5.2. Дайте определение ошибки декодирования.

5.3. Чем различаются стратегии декодирования по критериям максимального правдоподобия и максимума апостериорной вероятности?

5.4. В чем состоит теоретико-информационный смысл неравенств Фано и Файнстейна?

5.5. Сформулируйте прямую и обратную теорему Шеннона для кодирования в дискретном канале с помехами.

5.6. Рассмотрим двоичный симметричный канал с вероятностью ошибки $p = 0{,}1$.

а) Пусть для передачи четырех сообщений используется код, состоящий из следующих кодовых слов:

$$\mathbf{u}_1 = (00000),\ \mathbf{u}_2 = (01111),\ \mathbf{u}_3 = (10101),\ \mathbf{u}_4 = (11010)$$

с априорными вероятностями

$$p(\mathbf{u}_1) = 0{,}8,\ p(\mathbf{u}_2) = 0{,}1,\ p(\mathbf{u}_3) = 0{,}05,\ p(\mathbf{u}_4) = 0{,}05.$$

Указать разбиения на решающие области и выписать элементы этих областей при МАВ-декодировании и МП-декодировании. Найти условные, среднюю и максимальную вероятности ошибки декодирования.

б) Пусть для передачи двух сообщений используется код, состоящий из двух равновероятных кодовых слов $\mathbf{u}_1 = (010)$, $\mathbf{u}_2 = (101)$, и выбраны два варианта разбиения кодовых областей:

вариант 1: $A_1 = \{(010)\ (000)\ (110)\ (011)\}$, $A_2 = \{(101)\ (001)\ (111)\ (100)\}$;

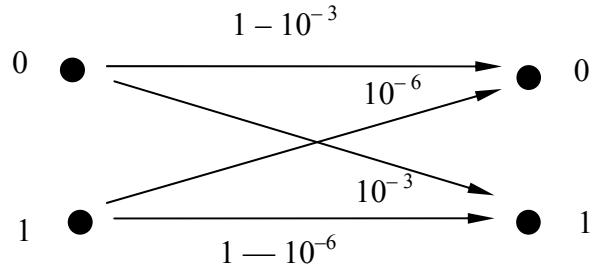
вариант 2: $A_1 = \{(000)\ (100)\ (010)\ (001)\}$, $A_2 = \{(110)\ (101)\ (011)\ (111)\}$.

Найти условные, среднюю и максимальную вероятности ошибки декодирования для обоих вариантов. Какое из двух разбиений решающих областей лучше и почему? Как должны быть выбраны кодовые слова, что-

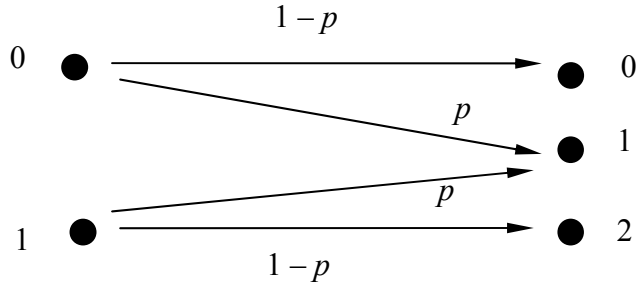
бы худшее разбиение стало лучшим: Зависит ли выбор лучшего разбиения от вероятности p?

5.7. Решить предыдущую задачу для случая использования двоичного несимметричного канала с переходными вероятностями $P(0/1) = 0{,}01$ и $P(1/0) = 0{,}1$.

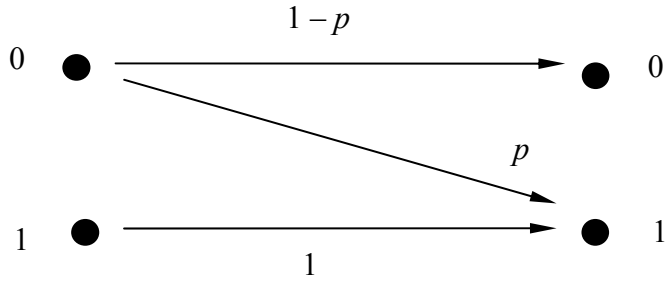
5.8. Найти и построить вид функции $g_k(s)$ (5.1.15) для следующих типов постоянных двоичных каналов:



а). Двоичный асимметричный канал



б). Двоичный стирающий канал



в). Z-канал

Для каждого случая провести минимизацию $g_k(s)$ на интервале [0; 1] и использовать эти результаты для получения верхней границы вероятности

ошибки при МП-декодировании, достижимой при использовании кода с двумя кодовыми словами, представляющими собой n-элементные последовательности нулей и единиц.

5.9. Показать, что для кода, состоящего из двух кодовых слов $\mathbf{u}_1$ и $\mathbf{u}_2$ вероятность ошибки при МП-декодировании не зависит от передаваемого слова и оценивается сверху величиной

$$\sum_{Y^n} \sqrt{p(\mathbf{y}/\mathbf{u}_1)\, p(\mathbf{y}/\mathbf{u}_2)}\,.$$

Далее, используя этот результат, показать, что для постоянного дискретного канала средняя вероятность ошибки декодирования, $\overline{\lambda(\mathbf{u}_1, \mathbf{u}_2)}$ вычисленная по ансамблю кодов, состоящих из двух кодовых слов, удовлетворяет неравенству

$$\overline{\lambda(\mathbf{u}_1, \mathbf{u}_2)} \le \left[\sum_{Y'} \left(\sum_{Y} \pi(y) \sqrt{p(y'/y)} \right)^2 \right]^n,$$

где $\pi(y)$ — априорные вероятности выбора кодовых слов.

5.10. Показать, что при МП-декодировании кода, состоящего из M кодовых слов $\mathbf{u}_1, \ldots, \mathbf{u}_M$, вероятность ошибки $P\{\varepsilon/\mathbf{u}_1\}$ при передаче слова $\mathbf{u}_1$ удовлетворяет неравенству

$$P\{\varepsilon/\mathbf{u}_1\} \le \sum_{q=2}^{M} \lambda(\mathbf{u}_1, \mathbf{u}_2),$$

где $\lambda(\mathbf{u}_1, \mathbf{u}_2)$ — вероятность ошибки для кода, состоящего из двух кодовых слов $\mathbf{u}_1$ и $\mathbf{u}_2$.

5.11. Построить экспоненту случайного кодирования и найти критическую скорость для двоичного симметричного канала, в котором вероятность ошибки равна p.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Шеннон К. Э.* Работы по теории информации и кибернетике / К. Э. Шеннон — М. : ИЛ, 1963.

2. *Фихтенгольц Г. М.* Курс дифференциального и интегрального исчисления / Г. М. Фихтенгольц. — М. : Физматлит, 2001.

3. *Эльсгольц Г. М.* Дифференциальные уравнения и вариационное исчисление / Л. Э. Эльсгольц. — М. : Эдиториал УРСС, 2000.

4. *Ильин В. А.* Линейная алгебра / В. А. Ильин, Э. Г. Позняк — М. : Физматлит, 2002.

5. *Колмогоров А. Н.* Элементы теории функций и функционального анализа / А. Н. Колмогоров, С. В. Фомин. — М. : Наука, 1972.

6. *Химмельблау Д.* Прикладное нелинейное программирование / Д. Химмельблау. – М.: Мир, 1975.

7. *Колесник В. Д.* Курс теории информации / В. Д. Колесник, Г. Ш. Полтырев. — М. : Наука, 1982.

8. *Вентцель А. Д.* Курс теории случайных процессов. / А. Д. Вентцель. — М. : Наука, 1975.

9. *Феллер В.* Введение в теорию вероятностей и её приложения. В 2-х томах / В. Феллер. — М. : Мир, 1984.

10. *Краснов М. Л.* Интегральные уравнения / М. Л. Краснов, А. И. Киселёв, Г. И. Макаренко. — М. : Эдиториал УРСС, 2003.

11. *Левин Б. Р.* Теоретические основы статистической радиотехники / Б. Р. Левин. – М. : Радио и связь, 1990.

12. *Турин В. Я.* Передача информации по каналам с памятью / В. Я. Турин. — М. : Связь, 1977.

Гельгор Александр Леонидович
Попов Евгений Александрович

ТЕОРЕТИКО-ИНФОРМАЦИОННЫЕ ОСНОВЫ ТЕЛЕКОММУНИКАЦИОННЫХ СИСТЕМ

Учебное пособие

Лицензия ЛР № 020593 от 07.08.97
Налоговая льгота – общероссийский классификатор продукции ОК 005-93, т. 2;
95 3004 – научная и производственная литература

Подписано в печать .2013 г. Формат 60×84/16. Печать цифровая.
Усл. печ. л. 18,0. Уч.-изд. л. 8,4. Тираж . Заказ

Отпечатано с готового оригинал-макета, предоставленного Корпоративным центром качества СПбГПУ,
в Цифровом типографском центре Издательства Политехнического университета.
195251, Санкт-Петербург, Политехническая ул., 29.
Тел.: (812) 550-40-14
Тел./факс: (812) 297-57-76