Ф.П. Тарасенко

ВВЕДЕНИЕ В КУРС ТЕОРИИ ИНФОРМАЦИИ

ИЗДАТЕЛЬСТВО ТОМСКОГО УНИВЕРСИТЕТА

ТОМСК-1963

Ф. П. ТАРАСЕНКО

ВВЕДЕНИЕ В КУРС ТЕОРИИ ИНФОРМАЦИИ

ИЗДАТЕЛЬСТВО ТОМСКОГО УНИВЕРСИТЕТА
Томск — 1963

Редактор проф. В. Н. Кессених

ПРЕДИСЛОВИЕ

Опыт чтения в течение нескольких лет курса теории информации на радиофизическом факультете Томского государственного университета им. В. В. Куйбышева убедил автора в целесообразности написания еще одной книги по теории информации, несмотря на наличие нескольких обобщающих монографий на эту тему. Пользование разнообразной и очень многочисленной периодической научной литературой по теории информации и ее приложениям требует хорошего знакомства с основами теории. Характер такой потребности зависит от того, в какой области работает данный специалист и на каком уровне он желает ознакомиться с теорией. Знание классической работы К. Шэннона и У. Уивера [7] необходимо каждому занимающемуся теорией информации, но вряд ли целесообразно рекомендовать эту работу для первоначального знакомства с теорией. Цель такого введения в современную теорию информации преследуется рядом обобщающих работ. Запросы математиков в значительной мере удовлетворяются работами А. Я. Хинчина [47, 48] и монографией А. Фейнстейна [81]. Инженеры-практики могут освоить понятия и методы теории информации по книгам А. А. Харкевича [79] и С. Гольдмана [66]. Потребностям широкого круга лиц, не имеющих специальной подготовки, отвечает книга А. М. Яглома и И. М. Яглома [63]. Но имеется также категория специалистов, которая, с одной стороны, тяготеет к прикладным вопросам, а с другой, — не чуждается и математических тонкостей, правда, не на таком уровне строгости, как чистые математики. К этой категории, в частности, относятся физики и радиофизики соответствующих специальностей университетского профиля, инженеры-разработчики и т. д. Автор полагает, что данная книга окажется полезной прежде всего именно таким специалистам. Такая установка определила характер книги: изложение ведется в основном на физическом уровне строгости,

но приводятся также те строгости доказательства, которые не требуют знания тонких математических фактов.

Построение книги имеет некоторые особенности. Во-первых, большое внимание уделено теории структуры сигналов, которая обычно считается чисто вспомогательным материалом. Во-вторых, хотя в современных исследованиях по теории информации центральное место занимает теория кодирования, в данной книге вопросы кодирования затрагиваются лишь в той мере, в какой это понадобилось для изложения соответствующих теорем Шэннона. Эти два момента могут послужить основанием для критики книги за отставание от современного уровня науки. Тем не менее, автор сознательно принял именно такое построение книги, рассчитанной на тех, кто впервые знакомится с проблематикой теории информации. Учебный характер книги и определил то, что главное внимание уделяется не столько современным результатам теории, сколько достаточно детальному изложению ее основ. Кроме того, теория кодирования сейчас уже развилась настолько, что заслуживает отдельного изложения в учебной и монографической литературе.

Автор считает приятным долгом выразить глубокую признательность члену-корреспонденту АН СССР А. А. Харкевичу и сотрудникам возглавляемого им Института Проблем Передачи Информации АН СССР М. С. Пинскеру, И. А. Овсеевичу, Э. Л. Блоху, Д. С. Лебедеву за ценные замечания, во многом способствовавшие улучшению книги. Автор выражает благодарность Р. Л. Добрушину за критическое обсуждение одного из вариантов рукописи, работникам кафедры электронной вычислительной техники и автоматики ТГУ и проблемной лаборатории счетно-решающих устройств ТГУ за дискуссии по отдельным разделам книги, а также проф. В. Н. Кессениху, впервые подавшему мысль написать эту книгу и взявшему на себя труд по ее редактированию.

Автор отдает себе отчет в наличии недостатков у данной книги и с благодарностью примет критику в свой адрес.

ВВЕДЕНИЕ

Теория информации первоначально возникла как теория, отвечающая непосредственно нуждам техники связи. Однако она быстро получила широкое признание со стороны самых разнообразных отраслей науки и техники, продемонстрировав целесообразность применения новых понятий и методов в различных областях: даже в тех случаях, когда применение теории информации не давало неизвестных ранее результатов, оно способствовало более глубокому пониманию установленных фактов и побуждало к дальнейшим исследованиям. Сейчас уже можно считать общепризнанным, что теория информации является одной из фундаментальных теорий, имеющих общее значение. Этому в значительной мере способствовало широкое обсуждение философских и физических аспектов теории информации (см., например, [51, 53, 54]), а также бурное развитие кибернетики, которая, по определению А. Н. Колмогорова, «занимается изучением систем любой природы, способных воспринимать, хранить и перерабатывать информацию и использовать ее для управления и регулирования». Теория информации является составной частью кибернетики.

Необходимо отметить (и это особенно важно иметь в виду начинающим изучение теории), что современная теория информации лишь на первый взгляд кажется устоявшейся и уже достигшей завершенности. Несмотря на наличие результатов, носящих законченный характер (теорема Котельникова, свойство асимптотической равнораспределенности, теоремы Шеннона, и пр.), теория в целом находится в состоянии развития. Это проявляется уже при обсуждении предмета теории информации. Можно было бы сказать, что предметом теории информации является изучение процессов передачи, хранения, преобразования и ис-

пользования информации. Однако такое определение далеко опережает действительное состояние теории и характеризует лишь конечную цель ее развития. Положение таково, что хорошо изученными (в рамках определенных ограничений) могут считаться процессы передачи и хранения информации; из процессов преобразования информации рассмотрены лишь наиболее простые (перекодирование, линейное предсказание, квантование и т. п.); вопросы использования информации остаются пока практически вне рамок теории информации. Одной из причин этого является отсутствие учета качественных особенностей информации: осмысленности, степени истинности, ценности и т. п. Иногда считают, что отказ от учета качества информации — это неизбежная цена за возможность изучения технических систем связи. Пожалуй, более верным будет считать, что это просто ограничение (по-видимому, временное), в силу которого пока поддаются рассмотрению лишь те свойства информационных систем, которые не связаны с качеством информации и которые определяются только степенью соответствия сигналов в разных частях систем. Поэтому предмет теории информации в ее современном виде можно определить как количественное описание соответствия между случайными объектами (процессами) произвольной природы.

Датой возникновения теории информации считаются 1947—1948 гг., когда К. Шэннон выполнил работу [7]; в этой работе теория информации сразу представлена как стройное, логическое построение. Однако ранее рядом исследователей независимо от Шэннона были получены результаты, органически вошедшие в теорию информации: Хартли (1928 г.), Купфмюллер (1924 г.), Котельников (1933 г.), Винер (1948 г.), Колмогоров (1941 г.), Таллер (1948 г.) и др. После 1948 г. теория информации стала объектом интенсивных исследований многих ученых. Наиболее заметный вклад в развитие теории внесли Шэннон, Макмиллан, Хинчин, Колмогоров, Фейнстейн, Габор, Фэно, Вудворд, Харкевич, Сифоров, Пинскер, Добрушин и др. Большое число работ других авторов по ряду частных вопросов также составляет ценный фонд новой теории.

Все проблемы, которыми занимается теория информации, более или менее отчетливо разделяются на несколько разделов. Совокупность вопросов, связанных со строением сигналов, образует первый раздел — теорию структуры сигналов. Проблемы второго раздела состоят в обосновании и детальном изучении основных понятий теории — энтропии и количества информации. В третий раздел входят вопросы описания и изучения основных свойств информационных систем. Четвертый раздел — теория построения оптимальных

и близких к оптимальным кодов. Многочисленные и разнообразные приложения можно отнести к пятой группе проблем теории информации. В данной книге освещены проблемы первых трех разделов; теория кодирования не рассматривается по мотивам, изложенным в предисловии; что касается прикладных результатов, то они приводятся лишь в той мере, в какой это необходимо для пояснения основ теории.

Часть I
ТЕОРИЯ СТРУКТУРЫ СИГНАЛОВ

ГЛАВА I

ОБЩИЕ ВОПРОСЫ ТЕОРИИ СТРУКТУРЫ СИГНАЛОВ

§ 1. ВВЕДЕНИЕ. ПРЕДМЕТ И СОДЕРЖАНИЕ ТЕОРИИ СТРУКТУРЫ СИГНАЛОВ

Теория структуры сигналов является разделом теории информации. Особенность этого раздела состоит в том, что здесь делается попытка отвлечься (насколько это оказывается возможным) от назначения сигналов и сосредоточить внимание на вопросе о том, *что такое сигнал, каковы его свойства* и как их описать математически. Основные проблемы теории структуры сигналов можно сформулировать следующим образом.

1. Анализ и уточнение понятия сигнала; разработка общего определения сигнала.
2. Вопросы классификации сигналов.
3. Разработка математических моделей сигналов различных типов.
4. Рассмотрение структурных (т. е. связанных со строением) свойств сигналов и описание этих свойств на языке предложенных математических моделей.
5. Обсуждение степени соответствия между разрабатываемой моделью сигналов и реальными сигналами.

§ 2. ПОНЯТИЕ СИГНАЛА. ОПРЕДЕЛЕНИЕ СИГНАЛА [2]

Сигналы самых разнообразных типов широко используются в повседневной жизни, поэтому уже интуитивное понятие сигнала имеет довольно определенное содержание. Тем не менее, это понятие стоит рассмотреть более подробно и дать определения, характеризующие сигнал с разных позиций и охватывающие все типы сигналов.

Характеризуя сигнал с точки зрения его функционального назначения, т. е. отвечая на вопрос, «для чего служит сигнал?», — можно дать следующее определение. Сигнал является отображением сообщения; сигнал есть материальный носитель информации. Каков бы ни был любой конкретный сигнал — звуковой, световой или радиосигнал, книга, грампластинка или кинофильм — весь смысл создания этого сигнала заключен в отображении определенной информации. В конечном счете, всякая информация, а следовательно, и всякий сигнал адресуются к получателю и представляют какую-то ценность только при наличии (или возможном наличии) получателя. Отправитель и получатель всегда разделены пространством или временем; сигналы обеспечивают общение между ними. Отсюда следует дополнение (или, скорее, пояснение) к данному выше определению: сигнал есть средство перенесения информации в пространстве и времени.

Данные выше определения не могут служить основой для теории структуры сигналов, так как они рассматривают сигнал с его служебной стороны и не связаны со строением сигнала. Бесконечное разнообразие сигналов, эквивалентность физически совершенно различных представлений одного сообщения — все это требует дать определение, отвечающее на вопрос: «Что такое сигнал?», т. е. определение, рассматривающее сигнал с точки зрения лица, интересующегося сигналом не как вспомогательным средством, а как объектом исследования.

Рассмотрение любых ситуаций, в которых участвуют сигналы, приводит к выводу о том, что хотя сигнал всегда связан с материальным объектом, большинство конкретных (физических, химических и пр.) свойств этого объекта несущественно. В конечном счете неважно, на какого сорта бумаге и какого состава чернилами написано данное письмо; от всех других писем оно отличается как сигнал состоянием распределения цвета по поверхности листа.

При осуществлении радиопередачи для отображения сообщения используется целый ряд физически различных объектов: машинописный текст передачи — голос диктора — электромагнитные волны — колебания тока в обмотке электромагнита — звук громкоговорителя — колебания барабанной перепонки слушателя — колебательные процессы в слуховом нерве слушателя. В качестве звеньев этой цепи можно включить запись и воспроизведение звука на магнитофоне и т. д. Общее, что связывает такое многообразие объектов, заключается в том, что все они служат для образования сигналов. В известном смысле можно сказать, что эти объекты сами «служат в качестве сигналов», однако более существен-

но то, что один и тот же объект (например, электромагнитное поле) может нести разные сигналы. Следовательно, в качестве сигналов используются не сами по себе объекты, а их состояния. Образование сигнала заключается в изменении состояния объекта. Это утверждение требует развития, так как, очевидно, обратное неверно: не всякое изменение состояния объекта является сигналом. Воздействие на объект, изменяющее его состояние, только тогда приведет к образованию сигнала, когда это воздействие производится по определенным правилам*). Наличие таких правил обеспечивает соответствие между сообщением и сигналом. Существование этого соответствия, в свою очередь, обеспечивает возможность извлечения сообщения из полученного сигнала. Эта возможность может быть реализована только в том случае, если правила изменения состояния объекта (т. е. правила образования сигнала) известны стороне, получившей сигнал, или известны частично, по крайней мере, до такой степени, чтобы, опираясь на эти частичные сведения и анализ сигнала, полностью определить эти правила. Теперь мы можем дать уточненное определение: сигнал есть изменение состояния материального объекта, произведенное по заранее определенным правилам (т. е. с помощью заранее определенного кода).

На первый взгляд может показаться, что это определение не охватывает сигналы, которые создаются случайным образом, т. е. так называемые «шумы» и «помехи». На самом деле это не так. Чтобы случайно созданное состояние объекта воспринималось как сигнал, необходимо, чтобы комплекс условий его создания (код помехи) перекрывался с комплексом условий создания (кодом) полезного сигнала**). Чем более близок код помехи к коду полезного сигнала, тем труднее различить их и тем более эффективна помеха. Что касается естественных шумов, возникающих как результат дискретности строения материи, то они тоже воспринимаются как сигналы лишь в той мере, в какой их характеристики совпадают с характеристиками полезного сигнала.

В заключение отметим, что к числу неотъемлемых, но информационно несущественных свойств динамического сигнала относятся и его энергетические характеристики. Существенно не наличие энергии, а изменение ее поступления.

---

*) В терминологии теории связи комплекс правил образования сигнала называется кодом.

**) Например, в случае создания помехи радиоприему необходимо, по крайней мере, чтобы полоса частот помехи перекрывалась с полосой частот полезного сигнала.

Это особенно наглядно видно на примере счета деталей на конвейере с помощью фотоэлемента, когда сигналом о прохождении детали служит прекращение фототока, прекращение поступления энергии.

§ 3. ДВА ОСНОВНЫХ КЛАССА СИГНАЛОВ

Поскольку сигналы служат для переноса информации в пространстве и времени, то для образования сигналов могут использоваться только такие объекты, состояния которых обладают достаточной устойчивостью по отношению к изменению времени или положения в пространстве. Количественные требования к устойчивости предъявляются в соответствии с конкретными условиями использования сигнала.

С точки зрения устойчивости все сигналы можно разделить на два класса. К первому классу относятся сигналы, в качестве которых используются устойчивые, стабильные состояния физических систем. Примерами сигналов такого типа могут служить: книга, фотографическое изображение, состояние пленки магнитофона, состояние ферритовой матрицы памяти электронной вычислительной машины, состояние регистра (системы триггеров) вычислительной машины, положение штанги железнодорожного семафора, расположение триангуляционной вышки и т. д. и т. п. Такие сигналы назовем статическими сигналами.

Во втором классе объединяются сигналы, в качестве которых используются динамические состояния силовых полей. Как было указано в предыдущем параграфе, сигнал возникает при изменении состояния объекта. В отличие от других материальных систем, поля характеризуются тем, что изменение их состояния не может быть локализовано в (неизолированной) части поля и приводит к распространению возмущения. При распространении возмущения в поле параметры конфигурации, строения этого возмущения обладают известной устойчивостью, что и позволяет использовать такие состояния поля в качестве сигналов. Примерами таких сигналов могут служить: звуковые сигналы (изменение состояния поля сил упругости в газе, жидкости или твердом теле), световые и радиосигналы (изменения состояния электромагнитного поля). Назовем сигналы второго класса динамическими сигналами.

В силу характерного различия динамических и статических сигналов их практическое использование тоже различно. Динамические сигналы используются преимущественно для передачи, а статические — для хранения информации. Однако эти функции нельзя полностью разделить. Динамические сигналы могут использоваться для хранения информа-

ции, как это имеет место, например, в запоминающих устройствах на ультразвуковых линиях задержки электронных цифровых вычислительных машин. В известном смысле можно сказать, что такие статические сигналы, как газеты и письма, в большей степени предназначены для передачи, чем для хранения информации.

§ 4. СТРУКТУРНЫЕ СВОЙСТВА СИГНАЛОВ. ПАРАМЕТРЫ СИГНАЛОВ

Одной из основных проблем теории структуры сигналов является достаточно полное описание сигналов. Структура сигнала есть структура соответствующих состояний объекта, используемых для отображения сообщений. Два момента выступают на первый план: 1) как полностью задать (определить) структуру конкретного сигнала (состояния объекта); 2) какое максимальное число различных сигналов можно отобразить с помощью данного объекта, т. е. каково число структурно различимых состояний объекта.

Для описания сигналов естественно воспользоваться разработанной в физике методикой описания состояний объектов (или систем объектов). С этой целью из бесконечного множества свойств объекта отбирается некоторое конечное число свойств, играющих существенную роль в рассматриваемой ситуации. Каждое из таких свойств характеризуется некоторыми параметрами; различие состояний объекта сводится, таким образом, к различию параметров объекта.

Рассмотрим теперь характерные особенности параметров сигнала и роль различных параметров сигнала в свойствах сигнала как носителя информации.

Необходимость выяснения числа различимых сигналов приводит нас к тому, чтобы ввести классификацию сигналов по мощности множества различных состояний объекта. Если множество значений, которые может принимать некоторый параметр сигнала, конечно или счетно, то сигнал называется дискретным по данному параметру. Если множество возможных значений параметра образует континуум, то сигнал называется непрерывным по данному параметру. В качестве примера непрерывных по амплитуде сигналов можно привести телефонный сигнал, сигналы в аналоговых вычислительных устройствах; с дискретными по амплитуде и времени сигналами мы встречаемся в цифровой вычислительной технике, телеграфии и пр.

Сигналы обычно характеризуются не одним, а несколькими параметрами. При этом возможны случаи, когда сигнал по одним параметрам является дискретным, по другим — непрерывным. Известны, например, системы, получающие данные

путем периодических кратковременных («мгновенных») измерений некоторой величины. Эти данные могут быть представлены сигналом, дискретным по времени; по значению измеряемой величины такой сигнал может быть непрерывным.

В некоторых случаях оказывается необходимым перейти от непрерывных сигналов к дискретным или наоборот. Например, данные для цифровой вычислительной машины необходимо представлять в дискретной форме, даже если первоначально они заданы непрерывным сигналом. Операция преобразования непрерывного сигнала в дискретный называется дискретизацией, или квантованием. С другой стороны, встречаются ситуации, когда по заданному дискретному сигналу необходимо построить соответствующий непрерывный сигнал. Такая ситуация, например, возникает, когда по нескольким экспериментальным точкам требуется построить непрерывную кривую.

Различные параметры объекта играют различную роль при использовании его в качестве сигнала. В связи с этим оказывается возможным классифицировать параметры сигнала по трем основным группам. В первую группу относятся те параметры, которые связаны с числом степеней свободы сигнала. Под числом степеней свободы сигнала понимается минимальное число координат, задание которых полностью определяет сигнал. Параметры, изменение которых влияет на число степеней свободы сигнала, будем называть структурными параметрами. Для динамических сигналов таким параметром является время (общая длительность сигнала); для статических сигналов структурными параметрами служат пространственные координаты.

Удобно, далее, ввести понятие параметров отбора (параметров различия). С помощью этих параметров осуществляется выбор из множества имеющихся в наличии сигналов тех, которые принадлежат интересующему нас источнику. К таким параметрам относятся: несущая частота для сигналов радиовещания, кодовые интервалы ключевой группы импульсов в командной системе управления, частота повторения при синхронном накоплении периодических импульсов, и т. п. Характерно, что параметры отбора обычно содержат только информацию о принадлежности данного сигнала к некоторому ансамблю.

Наконец, всякий сигнал имеет параметры, изменение которых осуществляется с целью отображения нужной информации. В радиовещательном сигнале таким параметром является амплитуда, в телеграфном сигнале — длительность элементарной посылки, в печатном тексте — форма символа, и т. п. Такие параметры сигнала целесообразно назвать информативными параметрами.

Следует иметь в виду, что рассмотренная выше классификация параметров сигнала (как, впрочем, и всякая классификация) не может быть абсолютной. Можно найти примеры, когда один и тот же параметр в разных случаях относится к различным типам, либо выполняет одновременно функции параметров разных типов. Например, частота модуляции амплитуды в некоторых командных системах служит одновременно и параметром отбора (сигнал проходит через узкополосный фильтр) и информативным параметром (прошедший через фильтр сигнал приводит в действие исполнительное устройство).

§ 5. ТИПЫ СИГНАЛОВ

Несмотря на огромное разнообразие сигналов, по способу их генерирования и извлечения из них сведений на приемном конце (т. е. по способу кодирования и декодирования) все сигналы разбиваются на три большие группы.

К первой группе относятся сигналы, которые можно назвать сигналами связи, или [1] прямыми сигналами. К числу таких сигналов относятся сигналы, используемые в телеграфе, телефоне, телевидении, телеуправлении, акустической и световой связи, письменные и печатные буквенные сигналы и т. п. Характерные особенности этой группы сигналов состоят в том, что, во-первых, всегда налицо отправитель и получатель сигнала и сигнал предназначен для передачи информации от первого ко второму; во-вторых, код полностью известен обеим связующимся сторонам; в-третьих, в той части, которая не затрагивает условий существования сигнала, код является условным, т. е. строится по соглашению связующихся сторон и по соглашению же может быть изменен.

Вторую группу образуют сигналы, с помощью которых производятся измерения. Измерение некоторой величины есть сравнение ее с соответствующим эталоном, поэтому при измерении всегда имеются два сигнала: эталонный и сравниваемый с ним. В некоторых ситуациях (например, в радиолокации) подлежащий сравнению сигнал есть измененный в процессе распространения эталонный («зондирующий» [1]) сигнал; в других случаях (например, при измерении длины линейкой) сравниваемый сигнал существует независимо от эталонного. Особенность эталонного сигнала в том, что о нем все известно, и сам он, следовательно, никакой информации не несет; для удобства сравнения широкого класса сигналов с эталоном последний обычно бывает периодическим, хотя это и не обязательно.

Сам по себе взятый сигнал, подлежащий сравнению, тоже не несет интересующей нас информации. Это видно хотя бы из того, что прием отраженных от самолета радиолокационных импульсов хотя и говорит о наличии самолета, но не дает возможности определить расстояние до него, если развертка приемного индикатора не синхронизируется зондирующими импульсами. Таким образом, измерительная информация заключена лишь в совокупности эталонного и сравниваемого сигналов.

Для измерительной ситуации характерно, что в конечном счете отправитель и получатель сигналов — одно и то же лицо. К числу особенностей измерительных сигналов относится также то, что если исключить некоторый произвол в выборе единиц измерения, в остальном код определяется физическими условиями эксперимента и должен быть известен лицу, производящему измерения.

В третью группу могут быть отнесены так называемые естественные сигналы. В § 2 мы уже отмечали, что сигналы выступают всегда как состояния физических объектов. В определенном смысле можно сказать, что любое состояние любого физического объекта можно рассматривать как сигнал даже в том случае, если приведение этого объекта в данное состояние вовсе не связано с передачей каких-либо сведений, а произошло в силу естественных причин. Можно сказать, что перед нами «сигнал с неполностью известным кодом».

К числу естественных сигналов относятся сигналы, изучаемые астрономией, радиоастрономией, физикой атмосферных разрядов и т. п. Световые сигналы, благодаря которым мы видим окружающий мир, тоже являются естественными (т. е. не генерируемыми с какой-то целью, а просто существующими) сигналами.

Следует указать на некоторую условность приведенной классификации. Можно привести примеры, иллюстрирующие трудности безоговорочного отнесения некоторых сигналов к какой-либо из трех групп. Зашифрованный сигнал, перехваченный противником, подвергается тщательному исследованию его структуры, подобно естественным сигналам. Сигналы, адресуемые к эмоциям человека (например, музыка, живопись и др.) лишь с большими натяжками могут быть отнесены к группе сигналов связи. Наконец, всякая операция кодирования и декодирования есть по существу операция сравнения с эталонами, т. е. своего рода «измерение» сопровождает обработку и сигналов связи.

ГЛАВА II

МАТЕМАТИЧЕСКИЕ МОДЕЛИ СИГНАЛОВ

§ 1. ИЗОМОРФИЗМ СИГНАЛОВ

Для изучения свойств сигналов математическим путем необходимо выдвинуть определенную модель сигнала, которая отражала бы существенные стороны реальных сигналов и, кроме того, допускала бы количественное описание.

Необходимость рассмотрения самых разнообразных видов сигналов привела к выдвижению нескольких математических моделей сигнала. Развитие этих моделей продолжается и в настоящее время; в дальнейшем изложении мы постараемся подчеркнуть, в каких направлениях идет это развитие. В данном параграфе рассмотрению конкретных моделей мы предпошлем математическое описание общего для всех моделей свойства сигналов, а именно — изоморфизма сигналов.

В предыдущей главе было отмечено, что одному и тому же сообщению может быть поставлен в соответствие целый ряд физически различных сигналов. При этом возможно построение длинной цепи сигналов (как в примере с радиопередачей), соответствие элемента которой с сигналом-первоисточником устанавливается не непосредственно, а путем последовательного построения сигнала одной физической природы по полученному сигналу другой природы. Сохранение информации обеспечивается взаимно однозначным соответствием сигналов.

Взаимное соответствие состояний физически разнородных объектов используется во всех случаях построения сигналов. Иногда такое соответствие устанавливается в силу объективных причинных связей, как например, соответствие колебаний тока в цепи микрофона и звуковых колебаний мембраны микрофона, соответствие отлитой матрицы и отпечатанной

с нее газетной страницы, соответствие фотоизображений на негативе и позитиве и т. п. Иногда же такое соответствие носит чисто условный характер и устанавливается в ходе построения кода. Один и тот же звук в русском алфавите обозначается символом «г »,в латинском — символом «g», в греческом — «γ». Другим примером установления общепринятого условного соответствия является математическая символика. Например, с помощью системы правил, входящих в код, установлено однозначное соответствие между геометрическими образами и аналитическими уравнениями, которым подчиняются координаты этих образов.

В теории множеств существует понятие изоморфизма множеств, придающее точный смысл несколько неопределенному понятию соответствия, примеры которого рассмотрены выше. Для простоты рассмотрим случай дискретных множеств.

Два множества X и Y, состоящие из элементов $x \in X$ и $y \in Y$, называются изоморфными, если выполняются следующие условия:

1. Каждый элемент $x_m \in X$ может быть взаимно однозначно сопоставлен с элементом $y_e \in Y$, т. е. $x_m \longrightarrow y_e$ и $y_e \longrightarrow x_m$;

2. Каждая операция f (из некоторого класса операций), преобразующая элемент $x_m \in X$ в $x_n \in X$ в множестве X, $f(x_m) = x_n$, может быть взаимно однозначно сопоставлена с операцией φ, преобразующей элемент $y_\kappa \in Y$ в $y_e \in Y$, $\varphi(y_\kappa) = y_e$, т. е. $f \rightarrow \varphi$, $\varphi \longrightarrow f$;

3. Если $x_m \in X$ соответствует $y_\kappa \in Y$ и $x_n \in X$ соответствует $y_e \in Y$, если $f(x_m) = x_n$ и $f \longrightarrow \varphi$, то для всех x, y, f $\varphi(y_\kappa) = y_e$.

Смысл первых двух условий очевиден; последнее условие обеспечивает удовлетворение того требования, чтобы элементы x_n и y_e, полученные из соответствующих друг другу элементов x_m и y_κ с помощью соответствующих друг другу операций f и φ, тоже соответствовали друг другу*).

В применении к теории сигналов каждое множество, упоминаемое в приведенном определении, представляет собой множество физически однородных сигналов (множество состояний одного физического объекта), элементами множества служат конкретные сигналы (состояния), операции f и φ являются операциями перекодирования. Переход из одного множества в другое эквивалентен переходу от сигнала одной физической природы к соответствующему сигналу другой природы. С этой точки зрения смысловое, или семан-

*) В связи с этим более строгое определение изоморфизма говорит об изоморфности множеств относительно некоторых операций.

тическое, содержание информации, несомой сигналом, исчерпывается изоморфным соответствием между сигналом и событием, отраженным данным сигналом. Здесь мы снова приходим к тому, что было сказано в главе I: если правила кодирования совершенно неизвестны, то изоморфность не может быть установлена, и информация не может быть извлечена из сигнала.

В свете понятия изоморфизма задача построения математической модели сигнала выглядит как задача построения множества математических образов, изоморфного множеству реальных сигналов.

§ 2. СЛУЧАЙНЫЙ ПРОЦЕСС — МОДЕЛЬ СИГНАЛА

Любой сигнал выступает как определенное состояние некоторой физической системы. Следовательно, описывать сигналы — значит описывать состояния этих объектов. Динамические сигналы (т. е. состояния силовых полей) в конечном счете регистрируются как временные процессы, происходящие в некоторой системе; т. е. моделью конкретного динамического сигнала может служить некоторая функция времени $f(t)$. Статические сигналы, наоборот, тем лучше выполняют свою роль, чем слабее зависимость от времени соответствующих состояний; т. е. в качестве модели статических сигналов должны рассматриваться абсолютно устойчивые состояния идеальных физических объектов. Можно, однако, усмотреть общность в описании как динамических, так и статических сигналов. Возьмем некоторый конкретный одномерный непрерывный статический сигнал, например, звуковую дорожку на киноленте. Если обозначить структурный параметр (в данном случае — расстояние от начала записи) символом t, то конкретная запись может быть выражена также функцией $f(t)$, f — ширина дорожки.

Однако описание сигналов с помощью однозначных функций структурного параметра $f(t)$ не исчерпывает всех существенных особенностей сигналов. Пусть, например, мы решили описывать динамический сигнал однозначной функцией времени $f(t)$. Задать такую функцию значит задать правило, по которому для каждого момента времени t можно найти соответствующее ему значение f. Однако, если такое правило известно на приемном конце, то необходимость передачи отпадает вообще: такой сигнал может быть построен в точке приема без обращения к линии связи. Если же это правило неизвестно, но существует, то имеет больший смысл передавать не $f(t)$, а сигнал-инструкцию, излагающую это правило. Например, вместо того, чтобы бесконечно генерировать синусоиду, разумнее сообщить правило ее по-

строения, что и делают обычно при изложении соответствующего раздела математики в школе.

Таким образом, одна-единственная однозначная функция структурного параметра не может служить моделью сигнала. Такая функция приобретает сигнальные свойства только тогда, когда она является одной из возможных функций. Следовательно, моделью сигнала может служить набор однозначных функций структурного параметра; в качестве конкретного сигнала используется какая-либо одна из этих функций. Такой модели соответствует понятие случайного процеса, определяемого [6, 7] как множество функций параметра t, на котором (множестве) определена вероятностная мера. Каждая конкретная функция называется реализацией случайного процесса.

Понятие случайного процесса является довольно сложной математической абстракцией, поэтому, кроме приведенного выше определения, имеются и другие, характеризующие это понятие с других сторон. Например, Дж. Дуб [4] определяет случайный процесс как произвольное семейство случайных величин $\{F_t, t \in T\}$. Если t — время (динамический сигнал), то $f(t)$ — то, что наблюдается в момент t, T — совокупность рассматриваемых моментов времени. Еще одна конкретизация понятия случайного процесса состоит в том, чтобы определить его как такую функцию времени, значение которой в каждый данный момент является случайной величиной. Тогда [8] полное определение случайной функции $f(t)$ сводится к заданию системы конечномерных совместных распределений вероятностей для n разделенных некоторыми временными интервалами значений $f(t)$. Итак, адекватным представлением сигналов является случайный процесс. Однако на реализации случайного процесса, моделирующего сигнал, должны быть наложены дополнительные ограничения. Эти ограничения вытекают из той особенности реальных информационных систем, что они могут оперировать только с конечным объемом передаваемых данных. Если допустить, что каждый момент времени приносит новые данные, то любой сигнал конечной длительности окажется способным дать бесконечное количество данных. Следовательно, к реализациям случайного процесса, используемого в качестве модели сигнала, должно быть предъявлено следующее требование: любой отрезок реализации конечной длительности должен нести конечное число независимых данных, т. е. иметь конечное число степеней свободы.

Выполнение этого требования еще не полностью решает проблему конечности объема данных. Так как в общем случае точное задание значения функции f в некоторый момент времени t может быть осуществлено только с помощью бес-

конечного ряда цифр, то мы приходим к выдвижению еще одного условия: реализация случайного процесса, моделирующего сигнал, не должна определяться с бесконечной точностью. Это может служить еще одним аргументом в пользу утверждения непригодности функций строгого анализа в качестве модели сигнала [3]. Однако пока не разработан математический аппарат, вполне удовлетворяющий этому требованию. Более или менее удовлетворительно можно обойти эту трудность двумя способами. Первый способ состоит в том, чтобы ввести фиксированные различимые уровни значений функции $f(t)$, и все значения, находящиеся между уровнями, относить к одному из них. В этом состоит принцип квантования, т. е. аппроксимации непрерывной функции времени дискретной функцией. Другой способ исходит из опыта представления экспериментальных данных о физически непрерывной величине. Он состоит в том, чтобы указать приближенное (округленное) значение функции $f(t)$ и указать, в каких пределах около этого значения находится истинное значение. Такое представление тоже можно называть квантованием, однако уровни квантования не фиксируются заранее, а смещаются вместе с истинным значением функции.

Применение указанных способов позволяет при анализе динамических сигналов пользоваться аппаратом функций строгого анализа.

§ 3. МАТЕМАТИЧЕСКИЕ МОДЕЛИ НЕКОТОРЫХ КОНКРЕТНЫХ ТИПОВ СИГНАЛОВ. КЛАССИФИКАЦИЯ СЛУЧАЙНЫХ ПРОЦЕССОВ

При теоретическом рассмотрении некоторых типов сигналов возникает необходимость конкретизировать некоторые характеристики случайного процесса, учтя специфические особенности моделируемого сигнала. Таким образом, все случайные процессы разбиваются на классы, каждый из которых требует особого описания и несколько отличных методов при их теоретическом изучении. Существует несколько признаков, по которым может производиться классификация случайных процессов. Кратко охарактеризуем основные классы случайных процессов.

Первым признаком классификации могут служить временные характеристики процесса. Прежде всего различаются *непрерывные* и *дискретные* процессы. Случайный процесс называется процессом с непрерывным временем, если его реализации определяются для всех моментов из некоторого (конечного или бесконечного) интервала T параметра t. Существуют ситуации, в которых достаточно определять реализацию случайного процесса лишь на малых

интервалах времени, разделенных сравнительно большими интервалами. Хотя физически сколь угодно малым каждый интервал определения реализации сделать нельзя, полезной математической абстракцией является понятие дискретного случайного процесса*), реализация которого определяется на дискретном ряде точек временной оси.

И непрерывные и дискретные случайные процессы по степени устойчивости статистических характеристик реализаций во времени разделяются на два класса: стационарные и нестационарные процессы. Случайный процесс называется стационарным в узком смысле, если для любого n конечномерные распределения вероятностей не изменяются со временем, т. е. выполняется условие.

$$p_n(x_0, t_0+\tau; x_1, t_1+\tau;\ldots; x_{n-1}, t_{n-1}+\tau) = p_n(x_0, t_0; x_1, t_1;\ldots; x_{n-1}, t_{n-1}). \quad (4.1)$$

Отсюда следует, что для стационарного в узком смысле процесса имеет место следующий ряд равенств:

$$p_1(x_0, t_0) = p_1(x_0),$$
$$p_2(x_0, t_0; x_1, t_1) = p_2(x_0, x_1; t_1 - t_0), \quad (4.2)$$

и т. д. В частности, P_n зависит только от $n-1$ разностей моментов времени.

На устойчивость статистических характеристик можно наложить более мягкие требования. В частности, случайный процесс называется стационарным в широком смысле (или в смысле Хинчина), если требования независимости от времени наложить только на первый и второй моменты процесса. Стационарность случайного процесса может характеризоваться и с помощью моментов, вычисляемых по конечным отрезкам реализации процесса (так называемых временных моментов). Например, в [5] предлагается требование стационарности определить как независимость от времени распределений вероятностей первых двух временных моментов. Это еще более расширяет класс стационарных процессов.

Следующим признаком классификации является признак эргодичности. Обрабатывая результаты наблюдения над одной из реализаций случайного процесса, можно получить количественные характеристики статистических свойств этой реализации: моменты, функции корреляции, распределения

*) Дискретные случайные процессы иногда называются случайными последовательностями, так как само слово «процесс» имеет оттенок непрерывности.

любых порядков. Случайный процесс называется эргодическим, если статистические характеристики одной реализации совпадают с соответствующими характеристиками любой другой реализации*). Как и в случае стационарности, можно определить эргодичность в узком смысле (если потребовать выполнения условия эргодичности для распределений вероятностей любого порядка), в широком смысле (накладывая эргодическое ограничение на первые два момента), или еще в каком-либо специальном смысле.

Признаками классификации могут служить также некоторые различия между распределениями вероятностей, характеризующими случайные процессы. К таким признакам относятся размерность и порядок распределения. Размерностью распределения называется число различных переменных (которым соответствуют физически различные величины) в совместном распределении. Например, распределение векторной величины задает совместную вероятность (или плотность вероятности) амплитуды и фазы вектора; следовательно, размерность этого распределения равна двум. Порядок распределения определяется числом значений одной и той же переменной, отстоящих друг от друга на заданные интервалы времени.

Можно также классифицировать случайные процессы по типу распределений, например, гауссов процесс, релеевский процесс, процесс с равномерным распределением и т. д.

Наконец, признаком классификации случайных процессов может быть степень случайности процесса. Степень случайности определяется статистической зависимостью между близкими значениями реализации.

Дальнейшую детализацию продемонстрируем на дискретных процессах. Дискретный случайный процесс называется марковским, если явная статистическая зависимость распространяется в прошлое только на один шаг, т. е.

$$p(x^n,\ t_n|x_{n-1},\ t_{n-1};\ x_{n-2},\ t_{n-2};\ldots) = \\ = p(x_n,\ t_n|x_{n-1},\ t_{n-1}). \qquad (4.3)$$

Если такая зависимость распространяется на k шагов в прошлое, т. е.

$$p(x_n,\ t_n|x_{n-1},\ t_{n-1};\ x_{n-2},\ t_{n-2};\ldots) = \\ = p(x_n,t_n|x_{n-1},t_{n-1};\ldots;\ x_{n-k},\ t_{n-k}), \qquad (4.4)$$

*) Строго говоря, эргодический ансамбль может включать в себя реализации, не удовлетворяющие этому условию, но суммарная вероятность таких реализаций должна быть сколь угодно близка к нулю.

то процесс называется обобщенным марковским процессом k-го порядка.

Характер дальнодействия статистической связи значений реализации может быть весьма своеобразным. Соответственно этому возникают новые классы случайных процессов. Например, Дж. Дуб [4] ввел в рассмотрение мартингалы, т. е. процессы, для которых математическое ожидание значения реализации при следующем шаге равно значению, осуществляющемуся на последнем шаге:

$$M(x_n,\ t_n|x_{n-1},\ t_{n-1};\ldots) = x_{n-1}. \tag{4.4}$$

Другим примером могут служить процессы с независимыми приращениями, для которых разности между соседними значениями случайной переменной $(x_n - x_{n-1})$, $(x_{n-1} - x_{n-2})$ и т. д. являются независимыми случайными величинами.

Приведенная классификация, конечно, не исчерпывает всех типов случайных процессов, но показывает, как велико число классов и насколько сильно они могут различаться. Некоторые из классов случайных процессов можно описать с помощью сравнительно простых математических моделей, которые позволяют изучить поведение соответствующих сигналов в реальных системах.

ГЛАВА III

ОБЗОР НЕКОТОРЫХ КОНКРЕТНЫХ МОДЕЛЕЙ СИГНАЛОВ

§ 1. ВВЕДЕНИЕ

Одной из основных задач теории структуры сигналов является адекватное математическое описание сигналов. В предыдущей главе было показано, что основные свойства сигналов охватываются математической моделью случайного процесса. Передаваемые сигналы рассматриваются как члены множества неслучайных функций времени, образующих в совокупности некоторый случайный процесс. Статистические характеристики этого процесса полностью описывают свойства сигналов.

Однако такой весьма общий подход в значительной мере затрудняет решение многих конкретных задач, например, задач о линейных и нелинейных преобразованиях сигнала, о свойствах заданных подмножеств реализаций процесса и т. п. Возникает необходимость более конкретного описания сигналов, разработки модели, достаточно полно отражающей свойства сигналов и вместе с тем достаточно простой. Естественный путь удовлетворения этих требований заключается в выдвижении различных моделей для различных классов сигналов: статических и динамических, непрерывных и дискретных, и т. д. Наконец, даже сигналы одного типа (например, дискретные во времени) могут столь сильно отличаться по (статистическим) свойствам, что для их описания требуются различные модели случайных процессов. В § 3 предыдущей главы были рассмотрены некоторые типы таких процессов. И все же, необходимо еще более упростить описание сигналов. Идея заключается в том, что трудности оперирования с реализациями в целом могли бы быть обойдены, если бы любую реализацию процесса можно было представить как совокупность более простых, «элементарных» процессов.

Однако, если возможность такого представления очевидна для дискретных сигналов, сама структура которых наводит

на мысль о существовании простейшего сигнала, то представление непрерывных сигналов в виде совокупности «элементарных» сигналов является весьма сложной проблемой, отдельные аспекты которой не решены окончательно и до сих пор. Вопросам разработки и методам использования конкретных моделей сигналов и посвящена данная глава. Здесь также обсуждаются важные вопросы сравнения различных моделей, а также вопросы полноты описания реальных сигналов этими моделями. Разнообразие и обилие материала сделали изложение несколько отрывистым: отдельные параграфы можно читать вне связи с остальными.

§ 2. ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ СИГНАЛОВ, ЯВЛЯЮЩИХСЯ МАРКОВСКИМИ ПРОЦЕССАМИ

Достаточно широкий класс реальных сигналов охватывается их моделью в виде марковского процесса. Пусть сложный сигнал строится из некоторого дискретного множества элементарных сигналов (например, в телеграфии такими элементарными сигналами являются точка, тире и пауза). Если источник генерирует в данный момент j-й элементарный сигнал, то будем говорить, что он находится в j-м состоянии. Предположение о марковости процесса сводится к тому, что его полное описание заключается в задании набора элементарных сигналов (состояний) $\{ \varepsilon_j \}$ и условных вероятностей p_{jk} перехода источника из состояния j в состояние k.

Для более широкой применимости модели включим в рассмотрение случай, когда вероятность перехода p_{jk} зависит от времени пребывания источника в состоянии j: вероятность перехода в состояние k в интервале $(\tau, \tau + d\tau)$ после перехода в состояние j равна $p_{jk}(\tau)\, d\tau$.

Этот сложный процесс может быть отражен очень простой графической моделью. Способ построения модели заключается в следующем. Состояниям источника ставятся в соответствие точки (узлы); возможность перехода из данного состояния в другое отображается наличием линии (ветви), соединяющей соответствующие узлы; направление перехода указывается стрелкой; вероятность перехода указывается числом около надлежащей ветви. Величины подчиняются очевидному соотношению: $\sum_k p_{jk} = 1$.

Полученный в результате график может выглядеть, например, как на рис. 1.

Такие графики, получившие специальное название графов, изучались с различных точек зрения: как абстрактные математические объекты [10], как отображение сложных радиотехнических [25, 26, 30] или вычислительных [24] схем

и, наконец, как отображение случайных сигналов [7, 18, 14, 35]. Было установлено существование преобразований графа, не изменяющих определенные свойства, но упрощающих его структуру. Рассмотрим некоторые преобразования графа, изображающего марковский случайный процесс.

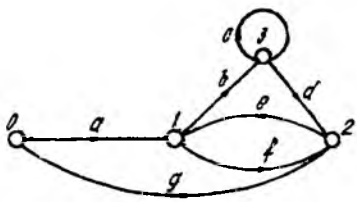

Рис. 1.

В некоторых случаях нас может интересовать только переход между двумя состояниями, хотя при этом источник обязательно проходит промежуточное состояние. Этому соответствует граф на рис. 2а. Задача состоит в том, чтобы

Рис. 2.

граф рис. 2а эквивалентно*) заменить графом рис. 2в; p_{02} должно при этом быть выражено через известные p_{01} и p_{12}. Из общих положений теории вероятностей следует, что

$$p_{02}(\tau) = \int p_{01}(t) p_{12}(\tau - t) dt = p_{01} * p_{12}, \qquad (2.1)$$

так как, по условию, переходы $0 \longrightarrow 1$ и $1 \longrightarrow 2$ являются независимыми случайными событиями. Так как для $\tau < 0$ все вероятности обращаются в нуль, то удобно воспользоваться преобразованием Лапласа

$$L[p_{jk}(\tau)] = P_{jk}(s). \qquad (2.2)$$

Тогда $P_{02}(s)$ запишется просто как

$$P_{02}(s) = P_{01}(s) \cdot P_{12}(s). \qquad (2.3)$$

*) Эквивалентность соблюдается лишь по отношению к интересующему нас переходу. В целом же граф преобразуется необратимо: промежуточное состояние при таком преобразовании выпадает.

При наличии двух параллельных, одинаково направленных ветвей (см. рис. 3) данный график можно заменить графом с одной ветвью с очевидным соотношением

$$P_{01} = P'_{01} + P''_{01}. \tag{2.4}$$

Рис. 3.

С помощью этих двух простейших правил легко получить более сложные преобразования графа. Например, граф на рис. 4 может быть заменен эквивалентным (по отношению к переходу 0→2) графом справа на том же рисунке. При этом, очевидно,

$$P_{02}(s) = P_{01}(s) \cdot P_{12}(s) + P_{03}(s) P_{32}(s). \tag{2.5}$$

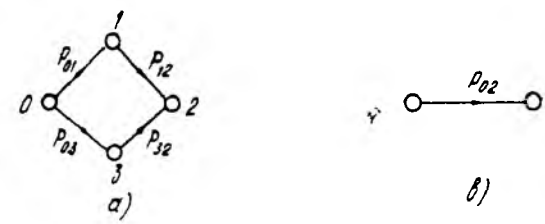

Рис. 4.

Из предыдущего явствует, что пользоваться преобразованиями Лапласа функций $p_{jк}(\tau)$ более удобно, чем самими функциями $p_{jк}(\tau)$. Кроме удобства при написании признаков ветвей графа, функции $P_{jк}(s)$ позволяют очень просто вычислить временные моменты сигналов. Пусть, например, нас интересуют только те сигналы, которые начинаются с состояния j и кончаются состоянием $к$. После соответствующих преобразований графа можно получить функцию $P_{jк}(s)$. Коэффициенты степенного ряда

$$P_{jк}(s) = a_0 + a_1 s + a_2 s^2 + \ldots \tag{2.6}$$

могут быть найдены, как обычно, по формулам

$$a_0 = \int p(\tau)\, d\tau = P(0),$$

$$a_1 = \int \tau p(\tau) d\tau = P'(0), \qquad (2.7)$$

$$2a_2 = \int \tau^2 p(\tau) d\tau = P''(0),$$

. . . . . .

Легко видеть, что a_0 дает безусловную вероятность осуществления перехода $j \to \kappa$, т. е. вероятность появления сигнала любой длительности, начинающегося с j-го символа и заканчивающегося κ-м. Величина $-\frac{a_1}{a_0}$ характеризует среднюю длительность такого сигнала; дисперсия длительности определяется величиной

$$\frac{2a_2}{a_0} - \left(\frac{a_1}{a_0}\right)^2.$$

Прежде чем переходить к примерам, приведем еще одно важное преобразование графов. Возможны случаи, когда ветвь графа окажется замкнутой на тот же узел, с которого она началась (рис. 5а). Желая найти выражение d через b и c, представим исходный граф в виде рис. 6, расщепив мультиплетный узел 2 на воображаемые узлы, в которых система может находиться лишь один раз. Пользуясь приведенными выше правилами, легко показать, что

$$d(s) = b[1 + c + c^2 + c^3 + \ldots] = \frac{b(s)}{1 - c(s)}. \qquad (2.8)$$

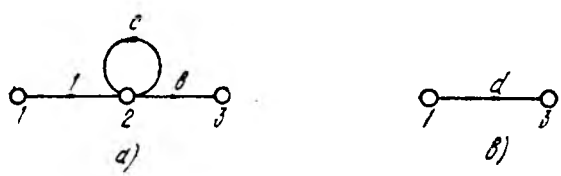

Рис. 5.

В этой сумме член n-й степени соответствует вероятности n оборотов по петле. Так как $b(s) + c(s) \leqslant 1$, т. е. в каждый данный момент переход может осуществиться или нет, то $d = \frac{b}{1-c}$ не может превзойти единицу, хотя на первый взгляд это кажется возможным.

Так как петля и ветвь, отходящая от нее, соединены последовательно, величину d формально можно разложить на два сомножителя: b и $1/(1-c)$.

При этом величина $1/(1-c) = U(s)$ —

характеризует только петлю. Анализ этой величины, проведенный Хаггинсом [35], показывает, что ее обратное преобразование Лапласа, $U(\tau) = L^{-1}[U(s)]$, имеет смысл интенсивности потока событий, заключающихся в последовательных оборотах по петле. Количественно интенсивность потока событий $U(\tau)$ характеризует плотность потока во времени, а именно: величина $U(\tau)\,d\tau$ равна среднему числу переходов по петле в интервале $(\tau, \tau+d\tau)$.

Изложенные выше правила и следующие из них преобразования графов приведены на рис. 7. Рассмотрим теперь, как и для каких целей можно пользоваться графом.

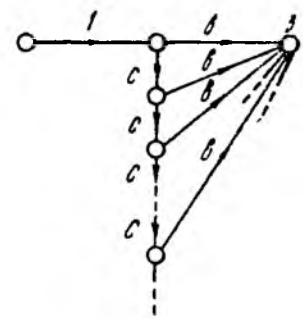

Рис. 6.

Из полного графа случайного процесса легко определить любую интересующую нас реализацию, выделив в нем соответствующую траекторию. По характеристикам ветвей траектории находится вероятность данной реализации и ее остальные временные статистические характеристики.

Наиболее эффективно граф может быть использован для нахождения статистических характеристик подмножества сигналов, выделяемого по какому-либо признаку. Продемонстрируем это на простых примерах сигналов, состоящих из двух символов, например 0 и 1. Пусть через каждые T секунд генерируется новый символ и вероятность появления каждого символа равна $\frac{1}{2}$. Следовательно,

$$p_{00}(\tau) = p_{10}(\tau) = p_{01}(\tau) = p_{11}(\tau) = \frac{1}{2}\cdot\delta(\tau - T);$$

преобразование Лапласа этой функции запишется как

$$\frac{1}{2}e^{-sT} = \frac{x}{2}.$$

В качестве первого примера рассмотрим ансамбль сигналов, отличающихся следующим признаком: последний символ в последовательности — единица, все предыдущие — нули. (В ансамбль включается сигнал длительностью в один символ, являющийся единицей). Граф ансамбля таких сигналов изображен на рис. 8. Реализация продолжается до тех пор, пока текущая точка «оборачивается» по петле; переходом по правой ветви в точку «1» сигнал заканчивается. Пользуясь правилами, приведенными выше, сразу получаем,

что преобразование Лапласа для распределения вероятностей интересующего нас перехода

$$P(s) = \frac{\frac{x}{2}}{1-\frac{x}{2}} = \frac{x}{2-x} = \frac{1}{2}x + \frac{1}{4}x^2 + \frac{1}{8}x^3 + \dots . \tag{2.9}$$

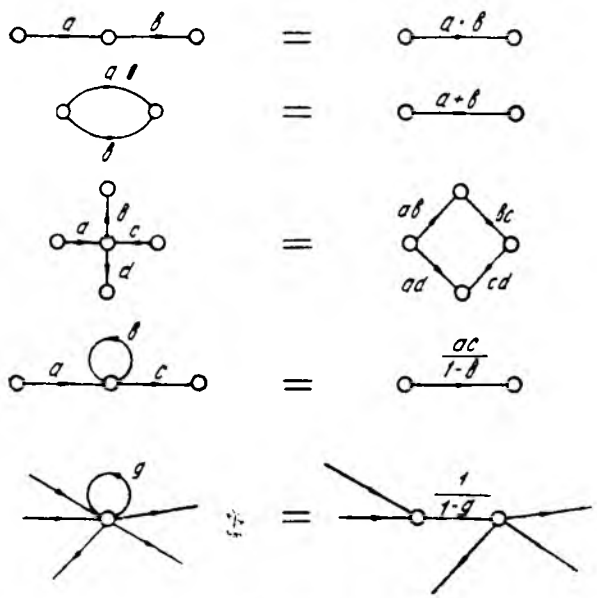

Рис. 7.

Переходя к разложению в ряд по s, получим

$$P(s) = 1 - 2Ts + 3T^2s^2 + \dots \tag{2.9'}$$

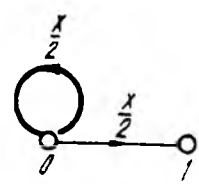

Рис. 8.

Отсюда можно сделать следующие выводы:

1. Так как $a_0 = 1$, то событие, заключающееся в возникновении сигнала с указанным признаком, достоверно.

2. Средняя длина такого сигнала равна $-a_1 = 2T$.

3. Дисперсия длины сигналов равна $2T^2$.

4. Трансформирование во время ряда (2.9) показывает, что коэффициенты ряда дают вероятности появления сигнала с заданным признаком и заданной длительности: вероятность такого сигнала длительности в n символов равна $(1/2)^n$.

Вторым примером может послужить рассмотрение ансамбля сигналов со следующим признаком: три последних символа в сигнале — единицы, предыдущие символы могут быть любыми, но не содержат трех единиц подряд. Исходный граф и постепенные его упрощения приведены на рис. 9. Из последнего графа сразу следует, что (так как $a = x/2$)

$$P(s) = \frac{a^3}{1 - a - a^2 - a^3} \cdot \frac{1}{8 - 4x - 2x^2 - x^3} =$$

$$= 1 - 14sT + \frac{667}{4}(sT)^2 + \ldots . \qquad (2.10)$$

Из (2.7) и (2.10) следует, что

1. Появление сигнала с заданным признаком является достоверным событием ($a_0 = 1$).

2. Средняя длительность такого сигнала равна $14\ T$ ($a_1 = -14\ T$).

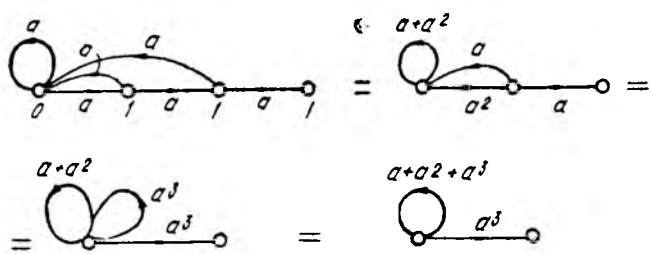

Рис. 9.

3. Дисперсия длительности равна $(11{,}9\cdot T)^2$.

4. Деля числитель (2.10) на знаменатель, получаем распределение вероятностей по длительностям сигналов с заданным признаком:

$$\{p_n\} = \frac{1}{8}\left\{0,\ 0,\ 0,\ \frac{1}{2},\ \frac{1}{2},\ \frac{1}{2},\ \frac{7}{16},\ \ldots\right\}. \qquad (2.11)$$

Можно получить асимптотическое распределение длительностей при больших n, разложив $P(s)$ на простые дроби и оценив слагаемые, медленно убывающие с ростом n. Так

как знаменатель имеет нуль в точке $x_1 = 1{,}087$ и два комплексных корня, $|x_{2;3}| = 2{,}71$, то

$$\frac{x^3}{8 - 4x - 2x^2 - x^3} = \frac{0{,}108}{1{,}087 - x} + \\ + \frac{N(x)}{7{,}36 + 3{,}087x + x^2}. \qquad (2.12)$$

Разложение в ряд по степеням x первого слагаемого дает асимптотическое распределение

$$p_n = 0{,}108 \left| \frac{1}{1{,}087} \right|^{n-1}. \qquad (2.13)$$

Приближение получается весьма хорошим: при $n = 7$ (2.13) дает 0,055 при точном $p_n = 0{,}0546$.

Этими примерами мы закончим рассмотрение графов, моделирующих марковские сигналы. Следует, однако, указать, что этим не исчерпываются возможности данной модели сигналов; Л. Задэ, например, показал [18], что применение графа может облегчить нахождение корреляционных характеристик дискретных марковских сигналов.

§ 3. ПРЕДСТАВЛЕНИЕ СИГНАЛОВ С ПОМОЩЬЮ МЕТОДА КАНОНИЧЕСКИХ РАЗЛОЖЕНИЙ

Среди методов изучения случайных процессов, моделирующих сигналы, известен развитый в последние годы В. С. Пугачевым [27] метод канонических разложений. Этот метод весьма эффективен при исследовании изменения статистических характеристик процесса при его преобразованиях, особенно линейных.

Первым шагом в применении метода канонических разложений является представление случайного процесса в виде суммы (конечной или бесконечной) элементарных случайных функций. Элементарной случайной функцией называется функция вида

$$x(t) = V \cdot \varphi(t), \qquad (3.1)$$

где V — обычная случайная величина, $\varphi(t)$ — обычная (неслучайная) функция времени. Элементарная случайная функция характерна тем, что две особенности случайного процесса здесь явно разделены: случайность вся сосредоточена в коэффициенте V, а зависимость от времени — в обычной функции $\varphi(t)$.

Рассмотрим свойства элементарной случайной функции. Математическое ожидание функции $x(t)$ равно

$$m_x(t) = M[V \cdot \varphi(t) = m_V \cdot \varphi(t), \tag{3.2}$$

где m_v — среднее значение случайной величины V. Если $m_V = 0$, то $m_x(t) \equiv 0$. Так как любую случайную функцию можно центрировать (т. е. привести к виду, для которого $m_x = 0$), то в дальнейшем будем рассматривать только центрированные элементарные случайные функции.

Функция корреляции для $x(t)$ выразится как

$$R_x(t, t') = M[x(t) \cdot x(t')] = \varphi(t)\varphi(t')MV^2 = \varphi(t)\varphi(t')DV, \tag{3.3}$$

где DV — дисперсия случайной величины V.

Пусть L — некоторый линейный оператор. Так как

$$Lx(t) = V \cdot L\varphi(t), \tag{3.4}$$

то при линейном преобразовании элементарной случайной функции все сводится только к линейному преобразованию известной функции времени $\varphi(t)$. Отсюда следует способ применения метода канонических разложений: если некоторый случайный процесс подвергается линейному преобразованию, то для облегчения расчетов его нужно представить (точно или приближенно) в виде суммы элементарных случайных функций, а затем подвергать преобразованию.

Пусть мы имеем сигнал, представимый в виде

$$y(t) = m_y(t) + \overset{\circ}{y}(t), \tag{3.5}$$

где $\overset{\circ}{y}(t)$ — случайный процесс со средним, равным нулю. Допустим, что нам удалось представить (точно или приближенно) в виде суммы по элементарным случайным функциям:

$$y(t) = m_y(t) + \sum_{k=1}^{n} V_k \varphi_k(t). \tag{3.6}$$

Такое представление называется разложением случайного процесса. Случайные величины $\{V_k\}$ называются коэффициентами разложения, а неслучайные функции $\{\varphi_k(t)\}$ называются координатными функциями.

Подвергнем теперь $y(t)$ линейному преобразованию L:

$$z(t) = Ly(t) = m_z(t) + \sum_{k=1}^{n} V_k \psi_k(t). \tag{3.7}$$

Здесь $m_z(t) = Lm_y(t)$ и $\psi_\kappa(t) = L\varphi_\kappa(t)$. Из (3.7) следует, что если случайный процесс $y(t)$, заданный разложением по элементарным случайным функциям, подвергается линейному преобразованию L, то коэффициенты разложения остаются неизменными, а математическое ожидание и координатные функции подтвергаются тому же линейному преобразованию.

Рассмотрим корреляционную функцию и дисперсию центрированного случайного процесса $\overset{\circ}{y}(t)$, заданного разложением. Пусть $\{V_\kappa\}$ — система случайных величин с математическим ожиданием нуль, а коэффициенты корреляции между ними равны $\{R_{\kappa l}\}$. Тогда

$$R_y(t,\ t') = M[\overset{\circ}{y}(t)\cdot\overset{\circ}{y}(t')] = M[\sum_\kappa V_\kappa\varphi_\kappa(t)\ \cdot\sum_l V_l\varphi_l(t')] =$$

$$= \sum_{\kappa,l} M[V_\kappa\cdot V_l]\ \varphi_\kappa(t)\varphi_l(t') = \sum_\kappa \varphi_\kappa(t)\varphi_\kappa(t')DV_\kappa +$$

$$+\sum_{\kappa\neq l}\varphi_\kappa(t)\varphi_l(t')R_{\kappa l}. \qquad (3.8)$$

При $t = t'$

$$R_y(t,\ t) = Dy(t) = \sum_\kappa [\varphi_\kappa(t)]^2 D\,V_\kappa + \sum_{\kappa\neq l}\varphi_\kappa(t)\varphi_l(t)R_{\kappa l}. \qquad (3.9)$$

Эти выражения приобретают особенно простой вид, если все величины $\{V_\kappa\}$ некоррелированы, т. е.

$$R_{\kappa l} = \begin{cases} 1, & \kappa = l, \\ 0, & \kappa \neq l. \end{cases}$$

Разложение, коэффициенты которого некоррелированы, называется каноническим разложением. Кроме очевидных преимуществ в простоте, канонические разложения имеют целый ряд весьма полезных свойств. Например, линейное преобразование канонического разложения случайного процесса приводит к каноническому же разложению преобразованного процесса. Кроме того, В. С. Пугачев доказал, что если задано каноническое разложение функции корреляции случайного процесса, то сама случайная функция имеет каноническое разложение с теми же координатными функциями.

Основную трудность в применении метода канонических разложений представляет нахождение координатных функций для заданного случайного процесса. В. С. Пугачевым разработан ряд методов нахождения координатных функций при задании различных характеристик процесса.

В качестве примера рассмотрим простой, но полезный случай нахождения одного из канонических разложений стационарного случайного процесса на конечном интервале времени $(-T, T)$. Для отыскания координатных функций воспользуемся теоремой о совпадении координатных функций разложений процесса и его функции корреляции.

Пусть задана корреляционная функция процесса $R_x(\tau)$. Так как $\tau = t - t'$ и $-T \leqslant t, t' \leqslant T$, то $\tau \in [-2T, 2T]$. Функция $R_x(\tau)$ является четной (для действительного процесса), поэтому ее можно разложить в ряд по косинусам в интервае ее определения:

$$R_x(\tau) = \sum_{\kappa=0}^{\infty} \Delta_\kappa \cos \omega_\kappa \tau; \tag{3.10}$$

$$\omega_\kappa = 2\pi \cdot \frac{\kappa}{4T},$$

$$\Delta_0 = \frac{1}{4T} \int_{-2T}^{2T} R_x(\tau) d\tau,$$

$$\Delta_\kappa = \frac{1}{2T} \int_{-2T}^{2T} R_x(\tau) \cos \omega_\kappa \tau d\tau.$$

Из четности $R_x(\tau)$ следует, что

$$\Delta_0 = \frac{1}{2T} \int_{0}^{2T} R_x(\tau) d\tau,$$

$$\Delta_\kappa = \frac{1}{T} \int_{0}^{2T} R_x(\tau) \cos \omega_\kappa \tau d\tau.$$

Выразив в (3.10) τ как $t - t'$, имеем:

$$R_x(t - t') = \sum_{\kappa=0}^{\infty} \Delta_\kappa (\cos \omega_\kappa t \cdot \cos \omega_\kappa t' + \sin \omega_\kappa t \cdot \sin \omega_\kappa t'). \tag{3.11}$$

Следовательно, каноническое разложение функции корреляции имеет координатными функциями синусы и косинусы следующих периодов:

$$4T, \ 2T, \ \frac{4}{3} T, \ T, \ \frac{4}{5} T, \ldots, \ \frac{4}{\kappa} T, \ldots . \tag{3.12}$$

Так как, согласно теореме Пугачева, каноническое разложение случайного процесса имеет те же координатные функции, что и каноническое разложение функции корреляции, то искомое разложение можно сразу записать в виде

$$x(t) = m_x(t) + \sum_{\kappa=0}^{\infty} (X_\kappa \sin \omega_\kappa t + Y_\kappa \cos \omega_\kappa t), \qquad (3.13)$$

где X_κ и Y_κ — некоррелированные случайные величины со средними значениями, равными нулю. Коэффициенты X_κ и Y_κ с одинаковыми номерами имеют одинаковую дисперсию Δ_κ.

Полученное разложение (3.13) в общем случае не является рядом Фурье для функции $x(t)$ в интервале $(-T, T)$. Только в том случае, когда в разложении (3.10) функции корреляции отсутствуют все нечетные гармоники, в разложении (3.13) останутся только гармоники периода $2T$, и (3.13) будет рядом Фурье.

При задании других условий, когда исходные данные задаются не в виде функции корреляции, способы нахождения канонического разложения усложняются; однако они разработаны в достаточной для практического применения степени. Других способов здесь мы рассматривать не будем: интересующихся можно отослать к монографии В. С. Пугачева [27]. В той же книге развита теория интегральных канонических разложений, т. е. представления процесса не в виде суммы, а через интеграл.

§ 4. СИГНАЛЫ, ОГРАНИЧЕННЫЕ ПО ШИРИНЕ СПЕКТРА ИЛИ ПО ДЛИТЕЛЬНОСТИ. ТЕОРЕМА КОТЕЛЬНИКОВА И ЕЕ АНАЛОГ В ЧАСТОТНОМ ПРЕДСТАВЛЕНИИ

Рассмотрим свойства сигналов, обладающих ограниченным спектром. Отложив до следующих параграфов обсуждение того, в какой степени предположение об ограниченности спектра сигнала соответствует реальности, будем считать, что можно указать некоторую частоту F, выше которой в спектре сигнала составляющие отсутствуют.

Для таких сигналов В. А. Котельников доказал следующую теорему:

Любая реализация случайного процесса со спектром, находящимся в интервале $(0, F)$, полностью определяется последовательностью ее значений в точках, отстоящих на $\frac{1}{2F}$ секунд друг от друга.

Пусть $f(t)$ — некоторая реализация случайного процесса, обладающая ограниченным спектром $S(\omega)$:

$$S(\omega) = \begin{cases} \int_{-\infty}^{\infty} f(t) e^{i\omega t}\, dt, & |\omega| \ll 2\pi F, \\ 0, & |\omega| > 2\pi F. \end{cases} \tag{4.1}$$

В интервале $(-2\pi F, 2\pi F)$ функцию $S(\omega)$ можно представить в виде ряда Фурье:

$$S(\omega) = \sum_{\kappa=-\infty}^{\infty} C_\kappa e^{-\frac{2\pi i}{4\pi F}\omega \cdot \kappa} = \sum_{\kappa=-\infty}^{\infty} C_\kappa e^{-i\frac{\kappa\omega}{2F}}. \tag{4.2}$$

Коэффициенты Фурье в этом случае запишутся в виде

$$C_\kappa = \frac{1}{4\pi F} \int_{-2\pi F}^{2\pi F} S(\omega) e^{i\frac{\kappa\omega}{2F}}\, d\omega = \frac{1}{2F} \cdot f\left(\frac{\kappa}{2F}\right). \tag{4.3}$$

Следовательно,

$$S(\omega) = \frac{1}{2F} \sum_{-\infty}^{\infty} f\left(\frac{\kappa}{2F}\right) e^{-i\frac{\kappa\omega}{2F}}. \tag{4.4}$$

Уже это соотношение доказывает теорему Котельникова: так как между $f(t)$ и $S(\omega)$ имеется однозначная связь, то $f(t)$, как и $S(\omega)$ (см. (4.4)), однозначно определяется отсчетами $\left\{f\left(\frac{\kappa}{2F}\right)\right\}$. Представляет, однако, интерес установить, каким именно образом по множеству отсчетов можно построить промежуточные значения функции $f(t)$. Воспользовавшись связью между данной реализацией и ее спектром, получим:

$$f(t) = \frac{1}{4\pi F} \int_{-2\pi F}^{2\pi F} S(\omega) e^{i\omega t}\, d\omega =$$

$$= \frac{1}{4\pi F} \int_{-2\pi F}^{2\pi F} \sum_{-\infty}^{\infty} f\left(\frac{\kappa}{2F}\right) e^{-i\omega\left(\frac{\kappa}{2F} - t\right)} d\omega =$$

$$= \frac{1}{2F} \sum_{-\infty}^{\infty} f\left(\frac{\kappa}{2F}\right) \cdot \frac{1}{2\pi} \int_{-2\pi F}^{2\pi F} e^{-i\omega\left(\frac{\kappa}{2F} - t\right)} d\omega =$$

$$= \sum_{-\infty}^{\infty} f\left(\frac{\kappa}{2F}\right) \cdot \frac{\sin(2\pi Ft - \kappa\pi)}{2\pi Ft - \kappa\pi} \quad . \qquad (4.5)$$

Таким образом, мы получили разложение реализации, координатными функциями которого служат функции вида $\sin x/x$, а коэффициентами разложения являются значения самой реализации, отстоящие на $\frac{1}{2F}$ *сек* друг от друга. Тем самым доказана теорема Котельникова и указан способ построения функции $f(t)$ по ее отсчетам.

Относительно спектра любого сигнала конечной длительности может быть высказано утверждение, аналогичное теореме Котельникова:

Если $S(\omega)$ есть спектр функции $f(t)$, тождественно равной нулю вне интервала (T_1, T_2), то $S(\omega)$ однозначно определяется последовательностью значений спектра в точках, отстоящих на $1/(T_2 - T_1)$ герц друг от друга.

Доказательство проводится аналогично доказательству теоремы Котельникова и приводит к разложению функции $S(\omega)$ по координатным функциям того же вида $\sin x\,x$. В простейшем случае, когда $T_1 = -T\,2$, $T_2 = T/2$, получим:

$$S(\omega) = \sum_{-\infty}^{\infty} S\left(\frac{2\pi}{T} \cdot n\right) \frac{\sin\left(\frac{\omega}{2} T - \pi n\right)}{\frac{\omega}{2} T - \pi n} \quad . \qquad (4.6)$$

Ввиду того, что как во временном, так и в частотном представлении теорема Котельникова приводит к разложениям по координатным функциям вида $\sigma(ax - \pi n) = \frac{\sin(ax - \pi n)}{ax - \pi n}$, интересно рассмотреть свойства таких функций. Читатель может легко доказать сам справедливость следующих утверждений:

1. Функции $\{\sigma(ax - \pi n)\}$ $(n = 0, \pm 1, \pm 2, \ldots)$ образуют семейство ортогональных функций, т. е.

$$\int_{-\infty}^{\infty} \sigma(ax - \pi n)\,\sigma(ax - \pi m)\,dx = \begin{cases} \frac{\pi}{a}, & n = m, \\ 0, & n \neq m. \end{cases} \qquad (4.7)$$

2. Фурье-преобразование функции $\sigma(ax)$ имеет вид:

$$\int_{-\infty}^{\infty} e^{-i\lambda x}\sigma(ax)dx = \begin{cases} \dfrac{\pi}{a}, & |\lambda| \ll a, \\ 0, & |\lambda| > a. \end{cases} \tag{4.8}$$

В частности, амплитудный спектр временной функции отсчета $\sigma(2\pi Ft - \pi\kappa)$ равномерен в интервале $(-2\pi F, 2\pi F)$.

Подчеркнем, что условия теоремы Котельникова не накладывают никаких ограничений на конкретный вид спектра внутри интервала (O, F). Интересно в связи с этим в качестве наглядного примера рассмотреть представление рядом Котельникова реализаций такого случайного процесса, который состоит из ансамбля синусоид с частотами $\{\omega_l\}$, $\omega_l \ll \Omega = 2\pi F$, $l = 1,2,\ldots$. Согласно теореме Котельникова любая синусоида частоты ω_l может быть представлена рядом:

$$\sin \omega_l t = \sum_{-\infty}^{\infty} \sin\left(\frac{\omega_l}{2F}\kappa\right) \cdot \sigma(\Omega t - \kappa\pi). \tag{4.9}$$

Такая запись на первый взгляд может вызвать недоумение: в левой части равенства стоит функция, амплитудный спектр которой есть $\delta(\omega - \omega_l)$, а справа — сумма функций отсчета, имеющих равномерные в интервале $(-\Omega, \Omega)$ амплитудные спектры. Противоречие здесь, конечно, только кажущееся. Учет фазовых соотношений между спектрами координатных функций дает результирующий спектр разложения, совпадающий со спектром реализации. Чтобы показать это, найдем спектр правой части равенства (4.9).

Функция $\sigma(\Omega t - \kappa\pi)$ имеет спектр

$$S_\sigma(\omega) = e^{i\frac{\omega}{2F}\kappa}, \quad |\omega| \ll \Omega. \tag{4.10}$$

Следовательно, спектр всей суммы запишется как

$$S_\Sigma(\omega) = \sum_{-\infty}^{\infty} \sin\left(\frac{\omega_l}{2F}\kappa\right) e^{i\frac{\omega}{2F}\kappa}. \tag{4.11}$$

Воспользовавшись соотношением [29]

$$\sum_{\kappa=1}^{\infty} p^\kappa \sin \kappa x = \frac{p \sin x}{1 - 2p\cos x + p^2}, \quad |p| < 1 \tag{4.12}$$

и заменой $\omega 2F = x$, получим, что

$$S_\Sigma(\omega) = \sum_{-\infty}^{\infty} (e^{ix})^\kappa \sin \kappa x_l =$$

$$= \sum_{\kappa=1}^{\infty} (e^{ix})^\kappa \sin \kappa x_l - \sum_{\kappa=1}^{\infty} (e^{-ix})^\kappa \sin \kappa x_l =$$

$$= \frac{e^{ix} \sin x_l}{1 - 2e^{ix} \cos x_l + e^{2ix}} - \frac{e^{-ix} \sin x_l}{1 - 2e^{-ix} \cos x_l + e^{-2ix}}. \qquad (4.13)$$

Приведя это выражение к общему знаменателю, получим, что знаменатель пропорционален величине $(\cos x_l - \cos x)^2$, а числитель тождественно равен нулю. Чтобы выяснить характер $S_\Sigma(\omega)$ в точке ω_l, рассмотрим ряд (4.11) при $\omega = \omega_l$:

$$S_\Sigma(\omega) = \sum_{-\infty}^{\infty} e^{i\kappa x_l} \sin \kappa x_l = \frac{1}{2} \left[\sum_{-\infty}^{\infty} e^{2i\kappa x} - \sum_{-\infty}^{\infty} 1 \right]. \qquad (4.12)$$

Легко показать, что сумма первого ряда равна нулю и, следовательно, $|S_\Sigma(\omega_l)| = \infty$. Таким образом, результирующий спектр разложения является δ-функцией, отнесенной к точке ω_l, как и спектр исходной реализации.

Одним из важных следствий теоремы Котельникова является определение числа степеней свободы сигнала с ограниченным спектром на заданном интервале времени: так как отсчеты следуют через $\frac{1}{2F}$ секунд, то их число на протяжении T *сек* равно $2FT$. Это, однако, не означает, что задание всех $2FT$ отсчетов на интервале T полностью определит сигнал на этом интервале: функции отсчета, относящиеся к точкам, не лежащим в интервале T, тоже дают вклад в значения функции для $t \in T$ (исключая непосредственно точки отсчета, в которых все функции отсчета, кроме одной, обращаются в нуль). Аналогично, если сигнал имеет конечную длительность T, то число степеней свободы, приходящееся на интервал частот F неограниченного спектра сигнала, также равно $2FT$.

Таким образом, если говорить строго, то теорема Котельникова позволяет сделать лишь следующее утверждение: имеющий ограниченный спектр сигнал, заданный в точках отсчета на всей оси времени, за исключением интервала длительностью T, имеет $2FT$ степеней свободы. При отсутствии каких-либо сведений о сигнале вне интервала T такого утверждения теорема Котельникова не позволяет сделать. Предположение о «равенстве нулю» сигнала вне T не спасает положения: при этом спектр становится неограниченным

и теорема Котельникова предписывает брать отсчеты сколь угодно близко друг от друга.

Можно оценить величину погрешности представления отрезка сигнала длительностью $(-T, T)$ конечным числом членов ряда Котельникова, т. е. только теми членами, моменты отсчета которых приходятся на интервал $(-T, T)$. Пусть на интервале $(-T, T)$ укладывается $2\kappa+1$ точек отсчета. Тогда можно показать [37], что

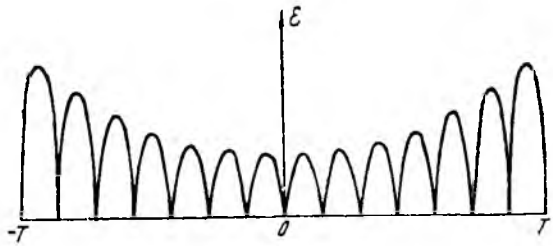

Рис. 10.

$$\left|f(t)-\sum_{-\kappa}^{\kappa} f\left(\frac{n}{2F}\right)\cdot\sigma(\pi Ft-\pi n)\right| \ll \frac{\sqrt{2}}{\pi}\cdot E\cdot \left|\sin \pi Ft\right| \sqrt{\frac{T/F}{T^2-t}} = \varepsilon E,$$

где $E=\int_{-F}^{F} S(\omega)\,d\omega$, т. е. E—полная энергия, которую несет функция $f(t)$. Множитель ε при E ведет себя как показано на рис. 10, т. е. ошибка равна нулю лишь в точках отсчета, имеет максимум между точками отсчета, и величина этих максимумов ошибки возрастает по мере приближения к границам интервала представления функции. Следует указать, что величина E является в общем случае неограниченной величиной для сигналов с ограниченным спектром, т. е. бесконечно длящихся. Поэтому ошибка рассматриваемого представления может быть неограниченной. Это еще раз указывает на неправильность часто встречающегося утверждения, будто отрезок сигнала с ограниченным спектром длительностью T «полностью характеризуется $2FT$ отсчетами».

§ 5. ОБОБЩЕНИЕ ТЕОРЕМЫ КОТЕЛЬНИКОВА НА СЛУЧАЙ СПЕКТРА, НЕ СОДЕРЖАЩЕГО НИЗКИХ ЧАСТОТ

При предположениях, допускающих ограниченность спектра сигнала, часто оказывается, что спектр можно огра-

ничить по частоте не только сверху частотой f_{max}, но и снизу ненулевой частотой f_{min}. Такое дополнительное ограничение, накладываемое на спектр сигнала, должно, очевидно, сказаться и на числе степеней свободы сигнала, т. е. на частоте следования отсчетов. Можно привести простое рассуждение, подтверждающее это. При разложении некоторой функции в ряд Фурье число степеней свободы будет определяться числом гармоник, но не тем, в какой части области частот лежат эти гармоники. Отсюда следует, что линейное смещение спектра не изменяет числа стпеней свободы сигнала, хотя и изменяет конкретный вид зависимости информативного параметра от времени. Если известно, что спектр сигнала не содержит частот ниже f_{min}, то можно, не изменив числа степеней свободы сигнала, с помощью гетеродинирования линейно сместить весь спектр влево (до нуля) на интервал f_{min}. При этом вступают в силу условия теоремы Котельникова (спектр сигнала лежит в интервале (O, F), $F = f_{max} - f_{min}$), согласно которой число степеней свободы такого сигнала, приходящееся на интервал T, равно $2FT$.

Несмотря на логичность этих рассуждений, результат, утверждающий, что при ограничении спектра снизу отсчеты можно брать реже, — кажется несколько парадоксальным. Обычно, стараясь интуитивно связать результаты теоремы Котельникова с привычными частотными представлениями, темп следования отсчетов ($\frac{1}{2F}$ в секунду) связывают [38] с тем, что за полпериода колебаний максимальной частоты суммарное колебание не может значительно измениться. Теперь же мы пришли к тому, что повышение нижней границы спектра без снижения верхней позволяет брать отсчеты реже, чем через полпериода колебания максимальной частоты. Это кажущееся противоречие легко разрешается, если учесть, что дополнительные сведения о величине f_{min} позволяют пользоваться в качестве координатных функциями более сложной структуры.

Рассмотрим комплексное колебание $u(t)$, спектр которого заключен внутри интервала положительных частот (f_{min}, f_{max}), или, обозначив среднюю частоту спектра через f_0, а ширину его — через F, в интервале $(f_0 - \frac{1}{2} f_0 + 1/2 F)$.

Проводя выкладки аналогично тому, как это делается при доказательстве теоремы Котельникова, получим, что

$$u(t) = \sum_{-\infty}^{\infty} u\left(\frac{n}{F}\right) \sigma(Ft - n) \exp\left[2\pi i f_0 \left(t - \frac{n}{F}\right)\right]. \tag{5.1}$$

Чтобы получить разложение вещественного колебания, необходимо взять вещественные части обеих частей равенства (5. 1). Полагая:
$u(t) = g(t) + ih(t)$, мы получим:

$$g(t) = \sum_{-\infty}^{\infty} g\left(\frac{n}{F}\right) \sigma(Ft - n) \cos 2\pi f_0 t -$$

$$- \sum_{-\infty}^{\infty} h\left(\frac{n}{F}\right) \sigma(Ft - n) \sin 2\pi f_0 t. \qquad (5.2)$$

Отсюда следует, что:

1) функциями отсчета являются колебания средней частоты f_0, модулированные по амплитуде функцией $\sigma(Ft)$;

2) отсчеты следуют с интервалом $\Delta t = \frac{1}{F}$;

3) в точках отсчета необходимо задавать две величины: значения функций g и h;

4) число степеней свободы сигнала на интервале времени T по-прежнему равно $2FT$. Необходимость задания двух величин в каждой точке отсчета физически означает, что в этих точках должны быть заданы мгновенное значение огибающей $\sqrt{g^2 + h^2}$ и мгновенное значение фазы несущей, arc tg(h/g).

Рассмотренная выше теорема определяет минимальную частоту следования отсчетов, при которой не происходит потери информации. Интересно отметить, что иногда повышение частоты следования отсчетов может привести к необратимому преобразованию, т. е. ухудшению представления непрерывного сигнала дискретным [11]. Это может произойти в тех случаях, когда дискретизацию непрерывного сигнала можно представить себе как процесс модуляции последовательности импульсов (с помощью которых задаются значения в точках отсчета) по высоте рассматриваемым сигналом. Спектр полученной таким образом последовательности есть свертка спектра исходного сигнала и спектра последовательности импульсов. Пусть нижняя частота спектра сигнала равна $f_{\min}$, а ширина его спектра равна F. При модуляции каждой частоте f из спектра непрерывного сигнала будут соответствовать частоты $|f + nf_s|$ в спектре дискретизированного сигнала; f_s — частота повторения дискретов, n — целое и принимает все значения из интервала $(-\infty, \infty)$. Следовательно, полосе частот $(f_{\min}, f_{\min} + F)$ в спектре дискрет-

ного сигнала будет соответствовать целый набор полос $|f_{min}|+nf_s,\ f_{min}+F+nf_s|$. Квантование будет обратимым (т. е. происходящим без потерь информации) процессом только в том случае, если ни одна из этих боковых полос не перекрывается со спектром исходного сигнала и, следовательно, с какой-либо другой боковой полосой. Это условие и накладывает ограничения на частоту следования отсчетов f_s.

Пусть $f_{min}=(N+K)F$, где N — целое, а $K<1$. Если $f_s>F$, то боковые полосы, соответствующие $n>0$, заведомо не перекроют спектр исходного сигнала. Рассмотрим случай $n<0$. Пусть n таково, что спектр исходного сигнала расположен между полосами, соответствующими n и $n+1$. Высшая частота n-й полосы равна

$$= nf \quad -f_{min}, \tag{5.3}$$

а низшая частота $(n+1)$-й полосы равна

$$f=(n+1)f_s-f_{min}-F. \tag{5.4}$$

Условие, чтобы эти частоты не лежали в полосе частот сигнала, запишется в виде

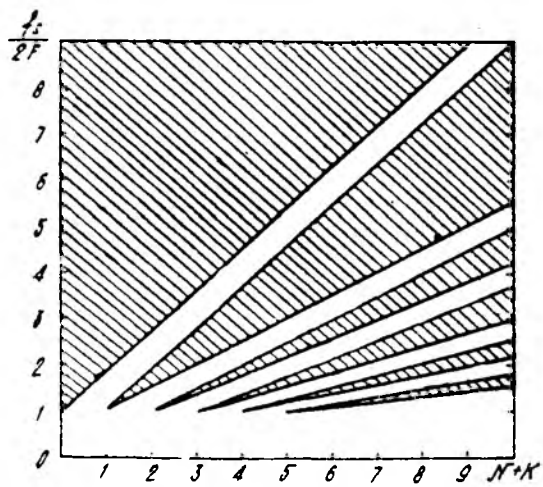

Рис. 11.

$$nf_s-f_{min}\ll f_{min}\leqslant(n+1)f_s-f_{min}-2F. \tag{5.5}$$

Границы областей допустимых частот f_s определяются, следовательно, равенствами:

$$f_{sn\ \min} = \frac{2f_{\min}}{n} = \frac{2F(N+K)}{n}, \tag{5.6}$$

$$f_{sn\ \max} = \frac{2f_{\min}+2F}{n+1} = \frac{2F(N+K+1)}{n+1}. \tag{5.7}$$

Максимальное допустимое число гармоник в спектре последовательности импульсов определится из следующих соображений: в интервале $(0, f_{\min})$ без перекрытия уложатся не более N полос; следовательно, N — максимальное значение, которое может принимать индекс n.

На рис. 11 заштрихованные области изображают области допустимых частот f_s. Легко видеть, что при заданном $N+K$ имеется $N+1$ областей допустимых частот следования отсчетов.

§ 6. ГЕОМЕТРИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ СИГНАЛОВ

В. А. Котельников [21] и К. Шэннон [38] разработали геометрическую трактовку соотношений, характеризующих сигналы. Основой геометрического представления сигналов служит тот факт, что совокупность чисел x_1, x_2,..., x_n, независимо от их происхождения, всегда может рассматриваться как совокупность координат точки в n-мерном пространстве. Согласно теореме Котельникова (§ 4), сигнал с ограниченным спектром полностью задается дискретным множеством равноотстоящих отсчетов. Совокупность чисел, характеризующих значение функций отсчета в соответствующих точках, можно также рассматривать как совокупность координат некоторой точки; таким образом, мы приходим к представлению сигнала как вектора (точки) в многомерном пространстве, которое можно назвать пространством сигналов.

Взяв в качестве координат отсчеты*) сигнала, мы добиваемся взаимной некоррелированности координат вектора сигнала. Размерность пространства сигналов равна при этом числу степеней свободы рассматриваемого сигнала. Как отмечалось в § 4, оценкой числа степеней свободы отрезка сигнала длительностью T и ограниченной шириной F спектра является число $2FT$.

Легко видеть, что в большинстве практических случаев число измерений пространства сигналов очень велико. Например, для телефонного разговора ($F = 10$ *кгц*) длительностью 10 *мин* пространство сигналов имеет 12000 измерений; телевизионная передача ($F = 5$ *мгц*) часовой длительности

*) Для краткости будем иногда называть значения функции в точках отсчета просто отсчетами.

представляется $3{,}6 \cdot 10^{10}$-мерной точкой. Хотя такие пространства не допускают наглядного изображения, аналитические соотношения геометрии значительно облегчают рассмотрение проблем связи.

Если считать, что все координатные оси взаимно перпендикулярны, то расстояния в пространстве сигналов получают очень наглядный смысл. Рассмотрим сначала расстояние d от начала координат до точки, изображающей сигнал:

$$d^2 = \sum x_n^2, \tag{6.1}$$

здесь x_n n-й отсчет и суммирование идет по всем степеням свободы. В силу ортогональности функций отсчета (см. § 4) интегрированием квадрата ряда Котельникова легко показать, что

$$\int_{-\infty}^{\infty} f^2(t)\,dt = \frac{1}{2F}\sum x_n^2. \tag{6.2}$$

Таким образом, если рассматривать $f(t)$ как динамический сигнал (например, колебания тока или напряжения), то из сравнения (6.1) и (6.2) следует, что квадрат расстояния от начала координат до данной точки есть энергия сигнала (выделившаяся на единичном сопротивлении), умноженная на $2F$:

$$d^2 = 2FE = 2FT \cdot P, \tag{6.3}$$

где P — средняя мощность сигнала за время T. Вообще квадрат расстояния между двумя точками пространства сигналов есть умноженный на $2FT$ квадрат разности между двумя соответствующими сигналами. Интересно также отметить, что скалярное произведение двух векторов в пространстве сигналов равно коэффициенту корреляции между двумя соответствующими сигналами, умноженному на $2FT$:

$$\sum f_n \cdot g_n = 2FT \cdot R = 2F\int_{-\infty}^{\infty} f(t)g(t)dt. \tag{6.4}$$

(Это соотношение легко доказывается при подстановке под интеграл разложений $f(t)$ и $g(t)$ в виде рядов Котельникова).

Пространство сигналов можно строить, не обязательно беря за координаты отсчеты сигнала. Можно, например, сигнал длительностью T разложить в ряд Фурье и в качестве координат взять коэффициенты при гармониках, следующих через $1/T$ *герц* друг за другом. Размерность пространства сигналов при этом не изменится, так как каждой частоте

соответствует два коэффициента (при синусе и косинусе). Более того, можно показать, что полученная таким образом система координат путем поворота может быть совмещена с координатной системой, построенной на отсчетах; следовательно, смысл расстояний остается тем же.

Простое геометрическое толкование имеют различные операции над сигналами. Например, если сигнал в канале искажается определенным образом, то пространство сигналов искривляется за счет определенного смещения каждой точки. Пропусканию сигнала через фильтр с полосой, меньшей ширины спектра, соответствует проектирование точки сигнала на некоторое подпространство, так как такая фильтрация уменьшает число степеней свободы сигнала. Наконец, сложение сигнала с помехой означает смещение точки сигнала на величину, пропорциональную среднеквадратичному значению помехи. Если помеха носит случайный характер, то она образует некоторую область неопределенности около каждой точки пространства сигналов. Действие приемников, рассчитанных на работу в условиях помех, основано на определенном разбиении пространства сигналов на области, связываемые с различными передаваемыми сигналами. Свойства таких приемников определяются способом разбиения пространства сигналов на области (в соответствующем разделе мы дадим этому более полное толкование).

Геометрическая модель позволяет дать наглядное изображение процессов, происходящих в линиях связи, и простые доказательства важным для теории связи теоремам.

§ 7. ДИНАМИЧЕСКИЙ СИГНАЛ КАК КОЛЕБАНИЕ СО СЛУЧАЙНЫМИ АМПЛИТУДОЙ И ФАЗОЙ

Понятие огибающей и фазы, введенные первоначально для гармонических колебаний, допускают обобщение на случай нерегулярных, случайных временных процессов. В связи с этим колебание со случайными амплитудой и фазой может служить во многих случаях как наглядная и удобная при вычислениях модель динамического сигнала. Такой моделью (при рассмотрении флуктуационных процессов) пользовался С. О. Райс [28]; В. И. Бунимович [12] очень подробно изучил свойства этой модели и развил способы ее применения для решения конкретных вопросов.

Согласно рассматриваемой модели реализация случайного процесса $u\ (t)$ представляется в виде

$$u\ (t) = E\ (t) \cdot \cos \Phi(t), \tag{7.1}$$

где $E\ (t)$ и $\Phi\ (t)$ — обобщенные «амплитуда» (огибающая) и «фаза», в свою очередь, являющиеся реализациями некото-

рых случайных процессов. Таким образом, процесс $u(t)$ представляется как синусоидальное колебание, случайно модулированное по амплитуде и фазе (или частоте). Особую наглядность приобретает такое представление в случае небольшой ширины спектра сигнала, когда (7.1) может быть переписано в виде

$$u(t) = E(t)\cos[\omega_0 t + \Theta(t)]; \qquad (7.2)$$

при этом $E(t)$ и $\Theta(t)$ являются медленно изменяющимися (по сравнению с $\cos\omega_0 t$) функциями времени.

При всей простоте такая модель позволяет довольно легко решать некоторые вопросы преобразования сигналов. Например, если нас интересует вопрос, как изменяется спектр при квадратичном детектировании такого сигнала, простые вычисления дают:

$$[u(t)]^2 = [E\cos\Phi]^2 = \frac{1}{2}E^2(t) + \frac{1}{2}E^2(t)\cos 2\Phi(t), \qquad (7.3)$$

откуда сразу же следует, что спектр на выходе детектора состоит из двух полос, соответствующих двум членам в правой части (7.3). Спектр функции $E^2(t)$ сосредоточен около нулевой частоты; вторая спектральная полоса лежит «в области удвоенных частот», вблизи $\omega = 2\omega_0$.

С помощью рассматриваемой модели можно изучать вопросы передачи сигналов при наличии шумов. Легко показать, что принимаемый сигнал, являющийся суммой полезного сигнала $S(t)\cos\omega_0 t$ и шума $N(t)\cos[\omega_0 t + \Theta(t)]$, можно представить в виде (7.2):

$$S(t)\cos\omega_0 t + N(t)\cos(\omega_0 t + \Theta(t)) =$$

$$= \sqrt{S^2 + N^2 + 2SN\cos\Theta}\cdot\cos\left(\omega_0 t + \operatorname{arctg}\frac{N\sin\Theta}{S + N\cos\Theta}\right). \qquad (7.4)$$

Для рассмотрения многих конкретных вопросов необходимо изучить статистические свойства огибающей и фазы результирующего сигнала (7.4). Найдем интересующие нас распределения вероятностей для случая весьма сильно упрощающих выкладки предположений. которые, однако, не являются слишком далекими от действительности:

1. Полезный сигнал s (t) модулирован только по амплитуде, его частота и начальная фаза постоянны:

$$s(t) = S(t)\cos\omega_0 t. \qquad (7.5)$$

2. Мешающий шум $n(t)$ является нормальным (гауссовским) процессом со стандартным отклонением σ.

3. Результирующий сигнал (смесь полезного сигнала с шумом) есть простая их сумма:

$$u(t) = s(t) + n(t). \tag{7.6}$$

4. Спектры сигнала и шума «узки», т. е. ω_0 значительно больше ширины этих спектров; для простоты будем считать их равными по ширине и совпадающими по положению на оси частот. Это позволяет записать $u(t)$ в виде

$$u(t) = R(t)\cos(\omega_0 t + \varphi(t)), \tag{7.7}$$

где $R(t)$ — мгновенное значение огибающей результирующего сигнала, ω_0 — несущая частота, $\varphi(t)$ — фаза результирующего колебания относительно полезного сигнала.

Воспользуемся широко известной в теории колебаний аналогией между векторами и синусоидальными колебаниями, которая появляется, если поставить в соответствие огибающую колебания и длину вектора, фазу колебания и угол, характеризующий направление вектора. При этом линейные операции над колебаниями соответствуют этим же операциям над векторами; в частности, суммарное колебание (7.6) изображается векторной суммой соответствующих векторов.

Выберем далее декартову систему координат с началом в точке $S = 0$, вращающуюся в положительном направлении углов с частотой ω_0; ось x направим по направлению вектора полезного сигнала. При этом вектор полезного сигнала имеет только x-овую составляющую. Гауссовость шума означает, что обе проекции вектора шума являются независимыми, нормально распределенными (с дисперсией σ^2 и средним нуль) случайными величинами. Проекции x и y результирующего колебания также являются независимыми, нормально распределенными величинами; только теперь среднее значение проекции на ось OX равно амплитуде полезного сигнала S. Отсюда следует, что совместное распределение вероятностей величин x и y можно записать в виде:

$$p(x, y)\,dx\,dy = \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{1}{2\sigma^2}[(x-S)^2 + y^2]\right\}dxdy. \tag{7.8}$$

Найдем теперь совместное распределение огибающей R и фазы φ смеси сигнала с шумом. Воспользовавшись очевидными соотношениями $x = R\cos\varphi$, $y = R\sin\varphi$, $dxdy = RdRd\varphi$, из (7.8) сразу получаем распределение

$$p(R, \varphi)dRd\varphi = \frac{R}{2\pi\sigma^2}\exp\left[-\frac{1}{2\sigma^2}(R^2 - 2RS\cos\varphi + S^2)\right]dRd\varphi, \tag{7,9}$$

полученное Д. Миддлтоном [23] несколько другим путем. Легко видеть, что если $S=0$ (т. е. рассматривается чистый шум), то

$$p(R, \varphi)dRd\varphi = \frac{R}{\sigma^2} e^{-\frac{R^2}{2\sigma^2}} \cdot dR \frac{d\varphi}{2\pi}, \qquad (7.10)$$

т. е. огибающая распределена по закону Релея, фаза—равномерно в интервале $(0,2\pi)$; R и φ независимы.

Распределения вероятностей отдельно для амплитуды и фазы сигнала с шумом получаются при интегрировании совместного распределения (7.9) по переменной, которую необходимо исключить.

Например, распределение огибающей будет:

$$p(R)dR = dR \int_0^{2\pi} p(R, \varphi)d\varphi =$$

$$= \frac{RdR}{\sigma^2} e^{-\frac{R^2+S^2}{2\sigma^2}} \frac{1}{2\pi} \int_0^{2\pi} e^{\frac{RS}{\sigma^2}\cos\varphi} d\varphi =$$

$$= \frac{R}{\sigma^2} e^{-\frac{R^2+S^2}{\sigma^2}} I_0\left(\frac{RS}{\sigma^2}\right) dR, \qquad (7.11)$$

здесь $I_0(x)$ — модифицированная функция Бесселя первого рода нулевого порядка. Это распределение называется обобщенным (или модифицированным) законом Релея, так как при $S \to 0$ оно переходит в обычное распределение Релея.

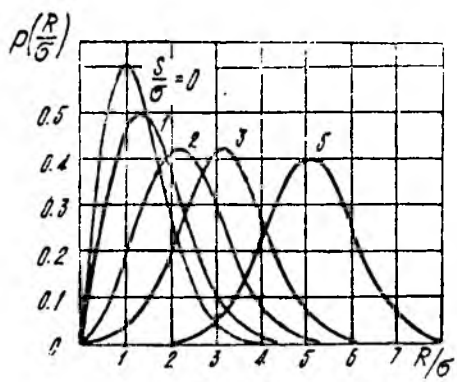

Рис. 12.

Семейство кривых $p(R)$ для различных S представлено на рис. 12.

Плотность распределения фазы находится интегрированием $p\ (R, \varphi)$ по R. В зависимости от того, какими соотношениями пользоваться в ходе вычислений, искомое распределение можно записать в виде [23]:

$$p(\varphi) = \frac{1}{2\pi} \exp\left(-\frac{S^2}{2\sigma^2}\sin^2\varphi\right)\left\{\sqrt{\frac{\pi}{2}}\,\frac{S}{\sigma}\cos\varphi + \right.$$

$$\left. + {}_1F_1\left(-\frac{1}{2}, \frac{1}{2}, \frac{S^2}{2\sigma^2}\cos^2\varphi\right)\right\},$$

либо в виде [9]

$$p(\varphi) = \frac{1}{2\pi}\, e^{-\frac{S^2}{2\sigma^2}}\left[1 + \right.$$

$$\left. + \sqrt{\frac{\pi}{2}}\,\frac{S}{\sigma}\cos\cdot e^{\frac{S^2}{2\sigma^2}\cos\varphi}\left[\Phi\left(\frac{S^2}{2\sigma^2}\cos\varphi\right) + 1\right]\right]. \quad (7.13)$$

Соответствующие кривые приведены на рис. 13.

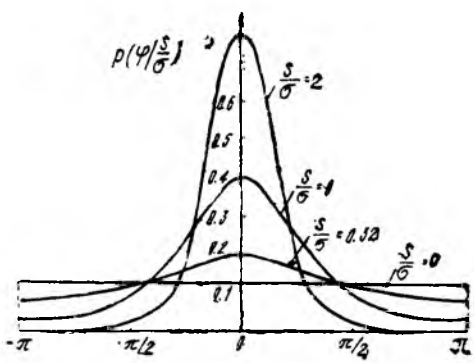

Рис. 13.

Полученные распределения позволяют пользоваться описываемой в этом параграфе моделью сигнала при количественном рассмотрении многих вопросов. Иногда, однако, помимо этих распределений необходимо знание распределений огибающей (или фазы), более детально описывающих сигнал, например, при условии, что известно, какое значение имела огибающая (или фаза) τ *сек* назад, или при учете коррелированности R и φ, и т. п. В случае необходимости читатель может найти соответствующие распределения в ра-

ботах С. Райса [28], В. И. Бунимовича [13], Д. Миддлтона [23] и др.

§ 8. ГАРМОНИЧЕСКИЙ АНАЛИЗ СИГНАЛОВ

Гармонический анализ временных функций, т. е. представление таких функций в виде рядов или интегралов Фурье, и вытекающее отсюда представление о частотных спектрах настолько широко и эффективно употребляются*), что, естественно, возникает необходимость использования гармонического анализа при изучении некоторых типов динамических сигналов. Однако такие сигналы являются случайными функциями времени, что и приводит к ряду особенностей спектрального представления сигналов.

Основная особенность и трудность связана с тем, что иногда необходимо выразить спектральные свойства ансамбля возможных сигналов в целом, а не каждого из них в отдельности. Получение спектра комплексных амплитуд каждой реализации в отдельности в принципе не представляет трудности. Но при этом мы получим столько же различных спектров, сколько реализаций образует ансамбль: преобразование Фурье обратимо, поэтому две различных функции времени не могут иметь одинаковых спектров.

Таким образом, случайному временному процессу с помощью Фурье-преобразования можно однозначно поставить в соответствие случайный процесс в частотной области, т. е. ансамбль спектров реализаций случайного процесса. В некоторых случаях более удобно рассматривать статистические свойства ансамбля спектров. При этом часто оказывается достаточным рассмотрение лишь некоторых суммарных характеристик, подобно тому, как при описании временных процессов часто достаточно рассматривать лишь некоторые моменты распределений, а не сами распределения.

Каким же образом осуществить такое «суммарное» описание, не чувствительное к некоторым индивидуальным особенностям спектров реализаций, но вскрывающее то общее**), что имеется между ними? Первым шагом к стиранию индивидуальности при спектральном описании является переход от спектров комплексных амплитуд к спектрам мощности реализаций. Спектр мощности нечувствителен к изменениям фазовых соотношений между компонентами, поэтому все

---

*) Детальное рассмотрение вопросов спектрального описания временных функций, в том числе и случайных, можно найти в книге А. А. Харкевича [36].

**) Прежде всего должно существовать то общее, что мы собираемся описывать; поэтому в дальнейшем будем иметь в виду только эргодические сигналы, заведомо обладающие общими свойствами.

реализации, отличающиеся только фазами (но не амплитудами!) гармоник имеют одинаковый спектр мощности. Это, конечно, не означает, что все реализации эргодического процесса вообще имеют одинаковые спектры мощности. В общем случае мы снова получим ансамбль спектров мощности.

Рассмотрим одно важное свойство спектра мощности. Спектр мощности $G_x(\omega)$ получается из спектральной функции $S_x(\omega)$ реализации $x(t)$ путем нахождения квадрата модуля функции $S_x(\omega)$:

$$G_x(\omega) = |S_x(\omega)|^2. \tag{8.1}$$

Функцию частоты $G_x(\omega)$ можно в свою очередь рассматривать как спектральную функцию некоторой временной функции $R_x(t)$. Представляет интерес выяснить, в каком соотношении находятся функции $x(t)$ и $R_x(t)$. Проделаем последовательно ряд преобразований, позволяющих связать $x(t)$ и $R_x(t)$ в явном виде:

$$R_x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} G_x(\omega)\, e^{i\omega t}\, d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} |S_x(\omega)|^2 e^{i\omega t} d\omega =$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_x(\omega)\,[S_x^*(\omega)\, e^{i\omega t}]\, d\omega =$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} S_x(\omega)\left[\int_{-\infty}^{\infty} x(t_1 - t) e^{i\omega t_1} dt_1\right] d\omega =$$

$$= \int_{-\infty}^{\infty} x(t_1 - t)\left[\frac{1}{2\pi}\int_{-\infty}^{\infty} S_x(\omega) e^{i\omega t} d\omega\right] dt_1 =$$

$$= \int_{-\infty}^{\infty} x(t_1 - t)\, x(t_1) dt_1. \tag{8.2}$$

Таким образом, функция $R_x(t)$ оказывается пропорциональной временной функции корреляции, вычисленной по реализации $x(t)$*).

Поскольку $R_x(t)$ есть функционал от реализации $x(t)$

*) То, что для некоторого класса случайных процессов интеграл (8.2) оказывается бесконечным, не лишает вывода общности. В этом случае достаточно рассмотреть предел, к которому будет стремиться отношение этого интеграла в конечном интервале к самому интервалу при стремлении последнего к бесконечности.

случайного процесса $X(t)$, то и само $R_x(t)$ есть реализация случайного процесса. Однако теоретические соображения, в полном согласии с экспериментами, приводят к выводу о том, что для эргодического процесса $X(t)$ функции корреляции отдельных реализаций «очень похожи», «близки» друг к другу, их различие носит характер случайных флуктуаций. Это является одним из проявлений закона больших чисел: подобно средним значениям, вычисленным по различным выборкам, функции корреляции, с одной стороны, содержат то общее, что есть в различных реализациях, а с другой стороны, принципиально не могут быть полностью лишены индивидуального отпечатка той реализации, по которой они вычислены.

Возвращаясь к исходной проблеме нахождения единой спектральной характеристики процесса $X(t)$ в целом, мы приходим к выводу о том, что такую характеристику можно найти двумя путями. Первый путь состоит в том, чтобы найти «среднюю функцию корреляции», около которой «флуктуируют» функции корреляции отдельных реализаций, а затем принять в качестве спектра мощности процесса в целом спектр средней функции корреляции. Второй способ — найти сразу средний спектр мощности, усреднив по ансамблю спектры мощности отдельных реализаций.

Иногда (а в практике — довольно часто) для стационарных эргодических процессов, рассмотренных на достаточно большом интервале времени, в качестве оценки спектра процесса в целом можно использовать спектр наблюденной реализации, подобно тому, как пользуются статистическими средними вместо математических ожиданий. Необходимо только ясно отдавать себе отчет, что при этом принципиально неизбежны «погрешности» флуктуационного характера, и при количественных измерениях необходимо оценивать эти погрешности наряду с остальными ошибками опыта.

§ 9. ЧАСТОТНО-ВРЕМЕННАЯ НЕОПРЕДЕЛЕННОСТЬ СИГНАЛОВ

В строгом классическом изложении спектральная функция (или просто «спектр») некоторого сигнала есть однозначное преобразование функции времени в функцию частоты. Заданной функции времени соответствует вполне определенная функция частоты и обратно; связь между ними определяется парой Фурье-преобразований. Сигнал может, следовательно, рассматриваться л и б о как функция времени, либо как функция частоты. При этом между масштабами во временной и частотной областях из свойств преобразования Фурье следует определенная связь [36]. Рассмотрим функцию

времени s (t), имеющую спектр S (ω). Изменим масштаб по оси времени в a раз и найдем спектр функции s (at):

$$S_a(\omega) = \int s(at)\, e^{-i\omega t}\, dt = \frac{1}{a} S\left(\frac{\omega}{a}\right). \qquad (9.1)$$

Множитель $\frac{1}{a}$ связан с ~~уменьшением полной энергии~~ изменением мощности сигнала, а вид функции $S_a(\omega)$ подобен виду функции $S(\omega)$. Существенно то, что для получения $S_a(\omega)$ из $S(\omega)$ необходимо изменить масштаб по оси частот обратно пропорционально a.

Для преобразования Фурье характерно также, что если сигнал имеет ограниченную длительность, то его спектр неограничен и, наоборот, сигнал с ограниченным спектром длится бесконечно долго. Однако практика указывает на то, что даже для ограниченных во времени сигналов всегда можно указать область частот, вне которой лежат спектральные компоненты, совершенно несущественно влияющие на информационные свойства сигналов. В связи с этим возникает вопрос: нельзя ли ввести в рассмотрение сигналы, обладающие ограниченным спектром и одновременно ограниченной длительностью? Формализм преобразований Фурье не допускает этого, но, не желая отказаться от удобного и широко распространенного аппарата Фурье-преобразований, можно все же ввести разумные допущения, позволяющие считать сигнал ограниченным и по спектру, и по длительности.

Оставаясь в рамках теории Фурье, такое ограничение можно провести лишь приближенно. Разумным критерием при этом является энергетический критерий: сигнал считается ограниченным по длительности, если можно указать интервал времени, в котором сосредоточена основная часть всей энергии сигнала; это не исключает ограниченности ширины спектра сигнала, которая определяется так же, как область частот, в которой сосредоточена часть энергии сигнала.

В силу того, что длительность сигнала и ширина его спектра определяются условно, сигнал теперь нужно рассматривать не просто как функцию времени **или** частоты, а как функцию времени **и** частоты. Так мы приходим к необходимости рассмотрения представления динамических сигналов в частотно-временной области. На плоскости «частота-время» ((f, t)-плоскости) в общем случае сигнал занимает бесконечную площадь. Энергетический критерий позволяет в некоторых случаях связать с заданным сигналом участок

(f,t)-плоскости, имеющий ограниченную площадь (см. рис. 14).

Очевидно, чем меньше занимаемая сигналом площадь на (f, t)-плоскости, тем выгоднее этот сигнал для передачи, т. е. одновременно сокращается и время передачи, и полоса занимаемых частот. Отсюда вытекает постановка задачи о нахождении сигнала с минимальной площадью на плоскости (f, t). Варьировать при этом нужно форму сигнала в целом, так как варьирование лишь длительности сигнала, согласно соотношению масштабов, дает сохранение площади. Майер

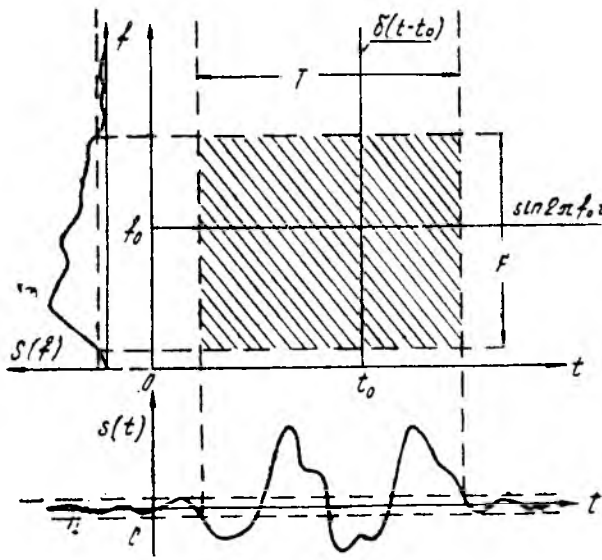

Рис. 4.

и Леонтович [22], а позднее Габор [15] показали, что свойством минимальности (f, t)-площади обладают синусоидальные колебания, модулированные по амплитуде гауссовой кривой («колокольным» импульсом, см. рис. 15); такой сигнал в комплексной форме записывается как

$$s(t) = e^{-\alpha^2(t-t_0)^2} e^{2\pi i f_0 t}. \tag{9.2}$$

Характерно, что Фурье-преобразование такого колебания выражается подобной функцией:

$$S(f) = e^{-\left(\frac{\pi}{\alpha}\right)^2 (f-f_0)^2} e^{-2\pi i f_0 t}. \tag{9.3}$$

Таким образом, исследования показывают, что (f, t)-площадь сигналов можно сжимать лишь до некоторого предела, что

может быть выражено соотношением

$$\Delta t \cdot \Delta f \geqslant \text{const} > 0, \tag{9.4}$$

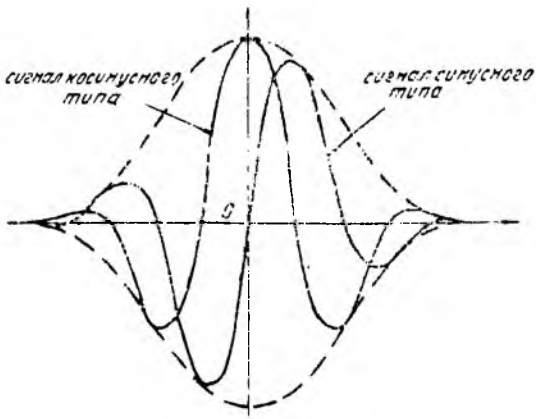

Рис. 15.

где Δt — длительность, Δf — ширина спектра сигнала. По аналогии с принципом неопределенности в квантовой механике, соотношение (9.4) называется принципом частотно-временной неопределенности сигналов.

§ 10. СИГНАЛЫ С ОГРАНИЧЕННЫМ СПЕКТРОМ КАК ПРИБЛИЖЕННАЯ МОДЕЛЬ СИГНАЛОВ С НЕОГРАНИЧЕННЫМ СПЕКТРОМ

Модель сигнала с ограниченным спектром может быть подвергнута критике по многим пунктам. Начать с того, что все реальные сигналы либо вообще существуют ограниченное время, либо, существуя практически бесконечно долго, воздействуют на регистрирующее устройство лишь в течение конечного интервала времени (как, например, свет в фотографии). В первом случае спектр сигнала в принципе неограничен; во втором — переходные процессы также не позволяют ограничить область частотных компонент, участвующих в процессе регистрации сигнала. С другой стороны, сигналы, возникающие в результате ударного возбуждения линейных систем, в силу физических причин имеют неограниченные спектры, что является прямым следствием требования физической реализуемости линейной системы. Имеются и другие веские основания считать, что абсолютно точно ограниченные спектры не соответствуют физической реальности; они вытекают из теории предсказания случайных процессов [20].

В связи с этим возникает необходимость ввести в рассмотрение такую математическую модель сигнала, которая

не накладывала бы строгого ограничения на ширину его спектра. Вместе с тем, совершенно реальным физическим фактом (который, кстати, и послужил основой для выдвижения представления об ограниченности спектра) является то, что любой заданной (меньшей единицы) доле полной энергии сигнала соответствует вполне ограниченный интервал спектра частот. Таким образом, спектр мощности реального сигнала достаточно быстро спадает вне интервала частот, на который приходится основная часть мощности. В связи с этим естественными являются попытки рассмотреть модель сигнала с ограниченным спектром как приближение к модели сигнала с неограниченным спектром. Этот вопрос исследован в ряде работ. Здесь мы кратко изложим одно из рассмотренных И. Т. Турбовичем [31—33] аналитических представлений сигналов с неограниченным спектром.

Пусть известно, что сигнал $f_0(t)$ имеет неограниченный спектр, но основная часть энергии сосредоточена в интервале частот $(0, \omega_c)$. Поэтому в качестве нулевого приближения $\varphi_0(t)$ естественно выбрать ряд Котельникова, коэффициентами которого будут отсчеты аппроксимируемой функции

$$\varphi_0(t) = \sum_{-\infty}^{\infty} f_0(k\tau)\,\frac{\sin\omega_c(t-k\tau)}{\omega_c(t-k\tau)}, \qquad (10.1)$$

здесь отсчеты следуют через интервал $\tau = \frac{\pi}{\omega_c}$.

В качестве следующего (первого) приближения нужно аппроксимировать функцию $f_1(t)$, равную разности $f_0(t)$ и $\varphi_0(t)$: $f_1(t) = f_0(t) - \varphi_0(t)$. Для обеспечения повышения точности аппроксимации необходимо повысить частоту отсчетов; удобно взять ее в два раза большей. Это будет соответствовать ряду Котельникова по функциям отсчета с максимальной частотой $2\omega_c$:

$$\varphi_1(t) = \sum_{-\infty}^{\infty} f_1\left(k\,\frac{\tau}{2}\right)\frac{\sin 2\omega_c\left(t-k\,\frac{\tau}{2}\right)}{2\omega_c\left(t-k\frac{\tau}{2}\right)}, \qquad (10.2)$$

так как $\varphi_0(k\tau) = f_0(k\tau)$, то $f_1\left(2k\,\frac{\tau}{2}\right) = 0$, т. е. все четные члены в (10.2) равны нулю. Далее необходимо представить в виде ряда Котельникова функцию $f_2(t) = f_1(t) - \varphi_1(t)$, взяв при этом отсчеты еще вдвое более частыми; и т. д. до бесконечности. Окончательно будем иметь:

$$f_0(t) = \sum_{n=0}^{\infty} \sum_{k=-\infty}^{\infty} f_n\left(k\frac{\tau}{2^n}\right) \frac{\sin 2^n \omega_c \left(t - k\frac{\tau}{2^n}\right)}{2^n \omega_c \left(t - k\frac{\tau}{2^n}\right)}, \qquad (10.3)$$

где

$$\begin{cases} f_{n+1}(t) = f_n(t) - \sum_{k=-\infty}^{\infty} f_n\left(k\frac{\tau}{2^n}\right) \frac{\sin 2^n \omega_c \left(t - k\frac{\tau}{2^n}\right)}{2^n \omega_c \left(t - k\frac{\tau}{2^n}\right)}, \\ f_n\left(2n\frac{\tau}{2^n}\right) = 0. \end{cases} \qquad (10.4)$$

Такое представление легко обобщить на случай, когда каждое следующее приближение получается при увеличении частоты следования отсчетов не вдвое, а в Q раз. Легко показать, что формулы (10.3) сохраняют тот же вид, только двойку нужно заменить на Q. Исследования И. Т. Турбовича [33] показали, что ряд (10.3) достаточно быстро сходится и является удобным для практических приближенных вычислений.

Более углубленный анализ модели сигнала с ограниченным спектром как приближенной модели сигнала с неограниченным спектром дан А. Н. Колмогоровым [20]. Этот анализ опирается на рассмотрение энтропийных характеристик случайного процесса и поэтому может быть рассмотрен лишь в последующих главах.

§ 11. О ВОЗМОЖНОСТИ ПОЛНОГО ОТКАЗА ОТ МОДЕЛИ СИГНАЛОВ С ОГРАНИЧЕННЫМ СПЕКТРОМ

В противоположность стремлению рассмотреть модель сигнала с ограниченным спектром как удобное упрощение или приближение, можно выступить с более решительных позиций полного отказа от предположения об ограниченности спектра и стационарности сигнала. Между тем именно благодаря этим двум предположениям теория структуры таких сигналов является простой и изящной. Появляются опасения, что переход к новой, возможно, более совершенной модели значительно усложнит аппарат теории сигналов, и его практическое применение натолкнется на серьезные затруднения.

В связи с этим Н. А. Железнов [5, 16, 17] выполнил ряд исследований предложенной им модели нестационарных сигналов с неограниченным спектром и показал, что при введении некоторых разумных ограничений теория хотя и несколько усложняется, но остается достаточно пригодной для практического использования.

Модель сигнала, развиваемая Н. А. Железновым, имеет следующие свойства:

1. Сигналы рассматриваются как нестационарный случайный процесс.

2. Длительность сигналов T конечна.

3. Энергетический спектр сигналов сплошной и отличен от нуля на всех частотах (за исключением, быть может, полосы меры нуль).

4. Интервал корреляции τ_0 ограничен, причем $\tau_{0\max} \ll T$. (Интервалом корреляции называется промежуток времени, за который полностью затухают корреляционные связи между отдельными частями сигнала).

В такой модели число степеней свободы сигнала длительностью T определяется величиной $N = T\,\tau_0$, если τ_0 — интервал корреляции, одинаковый для всех элементов сигнала. Нестационарность может выражаться зависимостью вида функции корреляции от времени.

Основным пунктом в теории структуры непрерывных сигналов является представление их через элементарные функции времени, т. е. принцип дискретизации. Математическое выражение принципа дискретизации дается соотношением:

$$f(t) \sim \Sigma \varphi_k \cdot u(t - t_k), \qquad (11.1)$$

которое утверждает, что непрерывной функции $f(t)$ может быть поставлена в соответствие взвешенная сумма известных функций времени $u(t)$, отнесенных к некоторым моментам времени t_k. В случае модели Н. А. Железнова эти моменты следуют через интервалы корреляции; коэффициенты разложения, φ_k, определяются значениями $f(t)$ в некоторых интервалах $T_k \leqslant T$; координатные функции разложения, $u(t-t_k)$, выбираются так, чтобы они не перекрывались друг с другом. (Это нужно для того, чтобы «будущие» значения не определяли «прошлого» сигнала).

Основной результат, получающийся при этом [17], сводится к тому, что дискретизация непрерывного сигнала в принципе приводит лишь к п р и б л и ж е н н о м у представлению его. Символ $\sim$ соответствия в (11.1) нужно понимать лишь как знак приближенного, а не точного равенства.

Н. А. Железновым установлено следующее важное и общее положение: дискретизация непрерывных сигналов, обладающих неограниченным спектром, с помощью физически реализуемых линейных фильтров*) (т. е. разложение по

*) Одно из условий физической реализуемости линейной системы может быть выражено как требование, чтобы отклик системы на внешнее возбуждение не начинался раньше самого возбуждения, поступающего на вход системы.

откликам таких фильтров) не может быть произведена со сколь угодно высокой точностью; существует предельная точность (верность) разложения, которая не может быть превзойдена никаким подбором коэффициентов и координатных функций (удовлетворяющих условиям физической реализуемости).

Это положение, на первый взгляд, «подрывает» основы техники передачи непрерывных сигналов методом импульсной модуляции, так как при этом в силу самого принципа дискретизации всегда будет происходить искажение передаваемого сигнала даже при отсутствии в линии связи помех и нелинейных искажений. Этот вопрос еще требует дальнейшего изучения, но, по-видимому, здесь вскрылось в явном виде противоречие, о котором шла речь в § 2 гл. 2: мы изображаем непрерывный сигнал как точную функцию времени $f(t)$, тогда как в действительности любая аппаратура, преобразующая сигналы (в том числе и «аппаратура естественная», т. е. органы чувств животных), никогда не фиксирует сигнал с бесконечной точностью.

§ 12. МАТЕМАТИЧЕСКОЕ ОПИСАНИЕ НЕКОТОРОГО КЛАССА НЕСТАЦИОНАРНЫХ СИГНАЛОВ

Можно показать, что аппарат теории эргодических и стационарных процессов накладывает на исследуемые сигналы такие ограничения, которым удовлетворяют далеко не все встречающиеся в практике сигналы. Рассмотрим, например, широко применяемые в технике связи амплитудно-модулированные колебания. Такие колебания представляются обычно в виде

$$x(t) = |A + f(t)| \cos \Omega t, \qquad (12.1)$$

где $|A + f(t)|$ — изменяющаяся во времени огибающая („амплитуда") колебаний несущей частоты Ω.

Отсюда сразу видно, что даже если $f(t)$ является стационарным процессом, распределение вероятностей, среднее по ансамблю, и корреляционная функция (вычисленная усреднением по ансамблю) процесса $x(f)$ будут зависеть от t. Следовательно, AM-колебание не является стационарным процессом.

Возникает необходимость такого описания сигналов, которое охватывало бы и нестационарные (в смысле теории случайных процессов) сигналы. Попытка построения такого описания сделана в работе В. В. Фурдуева [34].

Исходный пункт обобщенного подхода состоит в том, чтобы ввести количественные характеристики, достаточно полно описывающие как стационарные сигналы, так и некоторый класс нестационарных сигналов.

В качестве таких характеристик предлагается использовать текущие моменты реализации $x(t)$, вычисленные усреднением по интервалу времени T:

$$m(t, T) = \frac{1}{T}\int_{t-\frac{T}{2}}^{t+\frac{T}{2}} x(\xi)\, d(\xi), \tag{12.2}$$

$$R(\tau, t, T) = \frac{1}{T}\int_{t-\frac{T}{2}}^{t+\frac{T}{2}} x(\xi)\, x(\xi - \tau)\, d\xi. \tag{12.3}$$

$m(t, T)$ характеризует среднее значение сигнала на интервале T, $R(\tau, t, T)$ — степень когерентности сигнала и его запаздывающего на τ *сек* повторения; $R(o, t, T)$ дает величину средней мощности сигнала в интервале T.

Назовем структурно-однородными такие сигналы, для которых при возрастании T величины (12.2) и (12.3) стремятся к некоторым предельным значениям.

Понятие однородности есть обобщение понятия стационарности. Стационарные сигналы удовлетворяют условию однородности; однако этому же условию удовлетворяют некоторые типы нестационарных сигналов. Например, однородными являются все периодические сигналы

$$x = \sum_{k=1}^{\infty} C_k \cos(k\omega t + \varphi_k), \tag{12.4}$$

так как для них существует предельное значение

$$R(\tau) = \lim_{T\to\infty} R(\tau, T) = \sum_{k=1}^{\infty} \frac{1}{2} C_k^2 \cos k\omega t. \tag{12.5}$$

Амплитудно-модулированные колебания, не являющиеся стационарными даже если модулирующее колебание стационарно, будут однородными при однородности модулирующего сигнала. Покажем это. Пусть

$$x(t) = [A + f(t)] \sin \Omega t, \tag{12.6}$$

тогда

$$R(\tau, T) = \frac{1}{2} r(\tau, T) \cos \Omega t -$$

$$-\frac{1}{2T}\left\{\left[\int_{-\frac{T}{2}}^{\frac{T}{2}} F(t)\cos 2\Omega\, t dt\right]\cos\Omega\tau + \right.$$

$$\left. + \left[\int_{-\frac{T}{2}}^{\frac{T}{2}} F(t)\sin 2\Omega\, t dt\right]\sin\Omega\tau\right\}, \qquad (12.7)$$

где

$$r(\tau, T) = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} [A + f(t)][A + f(t-\tau)]\, d\tau,$$

$$F(t) = [A + f(t)][A + f(t-\tau)].$$

Заметим, что величины, взятые в (12.7) в квадратные скобки, при $T \to \infty$ дают значение спектральной плотности функции $F(t)$ на частоте 2Ω. Так как обычно ширина спектра модулирующего колебания значительно меньше несущей частоты Ω, то эти величины равны нулю или, во всяком случае, ограничены. Следовательно, существует предельное значение функции автокорреляции

$$R(\tau) = \lim_{T\to\infty} R(\tau, T) = \frac{1}{2} r(\tau)\cos\Omega\tau =$$

$$= \frac{A^2}{2}\cos\Omega\tau + \frac{1}{2} r_f(\tau)\cos\Omega\tau, \qquad (12.8)$$

где $r_f(\tau)$ — функция автокорреляции модулирующего сигнала. Если этот сигнал однороден, то и AM-сигнал тоже однороден, что и требовалось показать.

С точки зрения практических применений теории весьма важно, что между функцией автокорреляции $R(\tau)$ и спектром $\Phi(\nu)$ средней мощности сигнала сохраняется та же связь, что и в теории стационарных процессов:

$$\begin{cases} \Phi(\nu) = 4\int_0^\infty R(\tau)\cos\omega\tau\, d\tau, \\ R(\tau) = \int_0^\infty \Phi(\nu)\cos\omega\tau\, d\nu, \end{cases} \qquad (12.9)$$

$\omega = 2\pi\nu$. Предоставив провести самому строгое доказательство этих соотношений, интересующемуся читателю покажем, что они приводят к физически правильным результатам в применении к случаю заведомо нестационарного, но однородного процесса — AM-сигнала. Подставим функцию автокорреляции AM-сигнала (12.8) в (12.9) и вычислим спектр.

$$\Phi(\nu) = 2A^2\int_0^\infty \cos\Omega\tau\cdot\cos\omega\tau\cdot d\tau + 2\int_0^\infty r_f(\tau)\cos\Omega\tau\cdot\cos\omega\tau\cdot d\tau =$$

$$= \frac{A^2}{2}\delta(\omega-\Omega) + \int_0^\infty r_f(\tau)\cos(\Omega+\omega)\tau\cdot d\tau +$$

$$+ \int_0^\infty r_f(\tau)\cos(\Omega-\omega)\tau\, d\tau. \qquad (12.10)$$

Этот результат совпадает с общеизвестным представлением о спектре АМ-сигнала, который (спектр) состоит из колебания несущей частоты Ω с мощностью $A^2/2$, а по обе стороны несущей располагаются боковые полосы со смещенными спектрами модулирующего сигнала.

В. С. Пугачев предложил [27] иной способ анализа нестационарных процессов, аналогичных AM-колебаниям. Он показал, что метод канонических разложений (см. § 3) применим и к нестационарным процессам вида

$$y(t) = f(t)\cdot x(t) + g(t), \qquad (12.11)$$

где $x(t)$ – стационарный процесс, а $f(t)$ и $g(t)$ – известные неслучайные функции времени. Случайные процессы такого типа В. С. Пугачев предложил называть приводимыми к стационарным.

Покажем, как можно построить каноническое разложение нестационарного процесса (12.11). Математическое ожидание, корреляционная функция и дисперсия процесса $y(t)$, вычисленные усреднением по ансамблю, определяются формулами:

$$m_y(t) = f(t)\cdot m_x + g(t), \qquad (12.12)$$

$$K_y(t,t') = f(t)\cdot f(t')\cdot k_x(t-t'), \qquad (12.13)$$

$$D_y(t) = K_y(t,t) = f^2(t)\cdot k_x(0), \qquad (12.14)$$

где $k_x(\tau)$ — корреляционная функция стационарного процесса $x(t)$. На основании этих формул нормированная корреляционная функция процесса $y(t)$ равна:

$$R_y(t, t') = \frac{K_y(t, t')}{\sqrt{D_y(t) \cdot D_y(t')}} = \frac{k_x(t - t')}{k_x(0)} = r_x(t - t'), \quad (12.15)$$

т. е. совпадает с нормированной корреляционной функцией стационарного процесса $x(t)$. Это может служить признаком приводимости нестационарного процесса к стационарному.

Поскольку $k_x(\tau) = r_x(\tau) \cdot \text{const}$, координатными функциями разложения $k_x(\tau)$ являются синусы и косинусы (см. § 3). Следовательно, координатные функции разложения $K_y(t, t')$ запишутся как $v_k(t) = f(t) \cos \omega_k t$ и $w_k(t) = f(t) \sin \omega_k t$. Отсюда, в силу теоремы Пугачева, искомое каноническое разложение для $y(t)$ выразится в виде

$$y(t) = m_y(t) + \sum_{k=0}^{\infty} [V_k \cdot v_k(t) + W_k \cdot w_k(t)]. \quad (12.16)$$

С помощью данного разложения по обычным рецептам метода канонических разложений можно изучать изменение статистических характеристик процесса при его преобразованиях.

Часть II

ЭНТРОПИЯ, ИНФОРМАЦИЯ, КОЛИЧЕСТВО ИНФОРМАЦИИ

ГЛАВА IV

ЭНТРОПИЯ

§ 1. ВВЕДЕНИЕ

Рассматривая свойства сигналов (см. ч. I), мы пришли к выводу о том, что только случайные объекты (точнее, их состояния) могут служить в качестве носителей информации. Поэтому теория вероятностей и связанные с нею дисциплины (в особенности, теория случайных процессов, математическая статистика, исследование операций, теория игр и решений и т. д.) служат основой, на которой развивается теория информации. С другой стороны, понятия и методы теории информации оказывают влияние на развитие этих «базисных» наук. К числу таких понятий, выдвинутых теорией информации и имеющих выходящее за рамки этой теории значение, относится представление об энтропии случайного объекта.

Понятие энтропии возникло в связи с необходимостью ввести численную характеристику неопределенности случайного объекта на некотором этапе его рассмотрения. Все, что мы можем сказать априори о поведении случайного объекта, это указать множество его состояний и указать распределение вероятностей по элементам этого множества. Отложим пока обсуждение принципиально важного вопроса о том, на основании чего мы получаем возможность указать распределение и в каком соотношении оно находится с объективной реальностью. Обратим внимание на то, что различные распределения с различной неопределенностью характеризуют, какое из возможных состояний объекта должно реализоваться. Например, пусть некоторый объект имеет два возможных

состояния, A_1 и A_2; пусть при одних условиях распределение вероятностей характеризуется числами $p(A_1)=0{,}99$, $p(A_2)=0{,}01$; а в другом случае — $p(A_1)=p(A_2)=0{,}5$. Очевидно, что в первом случае результатом опыта «почти наверняка» будет реализация состояния A_1, во втором же случае неопределенность так велика, что естественно воздержаться от всяких прогнозов.

Желая сравнить между собой два (или более) распределения по их «размытости», неопределенности, мы должны ввести некоторую численную характеристику этого качества распределений. Если случайный объект допускает численное описание, т. е. его состояниям соответствуют некоторые количества, то в качестве числовых характеристик формы распределения могут служить различные средние, например, среднее значение, дисперсия, моменты высших порядков. Однако эти характеристики теряют всякую наглядность и удобство применения, если распределения являются резко асимметричными, многовершинными и т. п. И, наконец, моменты вообще теряют смысл, если случайный объект допускает лишь качественное описание, т. е. различным его «состояниям» соответствуют различные качества. Если взять, например, в качестве случайного параметра профессию (или национальность) человека из некоторой группы людей, то можно говорить о вероятности встретить железнодорожника (или, к примеру, украинца), но не имеет смысла вычислять «среднюю профессию» или дисперсию соответствующего распределения. Вместе с тем разнообразие признаков и в случае качественного объекта допускает количественную оценку, поскольку мы и здесь можем говорить о большей или меньшей неопределенности исхода опыта.

Так мы приходим к необходимости количественного описания неопределенности заданного распределения вероятностей. Понятие «неопределенность» естественно связывается с формой распределения, но не с множеством конкретных значений случайной величины. Поэтому первое требование к мере неопределенности состоит в том, что она должна быть функционалом, т. е. функцией от функций распределения вероятностей, и не зависеть от конкретных значений случайной величины. Кроме того, к мере неопределенности должен быть предъявлен еще целый ряд требований, таких как непрерывность относительно аргументов, наличие максимума и дополнительные требования, которые более подробно будут рассмотрены ниже. Важно подчеркнуть, что такой комплекс разумно выдвинутых требований к мере неопределенности допускает единственную форму функционала, который по ряду причин, подлежащих отдельному обсуждению, и назван энтропией случайного объекта.

§ 2. ЭНТРОПИЯ СЛУЧАЙНЫХ ОБЪЕКТОВ С ДИСКРЕТНЫМ МНОЖЕСТВОМ СОСТОЯНИЙ [7], [47].

В качестве меры неопределенности случайного объекта с конечным множеством возможных состояний A_1, A_2,..., A_n с соответствующими вероятностями p_1, p_2,..., p_n *) разумно взять функционал

$$H(A) = H(p_1, p_2, \ldots, p_n) = -\sum_{k=1}^{n} p_k \log p_k, \tag{2.1}$$

где логарифмы берутся при произвольном основании. Величину $H(A)$ называют энтропией случайного объекта A. В последующем параграфе мы докажем, что при некоторых весьма общих требованиях к мере неопределенности вид (2.1) функционала H является единственно возможным (с точностью до постоянного множителя). Здесь же мы рассмотрим особенности функции $H(p_1, p_2, \ldots, p_n)$ и убедимся, что она действительно обладает рядом свойств, которых мы склонны требовать от разумно заданной меры неопределенности конечной схемы.

1. $H(p_1, p_2, \ldots, p_n) = 0$ в том и только в том случае, когда из чисел $p_1, p_2, \ldots, p_n$ какое-нибудь одно равно единице (а следовательно, все остальные — нули). Это соответствует случаю, когда исход опыта может быть предсказан с полной достоверностью и когда отсутствует всякая неопределенность. Во всех других случаях энтропия положительна.

2. Естественно потребовать, чтобы при $p_1 = p_2 = \ldots = p_n = \frac{1}{n}$, т. е. в случае наибольшей неопределенности, функция H достигала наибольшего значения. Убедимся, что $H(p_1, p_2, \ldots, p_n)$ удовлетворяет этому требованию. Для всякой выпуклой функции $\varphi(x)$ имеет место неравенство

$$\varphi\left(\frac{1}{n}\sum_{k=1}^{n} a_k\right) \leqslant \frac{1}{n}\sum_{k=1}^{n} \varphi(a_k),$$

где $\{a_i\}$ — любые положительные числа. Полагая $a_k = p_k$ и $\varphi(x) = x \log x$, находим (учтя, что $\sum_k p_k = 1$):

$$\varphi\left(\frac{1}{n}\right) = \frac{1}{n} \log \frac{1}{n} \leqslant \frac{1}{n}\sum_{k=1}^{n} p_k \log p_k =$$

---

*) Иногда для кратности будем говорить, что задана конечная схема, если задано множество состояний $A_1, A_2, \ldots, A_n$ случайного объекта вместе с соответствующими вероятностями $p_1, p_2, \ldots, p_n$.

$$\frac{1}{n} H(p_1, p_2, \ldots, p_n).$$

Отсюда сразу следует, что

$$H(p_1, p_2, \ldots, p_n) \leqslant \log n = H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right),$$

что и требовалось доказать.

3. Пусть имеется два независимых случайных объекта A и B с числом состояний n и m соответственно. Естественно ожидать, что неопределенность случайного объекта AB, состояния которого образуются совместной реализацией состояний A_k и B_l, будет суммой неопределенностей исходных объектов. Пусть $H(A)$, $H(B)$, $H(AB)$ означают соответственно энтропии объектов A, B и AB. Так как вероятность π_{kl} состояния $A_k B_l$ в случае независимости A и B равна произведению $p_k q_l$, то

$$H(AB) = -\sum_{k,l} \pi_{kl} \log \pi_{kl} = -\sum_{k,l} p_k q_l (\log p_k + \log q_l) = -\sum_k p_k \log p_k \cdot \sum_l q_l - \sum_l q_l \log q_l \cdot \sum_k p_k =$$

$$= H(A) + H(B). \qquad (2.2)$$

Как видим, и это требование удовлетворяется функцией H.

4. Рассмотрим энтропию объекта AB при условии статистической зависимости между A и B. Обозначим через q_{kl} вероятность того, что реализуется состояние B_l объекта B, если известно, что реализовалось состояние A_k объекта A. Тогда $\pi_{kl} = p_k \cdot q_{kl}$. Энтропия объекта AB теперь равна

$$H(AB) = -\sum_{k,l} p_k q_{kl} (\log p_k + \log q_{kl}) =$$

$$= -\sum_k p_k \log p_k \cdot \sum_l q_{kl} - \sum_k p_k \sum_l q_{kl} \log q_{kl}.$$

Сумма $-\sum_l q_{kl} \log q_{kl}$ может рассматриваться как условная энтропия $H_{A_k}(B)$ объекта B, вычисленная при условии, что реализовалось состояние A_k объекта A. Так как $\sum_l q_{kl} = 1$ при любом k, то

$$H(AB) = H(A) + \sum_k p_k H_{A_k}(B).$$

Так как значение условной энтропии $H_{A_k}(B)$ определяется

тем, какое из состояний объекта A реализуется, то $H_{A_k}(B)$ является случайной величиной. Последний член в правой части представляет собой, следовательно, математическое ожидание величины $H_{A_k}(B)$ в схеме A, которое мы будем обозначать через $H_A(B)$. Таким образом, в самом общем случае (с учетом симметрии)

$$H(AB) = H(A) + H_A(B) = H(B) + H_B(A). \tag{2.3}$$

В частном случае независимости A и B это соотношение переходит в (2.2).

5. Заметим, что всегда $H_A(B) \leqslant H(B)$; это неравенство хорошо согласуется с интуитивным представлением о том, что знание состояния объекта A может только уменьшить неопределенность объекта B. Для доказательства воспользуемся тождественным неравенством для любой выпуклой функции $f(x)$:

$$\sum_k \lambda_k f(x_k) \geqslant f\left(\sum_k \lambda_k x_k\right),$$

при $\lambda_k \geqslant 0$, $\sum_k \lambda_k = 1$. Полагая $f(x) = x \log x$, $\lambda_k = p_k$, $x_k = q_{kl}$, имеем:

$$\sum_k p_k q_{kl} \log q_{kl} \geqslant \left(\sum_k p_k q_{kl}\right) \log \left(\sum_k p_k q_{kl}\right) = q_l \log q_l,$$

так как $\sum_k p_k q_{kl} = q_l$. Просуммируем обе части полученного неравенства по l. Слева имеем $-H(B)$, справа:

$$\sum_k p_k \sum_l q_{kl} \log q_{kl} = -\sum_k p_k H_{A_k}(B) = -H_A(B).$$

Отсюда следует, что

$$-H_A(B) \geqslant -H(B),$$

что и требовалось доказать. Таким образом, энтропия объекта B никогда не возрастает вследствие знания состояния объекта A. Она уменьшается, если только A и B не являются независимыми; в противном случае энтропия объекта B остается неизменной.

Свойства 1—5 функционала H показывают, что он действительно пригоден в качестве меры неопределенности случайного объекта с конечным числом состояний. Будем применять функционал энтропии как меру неопределенности и по отношению к объектам с бесконечным счетным числом возможных состояний, хотя соответствующей теоремы единствен-

ности в литературе нет. В приложениях это не приводит к недоразумениям, и кажется очевидным, что такое обобщение справедливо, но для завершенности теории интересно было бы дать строгое доказательство этого.

§ 3. ТЕОРЕМА ЕДИНСТВЕННОСТИ ФУНКЦИОНАЛА ЭНТРОПИИ КАК МЕРЫ НЕОПРЕДЕЛЕННОСТИ КОНЕЧНОЙ СХЕМЫ

В предыдущем параграфе было показано, что функционал энтропии

$$H = -\sum_{k=1}^{n} p_k \log p_k \tag{3.1}$$

обладает всеми свойствами, которые разумно потребовать от количественной меры неопределенности конечной схемы. Возникает естественный вопрос, является ли при этом вид функции H единственным совместимым с этими свойствами?

К числу основных свойств энтропии относятся следующие:

1. При данном n функция $H(p_1, p_2, \ldots, p_n)$ получает наибольшее значение, если $p_1 = p_2 = \ldots = p_n = \frac{1}{n}$, т. е.

$$H(p_1 p_2, \ldots, p_n) \leqslant H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right).$$

2. $$H(AB) = H(A) + H_A(B).$$

К этим двум свойствам добавим еще одно, которое очевидным образом должно выполняться при любом разумном определении энтропии.

3. Добавление к множеству состояний одного невозможного состояния (а значит, и любого числа таких состояний), не изменяет неопределенности объекта, т. е.

$$H(p_1, p_2, \ldots, p_n, 0) = H(p_1, p_2, \ldots, p_n). \tag{3.2}$$

Как показывают исследования [7, 47], этих условий достаточно для однозначного определения вида функционала H.

Это утверждается следующей теоремой единственности:

Если функция $H(p_1, p_2, \ldots p_n)$ при любом n непрерывна относительно совокупности своих аргументов и если она обладает свойствами 1, 2 и 3, то функция H имеет вид

$$H(p_1, p_2, \ldots, p_n) = -\lambda \sum_{k=1}^{n} p_k \log p_k, \tag{3.3}$$

где λ — постоянное положительное число.

Доказательство [47]. Положим для краткости $H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right) = L(n)$ и докажем, что $L(n) = \lambda \log n$, где $\lambda = \text{const} > 0$. В силу условий 3 и 1 имеем:

$$L(n) = H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}, 0\right) \leqslant$$

$$\leqslant H\left(\frac{1}{n+1}, \frac{1}{n+1}, \ldots, \frac{1}{n+1}\right) = L(n+1),$$

так что $L(n)$ есть неубывающая функция n.

Пусть m и r — натуральные числа. Рассмотрим m взаимно независимых объектов $S_1, S_2, \ldots, S_m$, каждый из которых имеет r равновероятных состояний, так что

$$H(S_k) = H\left(\frac{1}{r}, \frac{1}{r}, \ldots, \frac{1}{r}\right) = L(r),$$

$$(1 \leqslant k \leqslant m).$$

В силу свойства 2 с учетом независимости $\{S_i\}$,

$$H(S_1, S_2, \ldots, S_m) = \sum_{k=1}^{m} H(S_k) = mL(r).$$

Но объект $(S_1, S_2, \ldots, S_m)$ имеет, очевидно, r^m равновероятных состояний, так что его неопределенность равна $L(r^m)$. Отсюда, $L(r^m) = mL(r)$ и аналогично для любой другой пары натуральных чисел n и s: $L(s^n) = nL(s)$.

Пусть теперь числа r, s и n заданы произвольно, а число m определяется неравенствами

$$r^m \leqslant s^n \leqslant r^{m+1}. \tag{3.4}$$

Откуда

$$m \log r \leqslant n \log s \leqslant (m+1) \log r,$$

$$\frac{m}{n} \leqslant \frac{\log s}{\log r} \leqslant \frac{m}{n} + \frac{1}{n}. \tag{3.5}$$

В силу доказанной монотонности функции $L(n)$ из (3.4) следует

$$L(r^m) \leqslant L(s^n) \leqslant L(r^{m+1}),$$

и согласно свойству функции $L(n)$

$$mL(r) \leqslant nL(s) \leqslant (m+1)L(r),$$

откуда

$$\frac{m}{n} \leqslant \frac{L(s)}{L(r)} \leqslant \frac{m}{n} + \frac{1}{n}. \tag{3.6}$$

Из (3.5) и (3.6) вытекает, что

$$\left| \frac{L(s)}{L(r)} - \frac{\log s}{\log r} \right| \leqslant \frac{1}{n}.$$

Так как левая часть этого неравенства от n не зависит, а в правой n может быть взято сколь угодно большим, то

$$\frac{L(s)}{\log s} = \frac{L(r)}{\log r}$$

и, значит, ввиду произвольности r и s,

$$L(n) = \lambda \log n,$$

где λ — постоянная. В силу доказанной монотонности функции $L(n)$ мы имеем $\lambda \geqslant 0$, и для частного случая $p_1 = p_2 = \ldots = p_n = \dfrac{1}{n}$ теорема доказана.

Рассмотрим более общий случай, когда вероятности состояний объекта A, $p_1, p_2, \ldots, p_n$ — произвольные рациональные числа (причем, конечно, $p_k \geqslant 0$, $\sum\limits_k p_k = 1$). Пусть $p_k = g_k/g$, где все g_k — натуральные числа, $g = \sum\limits_k g_k$. Рассмотрим вторую схему B, зависимую от A и определяемую следующим образом: схема B содержит g состояний $B_1, B_2, \ldots, B_g$, которые мы подразделяем на n групп, содержащих соответственно $g_1, g_2, \ldots, g_n$ состояний. Если реализовалось состояние A_k объекта A, то все состояния κ-й группы объекта B (число которых равно g_k) получают одну и ту же вероятность $1/g_k$, а все состояния других групп получают вероятность 0 (становятся невозможными).

Таким образом, при любом состоянии A_k объекта A объект B имеет g_k равновероятных состояний, вследствие чего условная энтропия

$$H_{A_k}(B) = H\left(\frac{1}{g_k}, \frac{1}{g_k}, \ldots, \frac{1}{g_k}\right) = L(g_k) = \lambda \log g_k,$$

а значит

$$H_A(B) = \sum_k p_k H_{A_k}(B) = \lambda \sum_k p_k \log g_k =$$

$$= \lambda \sum_k p_k \log p_k + \lambda \log g. \qquad (3.7)$$

Обратимся теперь к объединенной схеме AB, образованной состояниями A_kB_l $(1 \leqslant k \leqslant n,\ 1 \leqslant l \leqslant g)$. Состояние A_kB_l имеет отличную от нуля вероятность только, когда B_l принадлежит k-й группе. Таким образом, число возможных состояний A_kB_l при данном k равно g_k; общее же число возможных состояний объекта AB равно $\sum_k g_k = g$.

Вероятность каждого возможного состояния A_kB_l равна, очевидно, $p_k \cdot \frac{1}{g_k} = 1/g$, т. е. одна и та же для всех состояний объекта AB. Отсюда следует, что

$$H(AB) = L(g) = \lambda \log g.$$

Пользуясь свойством 2 и соотношением (3.7), находим:

$$\lambda \log g = H(A) + \lambda \sum_k p_k \log p_k + \lambda \log g,$$

откуда

$$H(A) = H(p_1, p_2, \ldots, p_n) = -\lambda \sum_k p_k \log p_k. \qquad (3.8)$$

Соотношение (3.8), доказанное для рациональных $p_1, p_2, \ldots, p_n$, в силу предположений непрерывности функции $H(p_1, \ldots, p_n)$ должно остаться верным и при любых значениях аргументов. Таким образом, теорема единственности функционала энтропии конечной схемы доказана полностью.

§ 4. НЕОПРЕДЕЛЕННОСТЬ НЕПРЕРЫВНЫХ СЛУЧАЙНЫХ ВЕЛИЧИН. ДИФФЕРЕНЦИАЛЬНАЯ (ОТНОСИТЕЛЬНАЯ) ЭНТРОПИЯ

Введение количественной меры неопределенности случайного объекта оказывается настолько полезным, что возникает естественное желание обобщить понятие энтропии так, чтобы иметь возможность количественно оценивать неопределенность случайных объектов с континуумом возможных состояний. Здесь мы сразу сталкиваемся с целым рядом особенностей. Прежде всего, понятие «непрерывность» имеет смысл только для количеств, так что объект с континуумом возможных состояний — это по необходимости количественная случайная величина. Собираясь в дальнейшем вести обобщение по аналогии с дискретным случаем, мы замечаем, что роль

распределения вероятности по состояниям в непрерывном случае играет плотность вероятности, являющаяся в общем случае размерной величиной*). Желая не иметь дела с логарифмами размерных величин, введем в рассмотрение безразмерную случайную величину $x = x^*/x_0$, где x^* — размерная случайная величина, x_0 — единица ее измерения. Тогда и плотность вероятности $p(x)$ будет безразмерной функцией.

Попытаемся теперь непосредственно, путем предельного перехода, от дискретного случая перейти к непрерывному. Произведем для этого квантование значений непрерывной случайной величины x на счетное число уровней, т. е. разобьем всю область $(-\infty, \infty)$ возможных значений величины x на интервалы, разделенные отстоящими на равных расстояниях Δx друг от друга уровнями $(\ldots, x_{-1}, x_0, x_1, \ldots, x_k, \ldots)$. Будем теперь всякий раз, как реализуется значение $x \in (x_k, x_k + \Delta x)$ считать, что реализовалось значение x_k случайной величины x' (рис. 16).

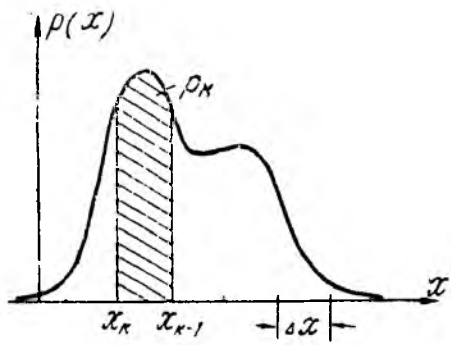

Рис. 16.

Полученная таким образом дискретная случайная величина x' характеризуется распределением, в котором вероятность k-го состояния равна

$$p_k = \int_{x_k}^{x_k + \Delta x} p(x)\, dx, \tag{4.1}$$

где $p(x)$ — плотность вероятности квантуемой непрерывной величины. С одной стороны, очевидно, что при $\Delta x \to 0$ кван-

*) Размерность плотности вероятности $p(x)$ обратна размерности x, так как величина $p(x)\, dx$, являющаяся вероятностью события $x \leq X \leq x + dx$, безразмерна.

тованная величина x' будет все более и более полно отражать все свойства непрерывной (квантуемой) величины x. С другой стороны, к дискретной случайной квантованной величине x' безо всяких оговорок применимо понятие энтропии. В связи с этим и появляется надежда получить выражение энтропии непрерывной величины x, рассмотрев предельное выражение энтропии дискретной величины x'. По определению имеем:

$$H(x') = -\sum_{-\infty}^{\infty} p_k \log p_k =$$

$$= -\sum_{-\infty}^{\infty} \left[\int_{x_k}^{x_k + \Delta x} p(x)\, dx\right] \log \left[\int_{x_k}^{x_k + \Delta x} p(x)\, dx\right]. \qquad (4.2)$$

Начнем теперь одновременно уменьшать все Δx. При достаточно малых Δx и достаточно гладкой $p(x)$

$$\int_{x_k}^{x_k + \Delta x} p(x)\, dx \approx p(x_k)\, \Delta x, \qquad (4.3)$$

поэтому

$$\lim_{\Delta x \to 0} H(x') \approx \lim_{\Delta x \to 0} \{-\sum_{-\infty}^{\infty} p(x_k)\, \Delta x \log [p(x_k)\, \Delta x]\} =$$

$$= \lim_{\Delta x \to 0} \{-\sum_{-\infty}^{\infty} p(x_k) [\log p(x_k)]\, \Delta x\} +$$

$$+ \lim_{\Delta x \to 0} \{-\sum_{-\infty}^{\infty} p(x_k) [\log \Delta x]\, \Delta x\} =$$

$$= -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx + \lim_{\Delta x \to 0} \{-[\log \Delta x] \sum_{-\infty}^{\infty} p(x_k)\, \Delta x\} =$$

$$= -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx - \lim_{\Delta x \to 0} \log \Delta x. \qquad (4.4)$$

Таким образом, энтропия величины x' (за счет второго члена) стремится к бесконечности при уменьшении интервала квантования Δx. Этого, вообще говоря, и следовало ожидать, так как даже в дискретном случае при бесконечном числе состояний энтропия не имеет верхней грани; да и интуиция подсказывает, что неопределенность реализации одного из бесконечного множества состояний может быть сколь угодно велика.

Убедившись, таким образом, что непрерывные случайные объекты не допускают введения конечной абсолютной меры

неопределенности, мы можем, тем не менее, ввести относительную количественную меру неопределенности и в непрерывном случае. В качестве стандарта для сравнения можно взять неопределенность какого-либо простого распределения, например, равномерного в интервале шириной ε. Производя квантование равномерно распределенной в ε величины, получим, что предел неопределенности образованной таким образом дискретной величины x'' запишется как

$$\lim_{\Delta x \to 0} H(x'') = \log \varepsilon - \lim_{\Delta x \to 0} \log \Delta x. \tag{4.5}$$

Будем характеризовать неопределенность непрерывной случайной величины x числом, к которому в пределе стремится разность энтропий квантованных величин x' и x'':

$$H_\varepsilon(x) = \lim_{\Delta x \to 0} [H(x') - H(x'')] =$$

$$= -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx - \log \varepsilon = -\int_{-\infty}^{\infty} p(x) \log \varepsilon p(x)\, dx. \tag{4.6}$$

Как видим, эта разность конечна. Если взять за стандарт неопределенность случайной величины, равномерно распределенной в единичном интервале ($\varepsilon = 1$), то запись величины $H_\varepsilon(x)$ упростится:

$$H_{\varepsilon=1}(x) = -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx. \tag{4.7}$$

В такой записи ε явно не фигурирует, однако, это нисколько не означает, что $H_\varepsilon(x)$ перестала быть относительной величиной. Число $H_\varepsilon(x)$ обычно и называют энтропией непрерывной случайной величины. Отличия от энтропии дискретных величин подчеркиваются в названии: благодаря связи $H_\varepsilon(x)$ с дифференциальным законом распределения вероятностей, ее часто называют дифференциальной энтропией; иногда употребляемый термин относительная энтропия указывает на условность, относительность этой характеристики. Укажем также, что «неабсолютность» величины $H_\varepsilon(x)$ усугубляется специфичностью ее вычисления: вообще говоря, можно задать такой закон квантования и такой способ стремления интервалов квантования к нулю, что предел соответствующих сумм будет отличаться от $H_\varepsilon(x)$. Однако, установив однажды способ вычисления какой-то характеристики и выяснив ее смысл, в дальнейшем можно пользоваться ею для сравнения различных распределений. Поэтому величина $H_\varepsilon(x)$, несмотря на ее относительность, имеет очень важное значение в теории информации.

§ 5. СВОЙСТВА ДИФФЕРЕНЦИАЛЬНОЙ ЭНТРОПИИ

Энтропия непрерывных величин обладает свойствами, во многом аналогичными, а в целом ряде случаев обобщающими свойства энтропии дискретных объектов.

1. Прежде всего отметим, что хотя при переходе к непрерывным распределениям случайные признаки объекта по необходимости сделались численными, тем не менее характерное свойство энтропии — независимость от конкретных значений случайной величины —сохранилось. Например, «параллельный перенос» распределения вероятностей на любой интервал a не изменяет дифференциальной энтропии. В самом деле,

$$H_\varepsilon(y = x - a) = -\int p(x-a)\log p(x-a)\,dx$$
$$= -\int p(x)\log p(x)\,dx = H_\varepsilon(x).$$

Не изменяют величину дифференциальной энтропии и такие преобразования, как «зеркальное отображение» и даже перераспределение вероятностей при сохранении множества дифференциальных вероятностей (см. рис. 17). В дискретном случае этим операциям соответствует простая перенумерация состояний. Укажем, что в непрерывном случае эти преобразования не изменяют масштаба переменной, что и обеспечивает сохранение величины дифференциальной энтропии.

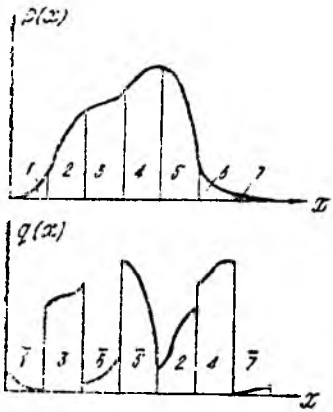

Рис. 17.

2. Энтропия и в непрерывном случае сохраняет свойство, выражаемое соотношением

$$H(x,y) = H(x) + H_x(y) = H(y) + H_y(x), \qquad (5.1)$$

где

$$H(x,y) = -\iint p(x,y)\log p(x,y)\,dx\,dy,$$
$$H_x(y) = -\iint p(x,y)\log p(y|x)\,dx\,dy, \qquad (5.2)$$
$$H_y(x) = -\iint p(x,y)\log p(x|y)\,dx\,dy.$$

Соотношение (5.1) легко доказывается простой подстановкой равенств (5.2) *) и является аналогом свойства 4 энтропии дискретного распределения.

*) Здесь и далее будем для краткости опускать индекс у $H_\varepsilon(x)$, если не будет необходимости специально учитывать зависимость дифференциальной энтропии от ε.

3. При любых двух случайных переменных x и y

$$H_\varepsilon(x, y) \leqslant H_\varepsilon(x) + H_\varepsilon(y), \tag{5.3}$$

причем знак равенства будет тогда (и только тогда), когда x и y независимы.

4. Прежде чем переходить к рассмотрению дальнейших свойств дифференциальной энтропии, докажем одно важное соотношение. Для любых плотностей вероятности $f(x)$ и $g(x)$ выполняется неравенство:

$$\int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)}\, dx \geqslant 0. \tag{5.4}$$

Для доказательства воспользуемся неравенством

$$\log_a u = \frac{\ln u}{\ln a} \geqslant \frac{1}{\ln a}\left(1 - \frac{1}{u}\right), \tag{5.5}$$

справедливым при любом $u \geqslant 0$. В самом деле, если $u \geqslant 1$, то $\frac{1}{t} \geqslant \frac{1}{u}$ при $1 \leqslant t \leqslant u$, и

$$\ln u = \int_1^u \frac{dt}{t} \geqslant \frac{1}{u}\int_1^u dt = \frac{1}{u}(u-1) = 1 - \frac{1}{u}.$$

Если $0 \leqslant u < 1$, то $-\frac{1}{t} \geqslant -\frac{1}{u}$ при $u \leqslant t < 1$, и

$$\ln u = -\int_u^1 \frac{dt}{t} \geqslant -\frac{1}{u}\int_u^1 dt = -\frac{1}{u}(1-u) = 1 - \frac{1}{u}.$$

Из неравенства (5.5) следует (с учетом $\int_{-\infty}^{\infty} p(x)\, dx = 1$)

$$\int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)}\, dx \geqslant \frac{1}{\ln a}\int_{-\infty}^{\infty} f(x)\left[1 - \frac{g(x)}{f(x)}\right] dx = 0. \tag{5.6}$$

Так как знак равенства в (5.5) имеет место только при $u = 1$, то и в (5.6) и (5.4) равенство имеет место только при $f(x) \equiv g(x)$.

5. Неравенство (5.4) позволяет доказать следующее важное свойство дифференциальной энтропии *): всякое сгла-

*) Свойство 5 дифференциальной энтропии, как будет показано ниже, имеет место и для энтропии дискретных объектов.

живание (усреднение) распределения вероятности $f(x)$ может привести только к возрастанию энтропии. Операция усреднения выражается соотношением

$$g(y) = \int_{-\infty}^{\infty} a(x, y) f(x)\, dx,$$

где $a(x, y)$ весовая функция, удовлетворяющая следующим условиям:

$$a(x, y) \geqslant 0, \int_{-\infty}^{\infty} a(x, y)\, dx = \int_{-\infty}^{\infty} a(x, y)\, dy = 1. \tag{5.7}$$

Функция $g(y)$ обладает всеми свойствами плотности вероятности:

$$g(y) \geqslant 0, \int_{-\infty}^{\infty} g(y)\, dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a(x, y) f(x)\, dx\, dy =$$

$$= \int_{-\infty}^{\infty} f(x) \left[\int_{-\infty}^{\infty} a(x, y)\, dy\right] dx = \int_{-\infty}^{\infty} f(x)\, dx = 1. \tag{5.8}$$

Определим, в каком соотношении находятся дифференциальные энтропии распределений $g(y)$ и $f(x)$.

$$H_\varepsilon(y) - H_\varepsilon(x) = \int_{-\infty}^{\infty} f(x) \log f(x)\, dx -$$

$$- \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a(x, y) f(x) \log g(y)\, dx\, dy =$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a(x, y) f(x) \log \frac{f(x)}{g(y)}\, dx\, dy =$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} a(x, y) f(x) \log \frac{a(x, y) f(x)}{a(x, y) g(y)}\, dx\, dy. \tag{5.9}$$

Функции $a(x, y) \cdot f(x)$ и $a(x, y) \cdot g(y)$ обладают свойствами плотностей вероятности (см. (5.7), (5.8)). Поэтому в силу неравенства (5.4), справедливого и для многомерных плотностей вероятности,

$$H_\varepsilon(y) - H_\varepsilon(x) \geqslant 0. \tag{5.10}$$

Таким образом, энтропия усредненного распределения $g(y)$ больше или равна энтропии исходного распределения $f(x)$.

6. Как мы видели в предыдущем параграфе, предельный переход от дискретной величины к непрерывной не мог дать выражения неопределенности последней в конечном виде. В связи с этим мы ввели относительную меру неопределенности непрерывной случайной величины, дифференциальную энтропию. Можно показать [27], что дифференциальная энтропия подходящим образом распределенной непрерывной величины в пределе переходит в обычную энтропию. Рассмотрим непрерывную случайную величину x, плотность вероятности которой $p(x)$ равна нулю всюду, кроме n-интервалов длины Δ каждый; причем в каждом из интервалов

$$p(x) = \text{const} = p_k \Delta, \quad (k = 1, 2, \ldots, n; \ \sum_k p_k = 1)$$

(см. рис. 18). Дифференциальная энтропия такой величины равна

$$H_\varepsilon(x) = -\int_{-\infty}^{\infty} p(x) \log [\varepsilon p(x)]\, dx = -\sum_{k=1}^{n} p_k \log \varepsilon \cdot \frac{p_k}{\Delta}. \quad (5.11)$$

Выберем в качестве стандарта $\varepsilon = \Delta$ и устремим одновременно ε и Δ к нулю. При этом в пределе мы получим дискретную случайную величину, неопределенность которой выразится обычной формулой

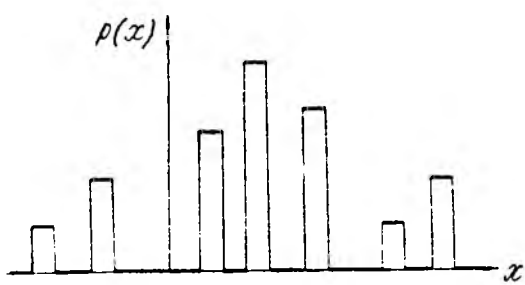

Рис. 18.

$$H(x) = -\sum_{k=1}^{n} p_k \log p_k, \quad (5.12)$$

совпадающей с $H_\varepsilon(x)$ при $\varepsilon = \Delta$.

Совпадение это не случайно, оно полностью оправдывается тем, что распределение вероятностей дискретной вели-

чины может быть представлено с помощью аппарата δ-функций:

$$f(x) = \sum_{k=1}^{n} p_k \delta(x - x_k),$$

а предел, к которому стремится функция $\varphi(x) = \begin{cases} \Delta^{-1}, & |x| \leqslant \frac{\Delta}{2}, \\ 0, & |x| > \frac{\Delta}{2}, \end{cases}$ при $\Delta \to 0$ является одним из представлений δ-функции.

Установленный таким образом переход от дифференциальной энтропии к энтропии дискретной величины еще раз подчеркивает их глубокую связь и общность свойств. В частности, свойство 5, доказанное для непрерывного случая, очевидно, сохраняется и при переходе к дискретному случаю. Следует, однако, всегда иметь в виду, что при всей их общности дифференциальная энтропия отличается от энтропии дискретной величины своей относительностью. Эта особенность в явном виде скажется при преобразованиях случайной величины, приводящих к изменению масштаба (см. § 7).

§ 6. ПРИНЦИП ЭКСТРЕМУМА ЭНТРОПИИ И ЭКСТРЕМАЛЬНЫЕ РАСПРЕДЕЛЕНИЯ

Довольно часто при решении теоретических, а иногда и практических вопросов встречается следующая ситуация. По каким-либо причинам нам известны лишь некоторые ограничения, накладываемые на случайную величину x; чаще всего это значения моментов, но могут быть заданы и другие условия, например, ограничение некоторыми пределами сверху и снизу области возможных значений случайной величины, или более сложные условия. Ставится задача: не зная ничего, кроме этих ограничений, не привлекая дополнительных сведений, — задаться для дальнейшего некоторым распределением вероятностей $p(x)$.

Такая задача возникает, например, при выборе «наилучшего» распределения вероятностей искусственно создаваемой помехи при заданной мощности генератора. Другим примером может служить задача статистической физики о подборе распределения вероятности по уровням энергии элементарных частиц, если известна лишь средняя энергия частиц [41]. Можно даже утверждать, что с такого рода ситуациями мы встречаемся чаще, чем принято думать: задаваясь, например, конечным числом моментов, мы тем самым в неявном виде останавливаем свой выбор на распределении определенного типа.

Нужно сразу уяснить себе, что заданному ограничению (или ограничениям) всегда удовлетворяет бесконечное множество различных распределений. Поэтому задача нахождения распределения вероятностей, удовлетворяющего данным ограничениям, фактически сводится к задаче выбора из данного множества некоторого «наиболее подходящего» распределения. Такой набор осуществить нельзя до тех пор, пока не будет задан точный критерий того, какое распределение считать «наиболее подходящим».

В качестве такого критерия теория информации предлагает принцип экстремума энтропии, который подписывает выбирать из множества характеризующихся заданными свойствами распределений то, которое обладает экстремальной, например, максимальной энтропией. Достаточным основанием для выдвижения этого принципа является то, что энтропия есть мера неопределенности случайной величины. Приписывая случайной величине некоторое распределение, пусть даже имеющее ряд свойств, совпадающих со свойствами действительного (неизвестного) распределения, мы должны отдавать себе отчет в том, что совершаем по необходимости некоторый произвол: выбор распределения однозначно определяет и те характеристики, которые нам неизвестны. Принцип максимума энтропии гарантирует максимальную неопределенность выбираемого распределения, и в этом смысле «минимальный» произвол. В тех же случаях, когда мы сами можем задавать распределение случайной величины (например, при создании помех), принцип максимума энтропии имеет еще более ясный смысл.

Продемонстрируем применение принципа максимума энтропии на примерах, которые имеют не только иллюстративное значение.

1. Пусть известно, что область возможных значений случайной величины x ограничена интервалом $[a,b]$, $a \leqslant x \leqslant b$. Найдем распределение, обладающее при этом максимальной энтропией. Для этого необходимо решить следующую вариационную задачу: найти функцию $p(x)$, обеспечивающую максимум функционала

$$H_x(x) = -\int_a^b p(x) \log p(x)\, dx \tag{6.1}$$

при дополнительном условии:

$$\int_a^b p(x)\, dx = 1, \tag{6.2}$$

Согласно известным теоремам вариационного исчисления для этого необходимо максимизировать функционал

$$\int_a^b F(x,p)\,dx = \int_a^b [-p(x)\log p(x) + \lambda p(x)]\,dx, \qquad (6.3)$$

что приводит к уравнению

$$\frac{\partial F(x,p)}{\partial p} = [-1 - \log p + \lambda] = 0, \qquad (6.4)$$

откуда

$$p = \exp(\lambda - 1). \qquad (6.5)$$

Пользуясь условием (6.2), определим неизвестную константу λ:

$$\int_a^b \exp(\lambda - 1)\,dx = \exp(\lambda - 1)\int_a^b dx =$$

$$= [\exp(\lambda - 1)](b - a). \qquad (6.6)$$

Отсюда непосредственно получаем:

$$p(x) = \begin{cases} 0, & x < a, \\ \dfrac{1}{b-a}, & a \leqslant x \leqslant b, \\ 0, & x > b, \end{cases} \qquad (6.7)$$

т. е. максимальной энтропией при ограниченности сверху и снизу области возможных значений x обладает равномерное в интервале $[a,b]$ распределение вероятностей.

Этот пример еще раз подчеркивает общность свойств дифференциальной энтропии с энтропией дискретной конечной системы: при задании числа n возможных состояний последняя достигала максимального значения $H = \log n$ при равномерном распределении; дифференциальная энтропия также имеет при ограничении значений x интервалом величины $b - a$ максимальное значение $H_\varepsilon(x) = \log(b - a)$ также при равномерном распределении.

Кроме того, этот пример оправдывает то, что в качестве эталона неопределенности при введении дифференциальной энтропии мы выбираем неопределенность равномерного в интервале распределения.

2. Рассмотрим случай, когда о случайной величине известно следующее: а) область возможных значений неограничена, $-\infty \leqslant x \leqslant \infty$; б) известно среднее значение величины x, $\bar{x} = a$; в) задана величина центрального момента второго порядка (дисперсии),

$\overline{(x-a)^2} = \sigma^2$. При этом задача сводится к нахождению максимума функционала

$$H_\varepsilon(x) = -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx. \tag{6.8}$$

при условиях

$$\int_{-\infty}^{\infty} (x-a)^2 p(x)\, dx = \sigma^2, \tag{6.9}$$

$$\int_{-\infty}^{\infty} x\, p(x)\, dx = a, \tag{6.10}$$

$$\int_{-\infty}^{\infty} p(x)\, dx = 1. \tag{6.11}$$

Уравнение для нахождения $p(x)$ с максимальной энтропией запишется в виде

$$-1 - \log p + \lambda_1 (x-a)^2 + \lambda_2 x + \lambda_3 = 0, \tag{6.12}$$

откуда

$$p = \exp\left[\lambda_1 (x-a)^2 + \lambda_2 x + \lambda_3 - 1\right]. \tag{6.13}$$

Воспользуемся условием (6.11):

$$\int_{-\infty}^{\infty} p\, dx = \exp(\lambda_3 - 1) \int_{-\infty}^{\infty} \exp\left[\lambda_1 (x-a)^2 + \lambda_2 x\right] dx = 1. \tag{6.14}$$

Отсюда следует, что для сходимости интеграла константа λ_1 должна быть отрицательной. Тогда

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

$$= \exp(\lambda_2 a) \int_{-\infty}^{\infty} \exp(\lambda_1 y^2 + \lambda_2 y)\, dy =$$

$$= \exp(\lambda_2 a) \exp\left(-\frac{\lambda_2^2}{4\lambda_1}\right) \sqrt{\frac{\pi}{-\lambda_1}} \tag{6.15}$$

(см. [29], 3.213). Таким образом, получаем:

$$\exp(\lambda_3 - 1) = \sqrt{\frac{-\lambda_1}{\pi}} \exp\left(\frac{\lambda_2^2}{4\lambda_1} - \lambda_2 a\right). \tag{6.16}$$

Второе условие (6.10) не позволяет получать дополнительных связей между константами λ_1 и λ_2, так как при подста-

новке (6.13) и (6.15) в (6.10) образуется тождество, выполняющееся при любых соотношениях постоянных λ_1 и λ_2. Отложим пока выяснение причин этого и обратимся к условию (6.9), которое с учетом (6.13) и (6.15) после несложных преобразований и приведения интегралов к табличным приводит к выражению вида

$$\frac{1}{-\lambda_1\sqrt{\pi}}\left[\frac{\sqrt{\pi}}{2}+\frac{\lambda_2^2\sqrt{\pi}}{4(-\lambda_1)}+\frac{\lambda_2}{2\sqrt{-\lambda_1}}\int_{-\infty}^{\infty}e^{-z}\,dz\right]=\sigma^2. \quad (6.17)$$

Легко видеть, что для конечности левой части уравнения константа λ_2 должна быть равной нулю. Тогда

$$\lambda_1=-\frac{1}{2\sigma^2}, \quad (6.18)$$

и, окончательно, получаем, что в рассматриваемом случае экстремальное распределение имеет вид

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-a)^2}{2\sigma^2}\right], \quad (6.19)$$

т. е. является нормальным распределением.

В ходе решения выяснилось, что условие (6.10) не дало никаких дополнительных, по сравнению с (6.9), сведений, т. е. оказалось просто «излишним». Дело здесь, конечно, не в том, что знания о среднем значении «не нужны»; просто эти сведения уже содержатся в условии (6.9), так как дисперсия является центральным моментом, т. е. моментом относительно среднего значения. При задании же начального второго момента условие (6.10), конечно, не было бы избыточным.

3. Изложенная выше методика позволяет находить экстремальные распределения (т. е. распределения с максимальной или минимальной энтропией) и в других случаях. Например, читатель может легко убедиться, что в случае, когда x ограничено половиной числовой оси [$p(x)=0$ при $x<0$], а первый момент x равен a, то максимум энтропии имеет место при

$$p(x)=\frac{1}{a}\exp\left(-\frac{x}{a}\right). \quad (6.20)$$

Совершенно аналогично решаются задачи и в случае дискретных x. Следует, однако, подчеркнуть, что вариационная методика позволяет оперировать лишь с однозначными

условиями, задаваемыми в виде строгих равенств. Мы не смогли бы, например, довести решение задачи 2 до конца если условие (6.9) было бы записано в форме

$$\int_{-\infty}^{\infty}(x-a)^2 p(x)\,dx \leqslant \sigma^2.$$

В подобных случаях необходимо привлекать дополнительные сведения (в данном случае—априорную статистику дисперсии), чтобы привести все же постановку задачи к требуемой форме.

§ 7. ИЗМЕНЕНИЕ ДИФФЕРЕНЦИАЛЬНОЙ ЭНТРОПИИ ПРИ ПРЕОБРАЗОВАНИЯХ КООРДИНАТ

Зависимость дифференциальной энтропии от выбора величины интервала ε в явном виде проявляется при преобразованиях координат. В самом деле, пусть y есть некоторая (для простоты—однозначная) функция случайной величины x, $y = y(x)$. Сравним энтропии величин x и y:

$$H_\varepsilon(x) = -\int p(x)\log p(x)\,dx, \tag{7.1}$$

$$H_\varepsilon(y) = -\int q(y)\log q(y)\,dy. \tag{7.2}$$

Между плотностями распределений вероятностей $q(y)$ и $p(x)$ имеется связь:

$$q(y) = p(x)\left|\frac{dx}{dy}\right| = p(x)\left|J\left(\frac{x}{y}\right)\right|, \tag{7.3}$$

где $J\left(\frac{x}{y}\right)$ — якобиан перехода от x к y. Следовательно,

$$H_\varepsilon(y) = -\int p(x)\log\left[p(x)\left|J\left(\frac{x}{y}\right)\right|\right]dx =$$

$$= -\int p(x)\log p(x)\,dx - \int p(x)\log\left|J\left(\frac{x}{y}\right)\right|dx =$$

$$= H_\varepsilon(x) - \overline{\log\left|J\left(\frac{x}{y}\right)\right|}. \tag{7.4}$$

Запись в одинаковой форме (7.1) и (7.2) дифференциальных энтропий величин x и y означает, что в обоих случаях мы

сравниваем неопределенность соответствующих распределений с неопределенностью одного и того же распределения, равномерного в единичном интервале ε. Преобразование $y = y(x)$ в общем случае приводит к соответствующему изменению масштаба, т. е. к изменению распределения вероятностей по единичным интервалам, а следовательно, и к изменению относительной неопределенности.

Лишь в тех случаях, когда $\overline{\log \left| J\left(\frac{x}{y}\right)\right|} = 0$, величина дифференциальной энтропии остается неизменной. К числу таких преобразований, в частности, относятся те, для которых $\left|J\left(\frac{x}{y}\right)\right| = 1$, т. е. преобразования переноса ($y = x \pm c$), „зеркальное" преобразование ($y = -x$), и более сложные, но состоящие из этих двух, как, например, на рис. 17—разбиение распределения на полосы и произвольная перестановка их. Интересно заметить, что этим не исчерпывается класс преобразований, сохраняющих энтропию: условие $\overline{\log \left|J\left(\frac{x}{y}\right)\right|} = 0$ допускает функции $\left|J\left(\frac{x}{y}\right)\right|$ иметь произвольные значения везде, за исключением окрестности произвольной точки, для которой $p(x_0) \neq 0$. В самом деле, задав произвольно $\left|J\left(\frac{x}{y}\right)\right| = f(x)$ везде, за исключением окрестности точки x_0, можно найти величину интеграла

$$I_1 = \int\limits_{-\infty}^{x_0 - \frac{\delta}{2}} p(x) \log f(x)\, dx + \int\limits_{x_0 + \frac{\delta}{2}}^{\infty} p(x) \log f(x)\, dx.$$

Затем можно произвольно доопределить $f(x)$ и на интервале $\left(x_0 - \frac{\delta}{2},\quad x_0 + \frac{\delta}{2}\right)$, выполнив лишь условие

$$I_2 = \int\limits_{x_0 - \frac{\delta}{2}}^{x_0 + \frac{\delta}{2}} p(x) \log f(x)\, dx = -I_1.$$

Все вышесказанное допускает очевидное обобщение и на случай многомерных случайных величин; при этом под x и y

в формулах (7.1)—(7.4) следует понимать многомерные величины $x=(x_1, x_2,..., x_n)$ и $y=(y_1, y_2,..., y_n)$, интегралы соответственно становятся кратными, а стандартом сравнения является неопределенность распределения, равномерного в единичном гиперобъеме $V=\varepsilon^n$.

Рассмотрим изменение дифференциальной энтропии в частном, но важном случае линейного преобразования; обратимся сразу к многомерному случаю:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ y_n = a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n. \end{cases} \tag{7.5}$$

Якобиан $J\left(\frac{y}{x}\right)$ будет просто определителем $|a_{ij}|$, а якобиан $J\left(\frac{x}{y}\right)$, фигурирующий в формуле (7.4), связан с $J\left(\frac{y}{x}\right)$ соотношением $J\left(\frac{y}{x}\right) = \left[J\left(\frac{x}{y}\right)\right]^{-1}$.

Отсюда следует, что при линейном преобразовании координат дифференциальная энтропия преобразованной величины связана с дифференциальной энтропией исходной величины формулой:

$$H_\varepsilon(y) = H_\varepsilon(x) + \log|a_{ij}|. \tag{7.6}$$

§ 8. ЭНТРОПИЙНЫЕ ХАРАКТЕРИСТИКИ ДИСКРЕТНЫХ СЛУЧАЙНЫХ ПРОЦЕССОВ

В предыдущих параграфах рассматривалась неопределенность и ее мера — энтропия — для случайных величин и событий. Необходимо теперь сделать следующий шаг — обобщить эти понятия таким образом, чтобы иметь возможность оценивать неопределенность случайных процессов.

Согласно представлениям, рассмотренным в § 2 гл. II, случайный процесс X_t можно рассматривать как множество случайных величин, определенных на множестве значений параметра t (для конкретности будем рассматривать только случай, когда t является временем). Исходя из этого, неопределенность случайного процесса можно определить как неопределенность всей совокупности случайных величин, об-

разующих процесс X_t. Однако на этом пути сразу же встречается ряд особенностей, требующих разъяснения.

Во-первых, нецелесообразно оценивать неопределенность случайного процесса на всей оси времени, от $-\infty$ до $+\infty$, так как в общем случае это сразу же приведет к бесконечным величинам. Естественно, возникает вопрос о возможности введения меры средней неопределенности случайного процесса, приходящейся на некоторый интервал времени, в частности — на единичный интервал.

Во-вторых, при вычислении энтропии множества реализаций случайного процесса конечной длины возникает необходимость учета статистической связи между случайными величинами, соответствующими различным значениям параметра t. Это вызывает ряд трудностей, особенно в случае непрерывного времени.

Имея в виду эти и другие особенности энтропийного описания непрерывных процессов, обратимся сначала к рассмотрению простого случая — стационарного процесса с дискретным временем и дискретным конечным множеством возможных состояний в каждый момент времени.

Рассмотрим множество реализаций такого случайного процесса длительностью в n элементов. Если множество возможных состояний каждого элемента насчитывает m состояний, то общее число отличающихся между собой реализаций будет равно m^n. Осуществление каждой реализации можно рассматривать как осуществление случайного события из m^n возможных событий. Зная распределение вероятностей возможных состояний каждого элемента и характер связей между элементами, можно вычислить вероятность каждой из m^n реализаций, $p(C)$. Теперь, в полном соответствии с тем, как это делалось для обычных случайных событий, мы можем вычислить энтропию множества n-членных реализаций:

$$H_n = -\sum_C p(C) \log p(C). \tag{8.1}$$

Величина H_n, конечно, зависит от того, каково n. Желая образовать унифицированную энтропийную характеристику случайного процесса, естественно ввести в рассмотрение величину

$$H = \lim_{n\to\infty} \frac{H_n}{n}, \tag{8.2}$$

если, конечно, указанный предел существует. Величина H будет характеризовать среднюю неопределенность, приходящуюся на один элемент процесса, и может быть названа энтропией процесса.

Докажем тот важный факт, что для каждого стационарного процесса предел (8.2) действительно существует [48].

Множество реализаций длиной $n+l$ элементов (где n и l — любые натуральные числа) можно рассматривать как некоторое объединение множеств реализаций длиной n и длиной l элементов. Согласно свойств 4 и 5 энтропии (см. § 2), при этом выполняются следующие соотношения:

$$H_{n+l} \equiv H(C_{n+l}) = H(C_n) + H(C_l \mid C_n);\ H(C_l \mid C_n) \leqslant H(C_l),$$

где через C_k обозначена реализация длиной в k символов. Отсюда

$$H_n \leqslant H_{n+l} \leqslant H_n + H_l.$$

Левое неравенство при $l = 1$ дает

$$H_n \leqslant H_{n+1}, \tag{8.3}$$

а правое очевидным образом распространяется на любое число слагаемых и, в частности, для любого натурального k дает:

$$H_{kn} \leqslant kH_1, \tag{8.4}$$

откуда, полагая $n = 1$, получаем:

$$H_k \leqslant kH_1. \tag{8.5}$$

Так как последнее соотношение справедливо для любого k, то

$$a = \lim_{n\to\infty} \inf \left(\frac{H_n}{n} \right) < +\infty.$$

Пусть теперь $\varepsilon > 0$ задано произвольно и пусть число q выбрано так, что

$$\frac{H_q}{q} < a + \varepsilon.$$

Для любого $n > q$ определим натуральное число $k > 1$ так, чтобы

$$(k-1)\, q < n \leqslant kq.$$

Тогда, в силу соотношения (8.3)

$$H_n \leqslant H_{kq}.$$

Учитывая (8.4), получим:

$$\frac{H_n}{n} \leqslant \frac{H_{kq}}{(k-1)\, q} \leqslant \frac{k}{k-1} \cdot \frac{H_q}{q} < \frac{k}{k-1}(a + \varepsilon),$$

и, следовательно, при достаточно большом n

$$a-\varepsilon<\frac{H_n}{n}<\frac{k}{k-1}(a+\varepsilon)<a+2\varepsilon,$$

а так как ε произвольно мало, то

$$\lim_{n\to\infty}\frac{H_n}{n}=a,$$

что и требовалось доказать.

§ 9. ФУНДАМЕНТАЛЬНОЕ СВОЙСТВО ЭНТРОПИИ ДИСКРЕТНЫХ ЭРГОДИЧЕСКИХ ПРОЦЕССОВ

Введенное в предыдущем параграфе понятие энтропии (на один элемент) случайного процесса играет весьма важную роль в теории информации. Особое значение эта величина приобретает благодаря специфической связи с вероятностями $p(C)$ отдельных реализаций конечной длины; обсуждение этой связи и является целью данного параграфа.

Как указывалось в § 8, осуществление конкретной реализации C длиною в n элементов можно рассматривать как случайное осуществление одного из m^n возможных событий. Для каждого из этих событий можно определить его вероятность $p(C)$. На множестве этих событий можно задать любую числовую функцию $f_n(C)$, которая будет, очевидно, случайной величиной. В частности, такой случайной величиной будет числовая функция:

$$f_n(C)=-\frac{1}{n}\log p(C).$$

Математическое ожидание этой функции находится обычным образом:

$$Mf_n(C)=\sum_C p(C)f_n(C)=-\frac{1}{n}\sum_C p(C)\log p(C). \tag{9.1}$$

Отсюда сразу следует, что

$$M\left[-\frac{1}{n}\log p(C)\right]=\frac{H_n}{n}. \tag{9.2}$$

Так как выше мы показали, что $\lim\limits_{n\to\infty}(H_n/n)$ существует,

и обозначили его через H, назвав энтропией процесса, то и

$$\lim_{n\to\infty} M\left[-\frac{1}{n}\log p(C)\right] = H, \qquad (9.3)$$

т. е. при $n\to\infty$ математическое ожидание случайной величины $f_n(C)=-\frac{1}{n}\log p\,(C)$ имеет пределом энтропию случайного процесса H [48]. Это соотношение между H и $p(C)$, 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 $f_n(C)$ при $n\to\infty$ имеет пределом H, но и сама эта величина $f_n(C)$ стремится по вероятности к H при $n\to\infty$. Другими словами, как бы малы ни были $\varepsilon>0$ и $\delta>0$, при достаточно большом n вероятность неравенства $|f_n(C)-H|>\varepsilon$ будет меньше, чем δ; близость $f_n(C)$ к H при больших n является почти достоверным событием.

Это фундаментальное свойство эргодических процессов впервые было обнаружено К. Шэнноном [7] на примерах процесса с независимыми элементами и простой марковской цепи. Тогда же Шэннон высказал предположение, что этим свойством обладают любые эргодические процессы, однако строгое доказательство этого было найдено Б. Макмилланом лишь в 1953 г. [43]; А. Я. Хинчин [48] окончательно отшлифовал это доказательство.

Для большей наглядности сформулированное выше фундаментальное свойство („свойство E", по терминологии Хинчина) эргодических процессов обычно излагают следующим образом. Для любых заданных $\varepsilon>0$ и $\delta>0$ можно найти такое n_0, что реализации любой длины $n\geqslant n_0$ распадаются на два класса:

1) группа реализаций, общая вероятность которых не превышает δ;

2) группа реализаций, вероятности которых удовлетворяют неравенству

$$\left|\frac{1}{n}\log p\,(C)+H\right|<\varepsilon. \qquad (9.4)$$

Первую группу реализаций называют «маловероятной», вторую — «высоковероятной».

Изложенное свойство эргодических процессов приводит к ряду важных следствий, из которых три заслуживают особого внимания.

В качестве первого следствия укажем, что из (9.4) немедленно следует, что все реализации высоковероятной группы приблизительно равновероятны, независимо от того, как распределены вероятности по возможным состояниям отдельного элемента и как конкретно связаны элементы процесса между собой. В связи с этим фундаментальное свойство энтропии иногда называют «свойством асимптотической равнораспределенности». Это следствие, в частности, означает, что по известной вероятности p (C) одной из реализаций высоковероятной группы может быть оценено число N_2 реализаций в этой группе:

$$N_2 = 1/p\,(C).$$

Второе следствие заключается в том, что согласно (9,4) и первому следствию, энтропия H_n n-членной реализации случайного процесса с высокой степенью точности равна логарифму числа N_2 реализаций в высоковероятной группе.

Третьим следствием является утверждение о том, что при больших n высоковероятная группа обычно (за исключением случая равновероятных и независимых состояний элемента, т. е. $H = \log m$) охватывает лишь ничтожно малую долю всех возможных реализаций. Действительно, из (9.4) и первого следствия имеем:

$$-\log p\,(C) = \log N_2 \approx nH,$$

откуда число N_2 реализаций в высоковероятной группе

$$N_2 = a^{nH}, \tag{9.5}$$

где a—основание логарифма. С другой стороны, общее число всех возможных реализаций

$$N = m^n = a^{n \log m}. \tag{9.6}$$

Доля высоковероятных реализаций в общем числе N характеризуется отношением

$$\frac{N_2}{N} = a^{n(H - \log m)}; \tag{9.7}$$

экспоненциальная зависимость от n и подтверждает сделанное выше высказывание. Пусть, например,

$a = 2$, $n = 100$, $H = 2{,}75$, $m = 8$. Тогда $N = 8^{100} = 2^{300}$,

$N_2 = 2^{100} \cdot 2{,}75 = 2^{275}$. Отсюда $\frac{N_2}{N^1} = 2^{-25} \approx (3.10^7)^{-1}$,

т. е. к высоковероятной группе относится лишь одна тридцатимиллионная доля всех реализаций!

Строгое математическое доказательство свойства E эргодических процессов требует привлечения весьма тонких понятий и теорем из теории случайных процессов; в связи с тем, что данная книга предназначена не для математиков, здесь уместно ограничиться простым указанием на существование такого доказательства [48]. Целесообразно, однако, отметить, что в простейшем случае отсутствия статистической зависимости между элементами процесса свойство E является простым следствием закона больших чисел. Действительно, пусть мы рассматриваем длинную n-членную реализацию такого процесса. Закон больших чисел (в форме теоремы Бернулли) утверждает, что с вероятностью, близкой к 1, в этой реализации около np_1 элементов будет находиться в состоянии 1, около np_2 элементов — в состоянии 2, и т. д. Следовательно, реализации высоковероятной группы имеют вероятности, близкие к

$$p = p_1^{np_1} \cdot p_2^{np_2} \dots p_m^{np_m}.$$

Отсюда

$$-\log p = -n \sum_{i=1}^{n} p_i \log p_i = nH,$$

что, собственно, и доказывает свойство E в этом простейшем случае. Аналогично, но с привлечением свойств цепей Маркова доказывается свойство E для марковских процессов [7].

§ 10. О НЕОПРЕДЕЛЕННОСТИ И ЭНТРОПИИ НЕПРЕРЫВНЫХ СЛУЧАЙНЫХ ПРОЦЕССОВ

Обратимся теперь к попыткам дать энтропийное описание случайных процессов с непрерывными параметрами.

Рассмотрим сначала более простой случай процесса с дискретным временем, но непрерывным параметром x [46]. Задавая плотности совместных распределений все более и более высоких порядков

$$p_1(x);\ p_2(x_1, x_2);\dots;\ p_n(x_1, x_2, \dots, x_n);\dots \qquad (10.1)$$

мы будем все более и более полно характеризовать свойства такого случайного процесса. Так как информативный (слу-

чайный) параметр процесса является непрерывным, для каждого n мы можем вычислить величину дифференциальной энтропии:

$$H_n = -\int\int \ldots \int p_n(x_1, x_2, \ldots, x_n) \log [\varepsilon^n p(x_1, x_2, \ldots x_n)]\, dx_1, dx_2 \ldots dx_n. \tag{10.2}$$

Проследим, как изменяется H_n с ростом n, а также при изменении статистической связи между отдельными значениями случайного параметра процесса.

Соответственно последовательности распределений (10.1) мы будем иметь последовательность численных значений энтропии процесса:

$$H_1;\ H_2; \ldots;\ H_n; \ldots \tag{10.3}$$

В случае статистической независимости случайных величин $\{X_t\}$ (которые для простоты мы будем считать одинаково распределенными)

$$p_n(x_1, x_2, \ldots, x_n) = \prod_{t=1}^{n} p_t(x_t). \tag{10.4}$$

При этом $H_n = H_n^0 = nH_1$, что является верхней оценкой дифференциальной энтропии n-го порядка. Дальнейшие свойства последовательности (10.3) рассмотрим для наглядности на конкретном примере нормального дискретного процесса. В общем случае нормальная n-мерная плотность вероятностей записывается (см., например, [42]) в виде

$$p_n(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}} \sqrt{D_n}}$$

$$\cdot \exp\left(-\frac{1}{2 D_n \sigma^2} \sum_{i,\, k=1}^{n} D_{ik}\, x_i\, x_k\right), \tag{10.5}$$

где σ^2 — дисперсия случайной величины $x(t)$, D_n — определитель корреляционной матрицы $\|R_{ik}\|$, R_{ik} — коэффициенты корреляции, D_{ik} — их алгебраические дополнения. Вычисление интеграла (10.2) при $\varepsilon = 1$ приводит к выражению

$$H_n = \log_2 [(2\pi\sigma^2 e)^{\frac{n}{2}} \cdot D_n]. \tag{10.6}$$

Поскольку значение D_n заключено между нулем и единицей ($D_n = 1$ при отсутствии коррелированности x_i и x_k), то

ясно, что наличие корреляции приводит только к уменьшению H_n; $H_n \leqslant H_n^0$.

Для получения более конкретных результатов сделаем дальнейшее предположение о свойствах процесса. Будем считать, что процесс стационарен и коэффициент корреляции между двумя значениями, отстоящими на τ секунд друг от друга, равен

$$R(\tau) = e^{-\frac{|\tau|}{\tau_0}}. \tag{10.7}$$

Кроме того, будем считать, что параметр t меняется дискретно через одинаковые интервалы T; при этом все R_{lk} являются целочисленными степенями числа r,

$$r = \exp\left(-\frac{T}{\tau_0}\right), \tag{10.8}$$

и определитель корреляционной матрицы равен

$$D_n = \begin{vmatrix} R_{11} R_{12} R_{13} \dots R_{1n} \\ R_{21} R_{22} R_{23} \dots R_{2n} \\ \dots \dots \dots \dots \dots \\ R_{n1} R_{n2} R_{n3} \dots R_{nn} \end{vmatrix} = \begin{vmatrix} 1 & r & r^2 \dots & r^{n-1} \\ r & 1 & r \dots & r^{n-2} \\ \dots & \dots & \dots & \dots \\ r^{n-1} & r^{n-2} & r^{n-3} & \dots 1 \end{vmatrix} = (1-r^2)^{n-1}. \tag{10.9}$$

С учетом этого

$$H_n = \log_2 \left[(2\pi\sigma^2 e)^{\frac{n}{2}} \left(1 - e^{-2\frac{T}{\tau_0}} \right)^{\frac{n-1}{2}} \right]. \tag{10.10}$$

Как видим, n явно входит в выражение для H_n. Желая охарактеризовать процесс в целом, введем понятие энтропии процесса на одну степень свободы, определив ее аналогично энтропии дискретного процесса как

$$H = \lim_{n \to \infty} \frac{H_n}{n}. \tag{10.11}$$

В нашем случае

$$H = \frac{1}{2} \log_2 (2\pi\sigma^2 e) + \lim_{n \to \infty} \left(1 - \frac{1}{n}\right) \log_2 \left(1 - e^{-2\frac{T}{\tau_0}}\right)^{\frac{1}{2}} =$$

$$= H_1^0 + \log_2 \left(1 - e^{-2\frac{T}{\tau_0}}\right)^{\frac{1}{2}}, \tag{10.12}$$

где $H_1^0 = \frac{1}{2}\log(2\pi\sigma^2 e)$ — энтропия, приходящаяся на один отсчет при отсутствии коррелированности соседних значений процесса. Поскольку второй член (10.12) всегда отрицателен, то, как это и следовало ожидать, коррелированность уменьшает неопределенность.

Обратим внимание на то, что величина H может иметь любой знак, в зависимости от того, какое из слагаемых (первое — положительное, второе — отрицательное) имеет большую абсолютную величину. В соответствии с этим и H_n, рассматриваемое как функция n, может как возрастать, так и убывать с ростом n. В самом деле,

$$H_n = n\log_2\left[(2\pi\sigma^2 e)\left(1 - e^{-2\frac{T}{\tau_0}}\right)\right]^{\frac{1}{2}} -$$

$$-\log_2\left(1 - e^{-2\frac{T}{\tau_0}}\right)^{\frac{1}{2}} = nH - \log_2\left(1 - e^{-2\frac{T}{\tau_0}}\right)^{\frac{1}{2}}, \quad (n>1) \tag{10.13}$$

и рост или падение H_n с возрастанием n определяется в первую очередь знаком H. Этот странный на первый взгляд факт легко разъясняется, если мы вспомним, что речь идет об относительной энтропии, т. е. о разности неопределенностей рассматриваемого распределения и распределения, равномерного в объеме $V = \varepsilon^n$, причем стандартное распределение характеризует некоррелированные величины. Отрицательность величины H означает лишь то, что неопределенность рассматриваемого процесса на один отсчет меньше неопределенности равномерного в ε распределения. Соответственно этому, убывание H_n при росте n означает, что неопределенность рассматриваемого распределения при увеличении n возрастает медленнее, чем неопределенность равномерного в объеме $V = \varepsilon^n$ распределения n некоррелированных величин.

Можно попытаться рассмотреть неопределенность конечного отрезка длины T_0 непрерывного случайного процесса. Для этого разобьем T_0 на n равных интервалов T, затем устремим n к бесконечности, а T к нулю так, чтобы $nT = T_0$. Из (10.10) следует, что при этом

$$H_n = \log_2\left[(2\pi\sigma^2 e)^{\frac{n}{2}}\left(1 - e^{-\frac{2T_0}{n\tau_0}}\right)^{\frac{n-1}{2}}\right] =$$

$$= \frac{n}{2}\log_2(2\pi\sigma^2 e) + \frac{n-1}{2}\log_2\left(1 - e^{-\frac{2T_0}{n\tau_0}}\right). \tag{10.14}$$

Легко убедиться в том, что при $n \to \infty$ $H_n \to -\infty$. Такое поведение H_n вполне объяснимо, так как вычисление по формуле (10.2) означает, что в качестве эталонного процесса служит абсолютно некоррелированный, равномерно распределенный в интервале ε, случайный процесс, который имеет несравнимо большую неопределенность, чем рассматриваемый процесс: последнее утверждается отрицательной бесконечностью предела H_n.

Может показаться, что изложенное выше указывает на бесполезность понятия энтропии для изучения непрерывных случайных процессов. Это, конечно, не совсем так. Полезность понятия энтропии на одну степень свободы не вызывает сомнений. С другой стороны, многие важные свойства процессов с непрерывным временем могут быть изучены при рассмотрении свойств множества мгновенных значений, взятых в дискретные моменты времени; некоторые непрерывные процессы полностью определяются дискретным множеством своих значений (теорема Котельникова). Кроме того, во многих случаях для оценки ситуации можно пользоваться энтропией мгновенного значения без учета его коррелированности с остальными.

§ 11. ЭНТРОПИЙНАЯ МОЩНОСТЬ СЛУЧАЙНЫХ ПРОЦЕССОВ

В некоторых случаях при рассмотрении процессов с непрерывным информативным параметром оказывается удобным пользоваться не непосредственно дифференциальной энтропией, а производной от нее величиной, которую К. Шэннон назвал [7] энтропийной мощностью.

Понятие энтропийной мощности случайного процесса вводится следующим образом. Легко подсчитать, что дифференциальная энтропия белого гауссова шума при $\varepsilon=1$ на один отсчет равна

$$H = \log_2 \sqrt{2\pi e \sigma^2} = \log_2 \sqrt{2\pi e N},$$

где N — средняя мощность шума. Согласно теореме Котельникова, среднее число степеней свободы процесса с шириной спектра F составляет $2F$ в 1 *сек*. Отсюда, энтропия на единицу времени для рассматриваемого нами шума равна

$$H_1^0 = F \log 2\pi e N.$$

Определим энтропийную мощность некоторого случайного процесса, имеющего ширину спектра F и энтропию H', как среднюю мощность белого гауссова шума с такой же ширине-

ной спектра и такой же энтропией на степень свободы. Другими словами, если H' есть дифференциальная энтропия рассматриваемого процесса, то его энтропийная мощность равна

$$\overline{N} = \frac{1}{2\pi e} \cdot e^{2H'}. \tag{11.1}$$

Рассмотрим некоторые свойства энтропийной мощности $\overline{N}$.

1. Энтропийная мощность случайного процесса с произвольными статистическими свойствами меньше или равна его действительной средней мощности:

$$\overline{N} \leqslant N. \tag{11.2}$$

Это следует из того, что при заданной средней мощности нормальный процесс обладает максимальной энтропией (см. §. 6). Таким образом, в тех случаях, когда вычисления энтропии трудны, либо невозможны из-за незнания распределений, и в то же время задача позволяет удовлетвориться не точным указанием рассматриваемой характеристики, а лишь указанием границ, в которых она может находиться, — энтропийная мощность может оказаться полезной величиной. Следующее свойство служит дополнительным аргументом в пользу сказанного.

2. Энтропийная мощность N суммы двух независимых эргодических процессов не больше суммы средних мощностей складываемых процессов и не меньше суммы их энтропийных мощностей:

$$\overline{N}_1 + \overline{N}_2 \leqslant \overline{N} \leqslant N_1 + N_2. \tag{11.3}$$

Следует указать, что строгого доказательства этого свойства энтропийной мощности пока не имеется. Доказательство, предложенное Шэнноном [7], доводится до конца лишь при дополнительном предположении нормальности складываемых процессов, когда обе границы неравенства (11.3) сливаются. Можно, однако, привести соображения, говорящие за то, что это свойство величины N в некоторых типичных случаях справедливо. Эти соображения таковы. Согласно первому свойству энтропийной мощности максимум энтропийной мощности суммарного процесса равен его средней мощности. С другой стороны, известно, что при сложении независимых (некогерентных) эргодических процессов средняя мощность суммарного процесса равна сумме средних мощностей слагаемых процессов. Тем самым показывается справедливость правой части неравенства (11.3). Однако, поскольку результирующий процесс будет строго нормальным лишь при очень частных

предположениях, то его энтропийная мощность будет в общем случае меньше суммы средних мощностей слагаемых процессов. Это ставит вопрос об оценке нижней границы N. Известно, что при многократном сложении произвольно распределенных случайных величин распределение суммы при весьма общих условиях стремится к нормальному. Этот эффект хотя и слабо выражен, но имеет место, конечно, и при сложении только двух величин. Поэтому естественно ожидать (хотя это не является, конечно, доказательством), что вследствие этой «нормализации», энтропийная мощность суммы ненормальных процессов будет больше суммы их энтропийных мощностей.

3. Энтропийная мощность гауссова шума, не являющегося белым и характеризующегося спектром мощности $N(f)$ в полосе F, выражается формулой

$$\overline{N} = \exp\left\{\frac{1}{F}\int_F \log N(f)\, df\right\}. \qquad (11.4)$$

Доказательство этого свойства энтропийной мощности будет дано в следующем параграфе.

§ 12. ВЫРАЖЕНИЕ ДИФФЕРЕНЦИАЛЬНОЙ ЭНТРОПИИ НА СТЕПЕНЬ СВОБОДЫ СТАЦИОНАРНОГО ГАУССОВА ПРОЦЕССА ЧЕРЕЗ ЕГО СПЕКТР

Из рассуждений, приведенных в § 10, следует, что энтропия на одну степень свободы „белого"*) гауссова процесса со средней мощностью $N = \sigma^2$ равна

$$H_1^0 = \frac{1}{2}\log(2\pi e \sigma^2). \qquad (12.1)$$

Если полоса частот, занимаемая спектром этого процесса, равна F, то энтропия на единицу времени выразится величиной

$$H_F^0 = F\log(2\pi e \sigma^2). \qquad (12.2)$$

Вычислим теперь энтропию на степень свободы для стандартного гауссова процесса, спектр мощности которого не является равномерным в заданном интервале частот F, а выражается функцией $N(f)$. Допустим, что $N(f)$ является дифференцируемой функцией (за исключением, может быть,

*) Белым называется процесс, имеющий равномерный спектр мощности.

множества точек меры нуль); тогда оказывается возможным разбить полосу частот F на малые «полоски» df, на протяжении которых $N(f)$ можно с высокой степенью точности считать константой. Теперь рассматриваемый нами процесс можно представить как совокупность множества белых гауссовых процессов. Величина энтропии на единицу времени для элементарного процесса, спектр которого лежит в интервале $(f, f+df)$, выразится, в соответствии с формулой (12.2), величиной

$$df \cdot \log [2\pi e N(f)]. \tag{12.3}$$

Так как энтропия совокупности независимых процессов равна сумме соответствующих энтропий, то для результирующего процесса энтропия на единицу времени может быть записана в виде

$$H_F = \int_F \log [2\pi e N(f)]\, df. \tag{12.4}$$

Отсюда окончательно следует, что величина дифференциальной энтропии на степень свободы гауссового процесса с произвольным спектром мощности $N(f)$ выражается формулой*)

$$H = \frac{1}{2F} \int_F \log [2\pi e N(f)]\, df. \tag{12.5}$$

Сопоставляя эту формулу с выражением H через функцию автокорреляции (см. § 10), можно получить полезные и интересные соотношения [44]. Из формул (10.6) и (10.11) следует, что

$$H = \lim_{n\to\infty} \frac{1}{n} \log [(2\pi e \sigma^2)^{\frac{n}{2}} \cdot \overline{D_n}]. \tag{12.6}$$

Кроме того, очевидно, что величину H можно также выразить как $H = \lim\limits_{n\to\infty} (H_n - H_{n-1})$, и, следовательно,

$$H = \lim_{n\to\infty} \frac{1}{2} \log \left[(2\pi e \sigma^2) \frac{D_n}{D_{n-1}} \right]. \tag{12.7}$$

Сопоставляя (12.5), (12.6) и (12.7), получаем

$$\lim_{n\to\infty} \frac{1}{2n} \log D_n = \frac{1}{2F} \int_F \log \frac{N(f)}{\sigma^2}\, df, \tag{12.8}$$

*) Отметим, что (12.5) получена в предположении независимости значений спектра в соседних интервалах; случай зависимых значений спектра должен рассматриваться дополнительно.

$$\lim_{n\to\infty}\frac{1}{2}\log\frac{D_n}{D_{n-1}}=\frac{1}{2F}\int_F\log\frac{N(f)}{\sigma^2}\,df. \tag{12.9}$$

Эти соотношения были получены математиками задолго до появления теории информации (Пойа, 1915; Сегё, 1921), но физическое содержание этих результатов оставалось не вскрытым. Полезность соотношений (12.8) и (12.9) проявляется при необходимости оценить величины n-мерных корреляционных определителей, трудно поддающихся непосредственному вычислению. Важно отметить, что если случайный процесс, имеющий корреляционную матрицу $\|R_{ik}\|$ и спектр $N(f)$, не является гауссовым, то в силу экстремальности энтропии гауссова процесса при заданных вторых моментах (см. § 6) правые части соотношений (12.8) и (12.9) могут служить верхней оценкой соответствующих выражений из n-мерных определителей, стоящих в левой части.

Отметим теперь, что соотношение (12.5) позволяет доказать свойство (11.4) энтропийной мощности, которое было дано в § 11 без доказательства. Подставляем в (11.1) выражение (12.5) для энтропии на степень свободы:

$$\overline{N}=\frac{1}{2\pi e}e^{2H}=\frac{1}{2\pi e}\exp\left\{\frac{1}{F}\int_F\log\left[2\pi e N(f)\right]df\right.$$

$$=\exp\left\{\frac{1}{F}\int_F\log N(f)\,df\right\}, \tag{12.10}$$

что и доказывает справедливость (11.4).

§ 13. ИЗМЕНЕНИЕ ДИФФЕРЕНЦИАЛЬНОЙ ЭНТРОПИИ ПРИ ЛИНЕЙНОЙ ФИЛЬТРАЦИИ

В предыдущем параграфе было показано, что дифференциальная энтропия на степень свободы для гауссова процесса полностью определяется его статистическим спектром $N(f)$ (см. формулу (12.5)). Это позволяет выразить количественно изменение при таких преобразованиях нормального процесса, которые приводят к изменению его спектра, но не изменяют тип распределения, оставляя его нормальным. К числу таких преобразований относится л и н е й н а я ф и л ь т р а ц и я, т. е. преобразование процесса при прохождении его через линейный частотный фильтр.

Действие линейного частотного фильтра на входной сигнал $x(t)$ сводится к тому, что каждая частотная компонен-

та амплитудного спектра функции $x(t)$ умножается на коэффициент передачи фильтра на данной частоте, $Y(f)$. $Y(f)$, рассматриваемое как функция частоты f, называется частотной характеристикой (или характеристикой передачи) фильтра. Условие линейности фильтра означает, что для любой частоты f коэффициент передачи фильтра $Y(f)$ не зависит от амплитуды соответствующей частотной компоненты входного сигнала.

Пусть $X(f)$ — спектр комплексных амплитуд, или спектральная функция входного сигнала $x(t)$. Тогда спектральная функция выходного сигнала фильтра выразится как $X_1(f) = X(f) \cdot Y(f)$. Спектр мощности сигнала $x(t)$ связан со спектральной функцией соотношением

$$N(f) = X(f) X^*(f) = |X(f)|^2. \tag{13.1}$$

Соответственно, спектр мощности выходного сигнала выразится формулой

$$N_1(f) = X_1(f) X_1^*(f) = N(f) \cdot |Y(f)|^2. \tag{13.2}$$

Воспользовавшись выражением (12.5) для энтропии на степень свободы нормального процесса, получаем:

$$H(x_1) = \frac{1}{2F} \int_F \log [2\pi e N_1(f)]\, df =$$

$$= \frac{1}{2F} \int_F \log [2\pi e N(f)]\ Y(f)|^2]\, df =$$

$$= \frac{1}{2F} \int_F \log [2\pi e N(f)]\, df + \frac{1}{2F} \int_F \log |Y(f)|^2\, df =$$

$$= H(x) + \frac{1}{2F} \int_F \log |Y(f)|^2\, df. \tag{13.3}$$

Таким образом, при прохождении нормального процесса через линейный фильтр с частотной характеристикой $Y(f)$ дифференциальная энтропия на степень свободы изменяется на величину

$$\Delta H = \frac{1}{2F} \int_F \log |Y(f)|^2\, df. \tag{13.4}$$

Если фильтр содержит только пассивные элементы (линейные сопротивления, конденсаторы и катушки самоиндукции), то $|Y(f)| \leqslant 1$, и следовательно, $\Delta H < 0$. Поэтому обычно говорят о „потере энтропии в линейных фильтрах (см. например, [7]). Однако в общем случае линейный четырехполюсник может содержать и линейные усилители; при этом ΔH может быть как отрицательным, так и положительным или нулем. Это лишний раз подчеркивает относительность дифференциальной энтропии: стандарт сравнения ε остается постоянным, тогда как ослабление или усиление сигнала изменяет его масштаб.

Хотя соотношение (13.4) получено в предположении нормальности фильтруемого процесса, важность этого соотношения несомненна, поскольку во многих практически важных случаях мы имеем дело с нормальными или близкими к ним процессами. Часто, однако, этот результат считают автоматически распространяющимся и на случай процессов с произвольным (не нормальным) распределением. К такому обобщению, проводимому без доказательства, следует отнестись весьма осторожно, если не отрицательно. Во-первых, нормальность процесса является существенным предположением в ходе рассуждений, из которых следует (13.4). Во-вторых, известно, что только в случае нормальности процесса задание его первых двух моментов (или их эквивалента — спектра мощности) однозначно определяет распределение вероятности и, следовательно, величину энтропии. Это свойство нормального процесса и позволило выразить H_ε только через $N(f)$. Для негауссова процесса по известному спектру мощности в общем случае нельзя однозначно определить распределение вероятностей и H_ε. Следовательно, можно полагать, что и изменение спектра при фильтрации, даже если оно известно, не может в общем случае однозначно определить изменение H_ε. Однако эти соображения лишь утверждают, что возможность применения (13.4) к фильтрации негауссовых процессов не доказана, но полностью не исключают такой возможности. Было бы интересно подвергнуть этот вопрос строгому рассмотрению.

§ 14. ТЕОРЕТИЧЕСКИЕ И ЭКСПЕРИМЕНТАЛЬНЫЕ ОЦЕНКИ ЭНТРОПИИ СЛУЧАЙНЫХ ВЕЛИЧИН И СТАЦИОНАРНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

В практике может возникнуть необходимость дать количественную оценку энтропии некоторой случайной величины или процесса, не пользуясь при этом конкретным видом распределений вероятностей. Такая проблема возникает, например, в тех случаях, когда либо функция $p(x)$ такова,

что значение интеграла энтропии $-\int p(x) \log p(x)\, dx$ может быть получено лишь путем громоздких приближенных вычислений; либо конкретный вид функции $p(x)$ вообще неизвестен, но известны некоторые свойства этой функции (например, ее моменты). В этих условиях можно поставить задачу нахождения численных оценок величины $H(x)$ теоретическим путем. С другой стороны, большое значение может иметь проблема измерения энтропии, т. е. нахождение экспериментальной оценки энтропии по данным опыта.

1. Теоретические оценки энтропии. Если конкретный вид функции $p(x)$ неизвестен, но некоторые характеристики этой функции заданы, то принцип экстремума энтропии (см. § 6) и другие соображения позволяют оценить границы для энтропии. Приведем некоторые оценки энтропии при различных предположениях.

а) Пусть известно, что $p(x)$ отлично от нуля лишь в интервале величиной D. Тогда, в силу максимальности энтропии равномерного в интервале D распределения,

$$H \leqslant \log D. \tag{14.1}$$

б) Известно, что максимальное значение $p(x)$ равно $P_{\max}$. Тогда

$$H \geqslant \log \frac{1}{P_{\max}}, \tag{14.2}$$

так как минимальную энтропию при этом условии имеет равномерное распределение.

в) Если среднее значение абсолютного отклонения x от среднего значения x_0 равно A, то

$$H \leqslant \log 2eA, \tag{14.3}$$

так как наибольшей энтропией обладает распределение

$$p(x) = \frac{1}{2A} \exp\left(-\frac{|x - x_0|}{A}\right).$$

г) Если задано стандартное отклонение σ случайной величины x, то, очевидно, в силу максимальности энтропии гауссового распределения при этом условии, имеем оценку

$$H \leqslant \frac{1}{2} \log 2\pi e \sigma^2. \tag{14.4}$$

Ряд интересных оценок энтропии в том случае, когда $p(x)$ известно, но интеграл энтропии не приводится к известным функциям, можно получить с помощью разложения $p(x)$ или $\log p(x)$ в ряды [39].

Разложим функцию $\log p(x)$ в степенный ряд Тейлора около точки $x_0 = \overline{x}$:

$$\log p(x) = \log p(x_0) + \log' p(x_0) \cdot (x - x_0) +$$

$$+ \frac{1}{2} \log'' p(x_0)(x - x_0)^2 + \ldots$$

Подставив это разложение в подынтегральное выражение, ограничившись тремя членами, получим верхнюю и нижнюю оценки величины $H(x)$:

$$-\log p(x_0) - \frac{1}{2}\sigma^2 \log''_{\max} p(x) \leqslant H \leqslant$$

$$\leqslant -\log p(x_0) - \frac{1}{2}\sigma^2 \log''_{\min} p(x). \qquad (14.5)$$

Здесь $\log''_{\max} p(x)$ и $\log''_{\min} p(x)$—экстремальные значения второй производной функции $\log p(x)$. Обе грани совпадают для равномерного, экспоненциального и нормального распределений.

Аналогичным способом можно получить [39] следующие оценки энтропии:

$$\frac{3}{2}\left(-\frac{2\pi}{\log''_{\min} p}\right)^{\frac{1}{2}} \cdot p_{\max} - \log e\, p_{\max} \leqslant H \leqslant$$

$$\leqslant \frac{3}{2}\left(-\frac{2\pi}{\log''_{\max} p}\right)^{\frac{1}{2}} p_{\max} - \log e\, p_{\max}, \qquad (14.6)$$

(если $\log''_{\max} p > 0$, то по этой оценке верхней границы энтропии нет).

$$\frac{3}{2}\left(\frac{\log''_{\max} p}{\log''_{\min} p}\right)^{\frac{1}{2}} - \log e\, p_{\max} \leqslant H \leqslant \qquad (14.7)$$

$$\leqslant \frac{3}{2}\left(\frac{\log''_{\min} p}{\log''_{\max} p}\right)^{\frac{1}{2}} - \log e\, p_{\max}.$$

2. Оценка энтропии по экспериментальным данным. Можно указать несколько способов получения

оценки энтропии путем обработки результатов опыта. Наиболее очевидный способ в дискретном случае состоит в определении относительных частот появления отдельных символов и последующем вычислении энтропии в предположении, что частоты являются хорошими оценками вероятностей. Очевидно, что чем больше объем выборки, по которой определены частоты, тем более точно оцениваются вероятности, и тем более больше точность полученной оценки энтропии. Этот метод оценки энтропии оказывается очень трудоемким, если число возможных символов велико или если большая часть вероятностей мала. Кроме того, довольно сложен вопрос о зависимости погрешности оценки энтропии от величины объема выборки; поэтому приходится брать выборку максимально большой в заданных условиях.

Другой способ оценки энтропии основывается на принципе экстремальности энтропии. С одной стороны, известно, что при заданной функции корреляции максимальной энтропией обладает гауссов процесс. С другой стороны, сейчас хорошо разработана техника экспериментального определения функции корреляции случайных процессов или коэффициентов корреляции случайных величин. Так как энтропия дискретного по времени нормального процесса равна (см. § 8)

$$H(x) = \log\left[(2\pi e)^{\frac{n}{2}}\sqrt{D}\right], \qquad (14.8)$$

(где D — определитель корреляционной матрицы), то, вычислив по полученным экспериментально коэффициентам корреляции величину D, мы получим с помощью (14.8) верхнюю оценку энтропии изучаемого процесса (или «точное» значение, если процесс гауссов).

Несмотря на приемлемость в отдельных случаях рассмотренных выше методов экспериментальной оценки энтропии, остается некоторая неудовлетворенность в связи с тем, что энтропия при этом «не измеряется непосредственно».

В первом случае, например, приходится оценивать сначала распределение вероятностей, а затем лишь находить его числовую характеристику H. Такая процедура представляется столь же неэкономной, как и, например, нахождение распределения с единственной целью — оценить затем среднее значение случайной величины. Естественно поставить проблему отыскания способов непосредственной оценки энтропии по характеристикам реализации, не требующих предварительных оценок вероятностей. Один из таких способов предложен Р. Л. Добрушиным [40]. Он показал, что характеристикой, по которой можно оценивать энтропию, является среднее значение интервала времени между появлениями одинаковых символов.

Рассмотрим метод Р. Л. Добрушина на примере последовательности независимых случайных величин. Введем следующие обозначения: ξ_k — значение случайной величины X, реализовавшееся на k-м шаге после нулевого; $\{x_e\}$ — множество возможных значений случайной величины X; x_e — одно из них; p_e — вероятность появления символа x_e; η — случайная величина, равная наименьшему положительному k, для которого $\xi_k = \xi_0$ (т. е. η — интервал времени между двумя соседними появлениями одинаковых символов). Очевидно, что условная вероятность

$$P(\eta > j | \xi_0 = x_e) = (1 - p_e)^j \cdot p_e. \tag{14.9}$$

Если вероятность p_e мала, то геометрическое распределение (14.9) будет близко к экспоненциальному:

$$P(\eta > t | \xi_0 = x_e) \approx \exp(-tp_e) \cdot p_e. \tag{14.10}$$

Рассмотрим теперь условное математическое ожидание случайной величины $\log \eta$. В силу (14.10) имеем:

$$M(\log \eta | \xi_0 = x_e) \approx p_e \int_0^\infty \log t \cdot \exp(-tp_e)\, dt = \tag{14.11}$$

$$= \int_0^\infty \lg y \exp(-y)\, dy - \log p_e = -C - \log p_e,$$

где $C = 0{,}577...$ — постоянная Эйлера. Легко видеть, что (в предположении, что все p_e малы) безусловное математическое ожидание случайной величины $\log \eta$ равно

$$M(\log \eta) \approx H - C, \tag{14.12}$$

где H — энтропия последовательности на один символ. Из соотношения (14.12) и вытекает метод непосредственной экспериментальной оценки энтропии. По данной реализации необходимо получить некоторое число N независимых наблюдений y_i $(i = 1,2,...N)$ величины η. Если N достаточно велико, то среднее арифметическое значение величины $\log y$ будет достаточно близко (закон больших чисел) к математическому ожиданию этой величины, и, следовательно,

$$\frac{1}{N}(\log y_1 + \log y_2 + ... + \log y_N) + C = H. \tag{14.13}$$

Левая часть формулы (14.13) и является искомой оценкой для энтропии.

Для практического применения этого метода необходимо уточнить количественно достаточное число слагаемых N и требование малости на вероятности p_e. Обычные способы оценки отклонения среднего статистического от математического ожидания позволяют вычислить число N, достаточное для достижения заданной точности. Что касается уточнения требований к малости вероятностей p_e то можно было бы исходить из оценки разности правой и левой частей соотношения (14.11). Однако Р. Л. Добрушин показал, как можно обойти необходимость такой оценки. Если среди вероятностей $\{p_e\}$ имеются как малые, так и относительно большие, то нужно, выбрав подходящим образом $\varepsilon > 0$, разбить все значения на два подмножества: исходы A, для которых $p_e > \varepsilon$, и исходы $\bar{A}$, для которых $p_e \leqslant \varepsilon$. Для $x_e \in A$ (таких значений не больше $1/\varepsilon$) надо оценить вероятности p_e путем нахождения частот. Образуем теперь случайную величину

$$\tilde{\eta} = \begin{cases} \log \eta, & \text{если } \xi_0 \in \bar{A} \\ -\log p_e, & \text{если } \xi_0 \in A,\ \xi_0 = x_e. \end{cases}$$

Легко видеть, что

$$M(\log \tilde{\eta}) \approx H - C,$$

что и позволяет обобщить описываемый метод оценки энтропии на случай произвольных вероятностей $\{p_e\}$.

Вопросы уточнения оценки энтропии при наличии статистической связи между элементами последовательности требуют дальнейшего рассмотрения. Некоторые соображения по этому поводу приводятся в заметке Р. Л. Добрушина [40].

ГЛАВА V

КОЛИЧЕСТВО ИНФОРМАЦИИ

§ 1. ИНФОРМАЦИЯ И КОЛИЧЕСТВО ИНФОРМАЦИИ

В основе всей теории информации лежит открытие, заключающееся в том, что информация допускает количественную оценку. Наиболее четко, вплоть до введения количественной меры информации, эта мысль, по-видимому, впервые была высказана Хартли в 1928 г. [70], а затем, уже на более высоком уровне, развита и обобщена Шэнноном, Винером, фон Нейманом, Фишером, Колмогоровым и другими. В данной главе приводится последовательное (несколько упорядоченное, но в общем соответствующее историческому развитию) изложение того, как наполнялось конкретным содержанием понятие количества информации. Однако, прежде чем непосредственно анализировать это понятие, целесообразно кратко обсудить соотношение его с понятием информации.

Как это ни парадоксально звучит, но для развития теории информации в ее современном виде вообще не требуется определения понятия информации как таковой; необходимым и достаточным для построения теории является понятие количества информации*). Поэтому употребление терминов «информация» и «количество информации» как синонимов не вызывает недоразумений в рамках самой теории [51, 63]. Но, имея в виду более широкое употребление этих терминов, необходимо рассматривать их как имеющие различное содержание.

Анализ понятия информации и разработка достаточно общего определения являются задачами внешними по отношению к теории информации, носящими методологический, философский характер. В целом ряде работ, посвященных кибернетике (см. [49—55, 58—60]), эти задачи обсуждаются

*) Это не должно казаться странным: такое положение характерно и для ряда других количественных теорий. Например, для изложения механики нужны лишь количественные характеристики движения, но не требуется анализа существа самого движения.

и находят свое разрешение. Несмотря на продолжающиеся дискуссии, можно считать, что база для разработки определения информации уже заложена. Во-первых, анализ этого понятия показал, что информация является не некоторой нематериальной субстанцией, а есть свойство материи. Во-вторых, стало ясно, что понятие информации в кибернетике родственно с понятием отражения, рассматриваемым диалектическим материализмом [56]. В соответствии с реальной действительностью диалектический материализм утверждает, что все материальные тела находятся во взаимосвязи, взаимодействии друг с другом. В силу этого взаимодействия всякое изменение состояния одного объекта приводит к изменению состояний других, взаимодействующих с ним объектов. Свойство отражения состоит в том, что между состояниями взаимодействующих объектов существует определенная связь, соответствие, изоморфность. Свойство отражения присуще не только объектам, но и процессам (т. е. изменениям объектов) и проявляется в наличии соответствия между отражающими друг друга процессами; это соответствие может носить иногда весьма сложный характер. С другой стороны, мы говорим, что «объект (или процесс) А содержит в себе информацию об объекте (или процессе) В», именно в тех случаях, когда между состояниями объектов (или процессов) А и В существует соответствие. Будем ли мы иметь в виду соответствие между нашими ощущениями и реальностью или соответствие между положением стрелки вольтметра и напряжением на его клеммах — во всем широчайшем диапазоне подобных ситуаций один объект отражает другой, один объект содержит информацию о другом.

Таким образом, понятие информации в кибернетике и понятие отражения в философии являются разными абстракциями одного и того же свойства материи. Информация и отражение базируются на соответствии между состояниями материальных объектов, на соответствии между процессами.

Это отнюдь не означает, что кибернетика охватывает или заменяет философскую теорию отражения. Здесь мы имеем яркий пример того, как определенный круг явлений и отношений рассматривается с разных точек зрения, — с позиций философии (теория познания, теория отражения) и с позиций естественных, точных наук (кибернетика, теория информации). Философию прежде всего интересуют качественные различия между типами отражения (информации); теория информации занимается количественным описанием этого свойства материи.

Но и резкое противопоставление этих подходов неправомерно. При всем различии методов изучения, при всей специфике каждого из подходов оба они изучают одно и то же

свойство материи — откуда следует их общность. Именно эта общность обеспечивает им взаимное обогащение, обоюдное стимулирование к развитию, что весьма наглядно проявляется во взаимной постановке новых проблем или вскрытии новых аспектов старых проблем. Так, перед теорией информации сейчас стоит задача учета качественных особенностей типов информации. К числу таких особенностей могут быть отнесены ценность информации, степень правдивости, осмысленность. Первые шаги в этом направлении уже сделаны (см. [57], [62]). В свою очередь, теория информации по-новому ставит перед теорией познания ряд проблем, например, проблему количества информации. До сих пор эта проблема рассматривалась лишь в общем гносеологическом плане как проблема адекватности отражения: признается наличие менее полного и более полного знания, менее правильного и более правильного отражения; утверждается, что познание в своем историческом развитии имеет тенденцию ко все более полному и более правильному отражению внешнего мира. Здесь вполне отчетливо признается количественная определенность информации. Однако лишь в последнее время стали обсуждаться более конкретные философские аспекты этого свойства информации.

При построении теории, описывающей информационные явления, — теории информации, — естественно начать с описания наиболее простых, элементарных информационных отношений, то есть отношений отражения между объектами неживой природы (в том числе и в первую очередь между техническими объектами). В этих отношениях обычно не проявляют себя (и, следовательно, могут не учитываться) такие свойства информации, как смысл, доброкачественность (степень правдивости), ценность, и т. д. — свойства, весьма существенные для информационных отношений с участием живых организмов. Таким образом, отказ от учета смысла, ценности, качества информации является не недостатком теории информации, не искусственным ограничением, накладываемым ради простоты (как это иногда представляют), а естественным следствием того, что рассматриваются лишь те информационные отношения, в которых эти свойства информации не проявляются. В связи с этим и количество информации является характеристикой, лишь с одной стороны описывающей информационные отношения в реальном мире. Однако именно эта сторона — соответствие состояний — играет главную роль в технических устройствах; поэтому-то теория информации имеет важное значение при современном подходе к проектированию и изучению различных технических информационных систем. Это, конечно, не означает, что высшие формы отражения не допускают коли-

чественного описания и что полученные сейчас теорией информации результаты не будут иметь значения при таком описании. Представляется логичным, что теория информационных отношений с участием высокоорганизованных систем явится обобщением современной теории информации, будет включать последнюю как частный, предельный случай, аналогично тому, как релятивистская механика обобщает и включает в себя классическую. Можно даже предположить, что существенную роль в этой общей теории будет играть и такая характеристика, как степень организованности рассматриваемой системы. Все это, однако, дело будущего, вряд ли стоит строить здесь далеко идущие догадки.

§ 2. ПРОСТЕЙШИЙ СЛУЧАЙ. КОЛИЧЕСТВО ИНФОРМАЦИИ ПО Р. ХАРТЛИ

Количественная мера информации должна, конечно, согласовываться с интуитивными представлениями о содержании информации в сигнале. Интуитивные представления, в свою очередь, обычно основываются на «ощущении» объективных закономерностей, хотя и не сформулированных полно и точно. Мы, например, хорошо понимаем, что чем длиннее телеграмма, тем больше информации она обычно содержит; следовательно, вводимая мера информации должна монотонно возрастать с увеличением длительности сигнала, которую естественно измерять числом символов в дискретном сигнале и временем передачи в непрерывном случае.

С другой стороны, интуитивно чувствуется, что количество информации связано не только с длительностью сигнала, но и с другими особенностями построения сигнала. Среди таких особенностей особо стоит отметить зависимость количества информации от числа употребляемых элементов сигнала. Например, очевидно, что при пятибалльной системе оценок полученная оценка более полно характеризует состояние знаний обучающегося, чем оценка по двухбалльной системе. Качество телевизионного изображения (а с ним и количество информации об изображаемом предмете) тем выше, чем больше градаций яркости между «черным» и «белым» способна передавать телевизионная система. Другими словами, количество информации на один элемент сигнала тем больше, чем больше число возможных элементов; этим свойством должна обладать и вводимая мера информации.

Имеются и другие факторы, влияющие на содержание информации в сигнале. Как было отмечено в § 2 гл II, всякий сигнал должен рассматриваться как случайный процесс. Очевидно, статистические характеристики такого процесса тоже должны влиять на содержание информации в сигнале.

Как видно, построение количественной меры информации, удовлетворяющей приводимым выше требованиям, является не простой задачей. Эта задача сводится к отысканию некоторого числа, монотонно возрастающего с увеличением длительности и увеличением числа возможных элементов сигнала, и подходящим образом изменяющегося при изменении статистических характеристик сигнала. Бросается, однако, в глаза, что таким числом, или основой для построения такого числа, по-видимому, может служить число различимых реализаций, образующих процесс, или (как иногда принято говорить) число различных сигналов. В самом деле, число различных реализаций процесса изменяется при варьировании указанных выше фактов именно так, как это требуется приведенными выше соображениями.

Итак, может быть высказано предположение. что основой для построения количественной меры информации может служить число N различимых реализаций случайного процесса, используемых в качестве сигналов.

Чтобы проверить это предположение и воспользоваться им, если оно окажется справедливым, необходимо выяснить, в какой степени число N удовлетворяет дополнительным требованиям к мере информации, о которых пока не шло речи. Естественно сделать такую проверку на каком-то частном примере, и логично обратиться к рассмотрению самого простейшего случая. Такой простейший случай определяется следующими условиями.

1. Состояние наблюдаемого объекта (сигнал) однозначно определяется состоянием объекта-оригинала, т. е. влияние других, не интересующих нас объектов на сигнал отсутствует или настолько несущественно, что им можно пренебречь. В частном случае техники связи это соответствует условию отсутствия помех, шумов и неоднозначных преобразований.

2. Сигнал дискретен как по времени, так и по информативным параметрам. Такой сигнал является последовательностью сменяющих друг друга различных состояний объекта; в технике дискретной связи эти состояния рассматриваются как символы.

3. Множество различимых состояний (т. е. множество символов, или алфавит) не только дискретно, но и конечно.

Смена состояний (появление новых символов) происходит таким образом, что

4) все состояния (символы) являются равновероятными,

5) вероятностные связи между различными символами от-

сутствуют, т. е. символы являются статистически независимыми.

В этом простейшем случае (впервые рассмотренном Р. Хартли [70] в 1928 г.) число N различных последовательностей длиною в n символов из алфавита, содержащего m символов, равно m^n. Хартли указал, что экспоненциальная зависимость N от n не позволяет использовать величину N в качестве меры информации: «Рассматривая физическую систему передачи, мы не обнаруживаем такого экспоненциального нарастания качеств, необходимых для передачи результатов последовательных выборов. Различные символы одинаково различаются на приемном конце как при первом выборе, так и при любом другом. Телеграф передает десятое слово известия не с большим трудом, чем предшествующее. Телефон, успешно передающий речь, продолжает и впредь это делать до тех пор, пока свойства системы остаются неизменными».

Таким образом, N не может служить непосредственно мерой количества информации. Однако логично использовать величину N как основу для построения меры, удовлетворяющей практическим требованиям. Эта мера, количество информации I, должна быть, следовательно, однозначно связана с N.

Очевидно, мера количества информации должна удовлетворять естественному требованию аддитивности, которое заключается в том, что при прочих равных условиях количество информации I пропорционально длине n сигнала;

$$I = K \cdot n, \tag{2.1}$$

где K — некоторая константа. В рассматриваемом нами случае (равновероятных и независимых символов) эта константа зависит только от m. Чтобы найти зависимость $K(m)$, рассмотрим передачу одного и того же количества информации I с помощью двух различных алфавитов, содержащих соответственно m_1 и m_2 символов. Поскольку передается одно и то же количество информации I, то в обоих случаях число возможных последовательностей будет одинаково:

$$N = m_1^{n_1} = m_2^{n_2}. \tag{2.2}$$

С другой стороны,

$$I = K(m_1) \cdot n_1 = K(m_2) \cdot n_2. \tag{2.3}$$

Из сопоставления (2.2) и (2.3) следует, что

$$\frac{K(m_1)}{\log m_1} = \frac{K(m_2)}{\log m_2}. \tag{2.4}$$

Следовательно,

$$K(m) = K_0 \log m, \tag{2.5}$$

где K_0 — произвольная константа и логарифм может иметь произвольное основание. Подставляя (2.5) в (2.1) и приняв $K_0 = 1$, получаем окончательно, что в рассматриваемом случае

$$I = n \cdot \log m = \log m^n = \log N. \tag{2.6}$$

Сформулируем, следуя Хартли, окончательное определение: в качестве меры количества информации принимается логарифм числа возможных последовательностей символов.

Чтобы впоследствии отличать это определение от других, будем называть его первым определением количества информации.

§ 3. ВЕРОЯТНОСТНЫЙ ПОДХОД К. ШЭННОНА К ОПРЕДЕЛЕНИЮ КОЛИЧЕСТВА ИНФОРМАЦИИ. СНЯТИЕ УСЛОВИЯ РАВНОВЕРОЯТНОСТИ СИМВОЛОВ

В соответствии с определением Хартли, вычисление количества информации I сводится к нахождению (и логарифмированию) числа возможных сигналов. Заметим, что в рассмотренном выше случае равновероятности и независимости символов при любом n все возможные сигналы оказываются также равновероятными; вероятность каждого из таких сигналов равна $p = 1/N$, т. е. обратно пропорциональна искомому числу N. В связи с этим можно дать второе определение количества информации.

При условии равновероятности возможных сигналов количество информации, несомое любым конкретным сигналом, равно минус логарифму вероятности отдельного сигнала*).

Это определение позволяет значительную часть комбинаторных вычислений заменить выкладками, типичными для теории вероятностей, что сильно облегчает расчеты.

Наша дальнейшая задача состоит теперь в обобщении полученных результатов путем постепенного снятия ограничений, наложенных условиями 1—5 предыдущего параграфа. Такое обобщение мы проведем на основе второго определения количества информации.

---

*) При этом по-прежнему предполагается выполнение условия отсутствия помех и шумов.

Первое ограничение, которое мы снимем, это – условие равновероятности символов (условие 4). Пусть вероятность i-го символа ($i=1, 2...., m$) равна p_i; символы образуют полную группу событий, т. е. $\sum_{i=1}^{m} p_i = 1$.

Чтобы иметь право воспользоваться вторым определением количества информации, необходимо выполнить весьма существенное требование равновероятности возможных сигналов. В нашем случае это требование сводится к тому, чтобы относительные частоты появления отдельных символов во всех возможных сигналах были равны. Строго говоря, при любом конечном n это условие не выполняется. Однако, в силу теоремы Бернулли, с ростом n относительные частоты n_i/n стремятся (по вероятности) к соответствующим вероятностям p_i. Это дает нам возможность утверждать, что при достаточно больших n условие равновероятности возможных сигналов будет приближенно выполняться.

Найдем теперь формулу для вычисления количества информации в дискретном сигнале из независимых неравновероятных символов.

В силу статистической независимости символов, вероятность сигнала длиной в n символов равна

$$p = \prod_{j=1}^{n} p_j, \tag{3.1}$$

где p_j — вероятность j-го символа, $j = 1,2,...,n$ — порядковый номер символа в сигнале. Если i-й символ повторяется в данном сигнале n_i раз, то

$$p = \prod_{i=1}^{m} p_i^{n_i}. \tag{3.2}$$

Как отмечалось выше, при достаточно больших n $n_i \approx np_i$. Поэтому

$$p = \prod_{i=1}^{m} p_i^{np_i}. \tag{3.3}$$

Отсюда окончательно получим:

$$I = -\log p = -n \sum_{i=1}^{m} p_i \log p_i. \tag{3.4}$$

Это соотношение носит название формулы Шэннона. Как и следовало ожидать, формула Хартли (2.6) является частным случаем формулы Шэннона. В самом деле, при равновероятных символах, т. е. при $p_i = 1/m$, формула Шэннона переходит в формулу Хартли:

$$I = -n\sum_{i=1}^{m} p_i \log p_i = -n\sum_{i=1}^{m} \frac{1}{m}\log\frac{1}{m} = n\log m. \qquad (3.5)$$

Обращает на себя внимание тот факт, что формула Шэннона (3.4) для количества информации совпадает с формулой для величины энтропии множества возможных сигналов. Это совпадение не должно казаться случайным и заслуживает того, чтобы быть рассмотренным более детально, что и будет сделано в § 5 настоящей главы. Здесь же мы пока ограничимся простой констатацией этого важного факта: при отсутствии ошибок при приеме среднее количество информации на сигнал численно равно энтропии множества возможных сигналов. Необходимо сразу же предостеречь, что из этого не следует, что энтропия и количество информации это одно и то же.

§ 4. ВЫЧИСЛЕНИЕ КОЛИЧЕСТВА ИНФОРМАЦИИ ПРИ УЧЕТЕ ЗАВИСИМОСТИ МЕЖДУ СИМВОЛАМИ

Следующим шагом на пути получения общих соотношений для вычисления количества информации является снятие условия статистической независимости между символами. Это диктуется необходимостью оценивать весьма частые практические ситуации, в которых зависимостью между символами нельзя пренебречь; ярким примером является человеческая речь.

Так как по-прежнему пока остается в силе условие отсутствия ошибок, естественно воспользоваться тем фактом, что в этом случае количество информации численно равно энтропии множества возможных сигналов. Из формулы (3.4) предыдущего параграфа легко видеть, что энтропия сигнала длительностью в n символов исчисляется увеличенной в n раз энтропией одного символа, которая в том случае была равна энтропии алфавита. Наличие статистической зависимости усложняет исчисление неопределенности отдельного символа в ряду других, однако во многих конкретных случаях вычисления могут быть доведены до конца. Рассмотрим, например, важный случай такой статистической зависимости, которая имеет место между элементами простой цепи Маркова.

Очевидно, что энтропия символа, который должен осуществиться, теперь зависит от того, какой символ осуществился только что перед ним. Пусть, например, последним символом был символ под номером i. Тогда энтропия следующего символа (при условии, что предыдущий известен) равна

$$H_i = -\sum_{j=1}^{m} p(j|i) \log p(j|i), \qquad (4.1)$$

где $p(j|i)$ — вероятность того, что после символа i осуществится символ j. Нас, однако, интересует безусловная энтропия H символа в цепи Маркова, т. е. средняя величина $\bar{H}_i$. По определению среднего,

$$H = \sum_{i=1}^{m} p(i) H_i = -\sum_{i=1}^{m} p(i) \sum_{j=1}^{m} p(j|i) \log p(j|i) =$$

$$= -\sum_{i,j=1}^{m} p(i,j) \log p(j|i). \qquad (4.2)$$

Эти соотношения и дают нам искомый результат.

§ 5. КОЛИЧЕСТВО ИНФОРМАЦИИ КАК МЕРА СНЯТОЙ НЕОПРЕДЕЛЕННОСТИ

Обсудим связь между количеством информации и энтропией, которая, в частности, проявилась в предыдущих параграфах. Первоначальные рассуждения при введении количественной меры информации проводились с явной тенденцией построить самостоятельную, независимую величину, поведение которой согласовывалось бы с интуитивными требованиями к мере информации. В ходе поисков такой величины было показано, что мерой количества информации может служить логарифм числа возможных сигналов; при выполнении конкретных вычислений оказалось, что эта величина численно совпадает с энтропией множества возможных сигналов.

Такое совпадение лишь на первый взгляд может показаться случайным. Фактически же это есть результат проявления свойства E энтропии случайных процессов (см. § 9 гл. IV), которое как раз и гласит, что логарифм числа реализаций в высоковероятной группе весьма близок к энтропии множества всех реализаций процесса. Действительно, в рассуждениях § 2 и 3 неявно речь шла именно о реализациях высоковероятной группы, которые фигурировали там под несколько неопределенным наименованием «возможных сигналов». В частности, предполагалась равновероятность возможных сигналов — свойство, присущее лишь реализациям высоковероятной группы (см. следствие 1, § 9 гл. V).

Таким образом, уже простейший анализ причин численного совпадения I и H в рассмотренных случаях дает важные результаты. Во-первых, совершенно точный смысл приобретает понятие возможного сигнала, которое отождествляется

с понятием реализации из высоковероятной группы. Невозможных сигналов нет, но подавляющее большинство сигналов обычно обладает столь малой вероятностью, что практически такие сигналы не встречаются. Во-вторых, как только было показано, что мерой количества информации может служить логарифм числа различных (и равновероятных) сигналов, так сразу же, исходя из свойства E, можно было утверждать, что в этом случае количество информации I по величине будет совпадать с энтропией сигнала H.

Возникают естественные вопросы: не является ли численное совпадение величин I и H в рассмотренных случаях проявлением глубокой общности самой сущности этих величин? Если это так, то в чем заключается эта общность, и в чем различие этих величин?

Чтобы ответить на эти вопросы, обратимся к рассмотрению самого процесса получения информации. До того как получатель наблюдал сигнал, несущий информацию об интересующем его объекте, было неизвестно, в каком из состояний находится объект, но считается известным распределение вероятностей $p\ (x_к)$ по возможным состояниям $\{x_к\}$. Неопределенность ситуации до приема сигнала характеризуется, следовательно, энтропией $H(x) = -\sum_к p\ (x_к) \log p\ (x_к)$. Далее получатель наблюдает объект-сигнал. Поскольку везде до сих пор предполагалось, что состояния интересующего нас объекта и объекта-сигнала находятся в однозначном соответствии (шумы отсутствуют), наблюдение сигнала дает совершенно точный ответ, в каком именно состоянии находится объект-оригинал. Следовательно, в этом случае после приема сигнала неопределенность относительно состояния объекта равна нулю.

Таким образом, в результате приема сигнала, с одной стороны, произошло уменьшение неопределенности с $H\ (x)$ до нуля, а с другой стороны — получено количество информации I, численно равное H. Это подводит нас к мысли о том, что количество информации равно разности энтропий объекта до и после приема сигнала. В этом нас убеждает и ряд других соображений. Действительно, если мы снимем допущение об отсутствии шумов, то даже при точном определении состояния объекта-сигнала мы не можем сделать однозначное заключение относительно состояния объекта-оригинала. При этом исчезает численное совпадение I и H; количество информации будет, очевидно, меньше, чем в случае отсутствия шумов, так как прием не уменьшает энтропии до нуля. Если состояния объекта-оригинала и объекта-сигнала статистически не связаны, то наблюдение сигнала вообще не изменяет неопределенности, и приток информации отсутствует, так как один объект не отражает другого.

На основании изложенного выше можно сформулировать третье определение количества информации: количество информации есть мера снятой неопределенности; численное значение количества информации о некотором объекте равно разности априорной и апостериорной энтропий этого объекта.

Третье определение количества информации окончательно разрешает вопрос об общности и различии понятий количества информации и энтропии. Прежде всего, в свете этого определения понятие энтропии является первичным, исходным, а понятие количества информации — вторичным, производным понятием: энтропия есть мера неопределенности, а количество информации — мера снятой неопределенности, мера изменения неопределенности. Несмотря на простоту связи этих величин, было бы неправильным считать, что введение специального понятия для разности энтропий является искусственным, надуманным и не вызванным необходимостью мероприятием. Количество информаци даже как простая разность априорной и апостериорной энтропий имеет вполне самостоятельное и важное значение. Яркой аналогией является соотношение между понятиями электрического потенциала и напряжения: хотя напряжение есть простая разность потенциалов, в большинстве технических и физических расчетов и экспериментов именно напряжение, а не сами по себе потенциалы, является существенной величиной.

§ 6. ВЫЧИСЛЕНИЕ КОЛИЧЕСТВА ИНФОРМАЦИИ ПРИ НАЛИЧИИ ШУМОВ

Определение количества информации как меры снятой неопределенности значительно облегчает дальнейшие обобщения. Вслед за условиями равновероятности и статистической независимости, снятыми в § 3 и 4, это определение позволяет легко снять условие отсутствия шумов и получить обобщение соответствующей формулы для вычисления количества информации.

Согласно третьему определению, количество информации I определяется через априорную энтропию H и апостериорную энтропию H_0 соотношением

$$I = H - H_0. \tag{6.1}$$

Задача, таким образом, состоит лишь в том, чтобы связать апостериорную энтропию с шумами и привести формулу (6.1) к удобному виду.

Рассмотрим достаточно простой случай, когда передача сигнала в шумах происходит при следующих условиях: 1) полезный сигнал является последовательностью статистически независимых символов, 2) искажение (случайная подмена) очередного символа является событием, статистически независящим от того, как исказился предыдущий символ, 3) шум и сигнал являются статистически независимыми случайными процессами. В этом случае достаточно рассмотреть энтропию одного символа.

Прием очередного символа теперь не полностью снимает неопределенность, имевшую место до приема, так как данному принятому символу соответствует (с разными вероятностями) несколько возможных отправляемых сигналов. Эта ситуация иллюстрируется рисунком (рис. 19), где $\{x\}$ — множество состояний объекта-оригинала (отправляемые символы), $\{y\}$ — множество состояний

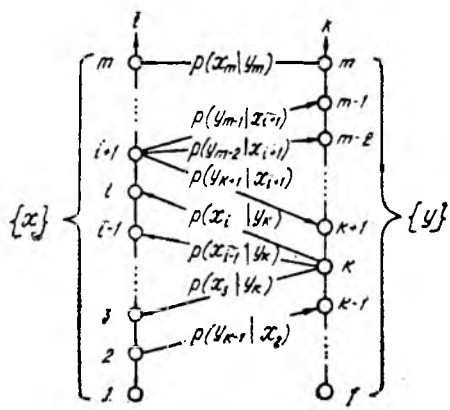

Рис. 19.

объекта-сигнала (принимаемые символы); $p(x_i|y_k)$ — вероятность того, что был отправлен символ x_i, если принят символ y_k; $p(y_k|x_i)$ — вероятность того, что будет принят символ y_k, если отправлен символ x_i.

Информационное описание рассматриваемой ситуации производится с помощью следующих энтропий:

$H(x)$ — энтропия множества отправляемых символов;

$H(y)$ — энтропия множества принимаемых символов;

$H(x,y)$ — энтропия множества всевозможных пар (x_i, y_k);

$H(x|y_k)$ — энтропия множества отправляемых символов, оставшаяся после приема символа y_k;

$H(y|x_i)$ — энтропия множества принимаемых символов при условии, что известен отправленный символ x_i;

$H(x|y)$ и $H(y|x)$ — математические ожидания величин $H(x|y_k)$ и $H(y|x_i)$ соответственно.

Количество информации, получаемое при приеме символа y_k, в соответствии с определением, запишется как

$$I_K = H(x) - H(x|y_k). \tag{6.2}$$

В общем случае I_k будет случайной величиной. Целесообразно поэтому для описания ситуации в целом вычислить среднее количество информации в объекте $\{y\}$ относительно объекта $\{x\}$, определив его как

$$I = MI_k = H(x) - MH(x|y_k) = H(x) - H(x|y). \tag{6.3}$$

Интересно, что $I(x, y)$ может быть аналогично (6.3) выражено через энтропии множества $\{y\}$. Действительно, из свойства 4 энтропии (см. § 2 гл. V) следует:

$$H(x) = H(x, y) - H(y|x) = H(y) + H(x|y) - H(y|x).$$

Подставив это выражение в (6.3), получим, что

$$I(x, y) = H(y) - H(y|x). \tag{6.4}$$

Таким образом, средняя неопределенность того, какой символ будет получен, снимаемая при посылке конкретного символа, равна средней неопределенности того, какой символ был отправлен, снимаемой при приеме символа. Смысл такой симметричности будет обсужден в следующем параграфе.

Получим теперь выражение $I(x, y)$ через соответствующие вероятности.

$$I(x, y) = H(x) - H(x|y) = -\sum_{i}^{m} p(x_i) \log p(x_i) +$$

$$+ \sum_{k}^{m} p(y_k) \sum_{i}^{m} p(x_i|y_k) \log p(x_i|y_k) =$$

$$= -\sum_{k}^{m}\sum_{i}^{m} p(x_i, y_k) \log p(x_i) +$$

$$+ \sum_{k}^{m}\sum_{i}^{m} p(x_i, y_k) \log p(x_i|y_k) =$$

$$= \sum_{k}^{m}\sum_{i}^{m} p(x_i, y_k) \log \frac{p(x_i|y_k)}{p(x_i)}. \tag{6.5}$$

Аналогично, из (6.4) можно получить:

$$I(x, y) = \sum_{k}^{m}\sum_{i}^{m} p(x_i, y_k) \log \frac{p(y_k|x_i)}{p(y_k)}. \qquad (6.6)$$

§ 7. КОЛИЧЕСТВО ИНФОРМАЦИИ КАК МЕРА СООТВЕТСТВИЯ ДВУХ СЛУЧАЙНЫХ ОБЪЕКТОВ

Отмеченная в предыдущем параграфе возможность выражения среднего количества информации через энтропии как отражающего объекта $\{y\}$, так и отражаемого объекта $\{x\}$, заслуживает весьма пристального внимания. Такая симметричность является новым свойством количества информации, установить и обсудить которое мы еще не имели случая.

Прежде всего заметим, что выражениям (6.5) и (6.6) легко может быть придана совершенно симметричная форма. Действительно, умножив и разделив логарифмируемое выражение в (6.5) на $p(y_k)$, а в (6.6) на $p(x_i)$, сразу получим:

$$I(x, y) = \sum_{i=1}^{m}\sum_{k\ 1}^{m} p(x_i, y_k) \log \frac{p(x_i, y_k)}{p(x_i)\cdot p(y_k)}. \qquad (7.1)$$

Здесь уже никак не сказывается тот факт, что при выводе считалось, что исходной величиной является $\{x\}$, а производной, сигналом первой, является $\{y\}$. По виду формулы (7.1) невозможно заключить, какой объект мы считаем отражаемым, а какой — отражающим. Поскольку перемена местами аргументов в (7.1) не меняет величины $I(x, y)$, формулу (7.1) можно толковать следующим образом: количество информации об объекте $\{x\}$, заключающееся в объекте $\{y\}$, равно количеству информации об объекте $\{y\}$, заключающемуся в объекте $\{x\}$.

Этот факт весьма существенен, так как позволяет осветить понятие количества информации еще с одной стороны. То, что $I(x, y)$ одинаковым образом зависит как от $\{x\}$, так и от $\{y\}$, говорит, что количество информации является не характеристикой одного из этих объектов, а характеристикой их связи, характеристикой соответствия состояний объектов $\{x\}$ и $\{y\}$. Подчеркивая это, можно сформулировать четвертое определение количества информации: среднее количество информации есть мера соответствия двух случайных объектов; численное значение ее определяется соотношением (7.1).

Это определение подводит нас к тому, чтобы окончательно связать понятие количества информации с понятием информации. Информация есть отражение одного объекта другим, проявляющееся в наличии соответствия их состояний. Один объект может быть отражаем несколькими другими; причем часто одними лучше, чем остальными. Среднее количество информации и есть численная характеристика степени отражения, степени соответствия.

Поскольку четвертое определение формально может рассматриваться как независимое, не связанное, в частности, с понятием энтропии, оно может быть принято в качестве исходного при аксиоматическом построении всей теории информации. Можно показать (см., например, [63, 20]), что построенная таким образом теория будет также внутренне непротиворечивой. При этом понятие энтропии может быть представлено как производное от понятия количества информации, исходя из того, что при отсутствии шума количество информации равно энтропии (см. § 3). Едва ли можно найти возражения логического порядка против такого принципа построения теории, особенно если излагать ее аксиоматически, как сугубо математическую дисциплину. Если же обращать внимание в первую очередь на прикладное значение теории информации и рассматривать ее как физическую теорию, то нельзя не заметить, что принятие количества информации за первичное понятие несколько обедняет физическое содержание теории, в частности, содержание понятия энтропии: энтропия выступает при этом не как мера неопределенности, а как количество информации, заключенное в объекте о самом себе (либо в объекте, однозначно связанном с первым).

Приводя соображения против того, чтобы принимать четвертое определение за исходное при построении всей теории, мы не должны, тем не менее, упускать из виду очень важное значение этого определения, поскольку оно подчеркивает новый аспект понятия количества информации, делает это понятие более ярким и содержательным.

Отметим еще одну особенность понятия количества информации, связанную с обсуждаемым определением. В функционале (7.1), характеризующем количество информации, как отражаемый, так и отражающий объекты выступают совершенно равноправно. С одной стороны, это подчеркивает обоюдность отражения: не только один объект отражает другой, но оба отражают друг друга, содержат информацию друг о друге. Это представляется естественным, поскольку отражение есть результат взаимодействия, т. е. взаимного, обоюдного изменения состояний. С другой стороны, фактически всегда одно явление (или объект) выступает как

причина, другое — как следствие. Негатив и отпечатанные с него фотокарточки, гербовая печать и ее оттиски на бумагах; картина древнего мастера и копии, написанные с нее; сигнал, переданный телестудией, и изображение на экране телевизора — содержат одинаковую информацию друг о друге; но во всех этих случаях один объект явно выступает как оригинал, причина, а другой — как его отражение, копия, следствие. Этот факт никак не учитывается определениями количества информации, рассмотренными выше.

Таким образом, понятие среднего количества информации опирается на отвлечение от причинно-следственных связей, на абстрагирование от процесса установления соответствия состояний объектов и описывает лишь конечный результат этого процесса — наличие соответствия. Такой подход позволяет без специальных оговорок применять теорию информации в случаях, самых различных по физическим, химическим и другим условиям связи между объектами; это, конечно, является одним из достоинств теории. Тем не менее, существование причинно-следственных отношений, несимметричность времени, реальная направленность потоков информации являются общими свойствами природы; соответствующее обобщение понятий будет, по-видимому, одной из важных проблем в дальнейшем развитии теории информации.

§ 8. КОЛИЧЕСТВО ИНФОРМАЦИИ В НЕПРЕРЫВНЫХ ОБЪЕКТАХ

Из пяти ограничений, в рамках которых было дано первое определение количества информации, в ходе обобщения осталось не снятым только одно; до сих пор речь шла об объектах с дискретным множеством состояний. Займемся рассмотрением вопросов, возникающих в связи с необходимостью обобщения понятия количества информации на объекты с континуумом возможных состояний. Переход от дискретного случая к непрерывному естественно вести через понятие дифференциальной энтропии, опираясь на третье определение количества информации.

Пусть оба отражающих друг от друга объекта $\{x\}$ и $\{y\}$ непрерывны. Снятая неопределенность измеряется теперь разностью априорной и апостериорной дифференциальных энтропий одного из объектов:

$$I(x, y) = H_\varepsilon(x) - MH_\varepsilon(x|y) = H_\varepsilon(y) - MH_\varepsilon(y|x) =$$

$$= -\int p(x) \log p(x)\, dx - \log \varepsilon +$$

$$+ \int p(y) \left[\int p(x|y) \log p(x|y)\, dx + \log \varepsilon \right] dy =$$

$$= \iint p(x,y) \log \frac{p(x,y)}{p(x)\cdot p(y)}\, dx\, dy. \qquad (8.1)$$

Полученная формула является обобщением формулы (7.1) для количества информации в дискретном случае.

Пусть теперь один из объектов, например $\{x\}$, дискретен, а другой — непрерывен. И в этом случае формула для вычисления количества информации может быть получена на основе третьего определения — либо как разность априорной и апостериорной энтропии дискретного объекта $\{x\}$, либо как разность априорной и апостериорной дифференциальных энтропий непрерывного объекта $\{y\}$:

$$I(x, y) = H(x) - MH(x|y) =$$

$$= H_\varepsilon(y) - MH_\varepsilon(y|x_i) =$$

$$= \sum_i \int p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)\cdot p(y)}\, dy. \qquad (8.2)$$

Отметим важный факт, вытекающий из формул (8.1) и (8.2): хотя $I(x, y)$ определяется в непрерывном и полунепрерывном случаях через соответствующие дифференциальные энтропии, само количество информации от величины стандарта сравнения ε не зависит. Это может служить еще одним аргументом в пользу толкования количества информации как меры соответствия двух случайных объектов: степень их соответствия, очевидно, не зависит от того, с каким стандартом сравнивается каждый из них в отдельности.

Следует также подчеркнуть, что количество информации, содержащееся в одной непрерывной величине относительно другой (не обязательно непрерывной), не зависит от того, в какой системе координат рассматривается непрерывная случайная величина. Действительно, как бы не преобразовывалась система координат, якобианы преобразования плотностей под знаком логарифма будут сокращаться, что и доказывает сделанное утверждение. Это также хорошо согласуется с толкованием количества информации как характеристики связи двух величин, которая, естественно, не зависит от того, в какой системе рассматривается одна из величин.

Итак, формула (8.1) и ее частные случаи (7.1) и (8.2) позволяют вычислять количество информации в любом случае, если соответствующие интегралы и суммы существуют. Мыслимы, однако, случайные объекты, для которых не суще-

ствует функция плотности вероятностей; это требует дальнейшего математического обобщения (см., например, [20]), которое мы здесь не будем рассматривать, поскольку подавляющее большинство прикладных задач решается в рамках обычных предположений.

§ 9. ОСНОВНЫЕ СВОЙСТВА КОЛИЧЕСТВА ИНФОРМАЦИИ

Рассмотрим некоторые основные свойства величины $I(x, y)$, являющейся функционалом распределений вероятностей случайных величин x и y. Заметим, что все свойства сохраняются и в том случае, когда x и y многомерны.

I. Количество информации в случайном объекте x относительно объекта y равно количеству информации в y относительно x:

$$I(x, y) = I(y, x). \tag{9.1}$$

Доказательство следует из симметричности функционала $I(x, y)$ относительно переменных x и y.

II. Количество информации неотрицательно:

$$I(x, y) \geqslant 0. \tag{9.2}$$

Приведем доказательство [7; 47, 65] этого свойства для случая дискретных x и y, $x = (x_1, x_2, \ldots, x_n)$, $y = (y_1, y_2, \ldots, y_m)$. Учитывая, что

$$\left.\begin{aligned} \sum_{i=1}^{n} p(x_i, y_k) &= p(y_k), \\ \sum_{k=1}^{m} p(x_i, y_k) &= p(x_i), \end{aligned}\right\} \tag{9.3}$$

можно записать:

$$I(x, y) = \sum_{i=1}^{n}\sum_{k=1}^{m} p(x_i, y_k) \log \frac{p(x_i, y_k)}{p(x_i)\cdot p(y_k)} =$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{m} p(x_i, y_k) \log \frac{p(x_i, y_k)}{p(x_i)} - \sum_{k=1}^{m} p(y_k) \log p(y_k) =$$

$$= \sum_{k=1}^{m}\left\{\sum_{i=1}^{n} p(x_i) \Phi\left[\frac{p(x_i, y_k)}{p(x_i)}\right] - \Phi[p(y_k)]\right\}, \tag{9.4}$$

где $\Phi(t) = t \log t$. Так как $\Phi(t)$ при $t \geqslant 0$ является выпуклой функцией, то

$$\sum_{i=1}^{n} \lambda_i \Phi(t_i) \geqslant \Phi\left(\sum_{i=1}^{n} \lambda_i t_i\right), \text{ при } \sum_{i=1}^{n} \lambda_i = 1,$$

причем равенство имеет место тогда и только тогда, когда все t_i совпадают между собой (см. [69], стр. 96). Полагая $\lambda_i = p(x_i)$, $t_i = p(x_i, y_k)/p(x_i)$, в силу (9.3) имеем:

$$\sum_{i=1}^{n} p(x) \Phi\left[\frac{p(x_i, y_k)}{p(x_i)}\right] \geqslant \Phi(p(y_k)). \tag{9.5}$$

Для того, чтобы это неравенство обратилось в равенство, необходимо и достаточно, чтобы все величины

$$t_i = \frac{p(x_i, y_k)}{p(x_i)} = p(y_k/x_i), \quad i = 1, 2, \ldots, n,$$

совпали между собой. Это условие равносильно условию статистической независимости x и y. Таким образом (9.5) в совокупности с (9.4) доказывают рассматриваемое свойство.

III. Количество информации в объекте x о самом себе равно его энтропии:

$$I(x, x) = H(x).$$

Если x — непрерывный объект, то $I(x, x) = H_\varepsilon(x)$ при $\varepsilon = 1$. Свойство III доказывается простой проверкой.

IV. Количество информации в одном объекте относительно другого не больше энтропии любого из этих объектов:

$$I(x, y) \leqslant H(x), \quad I(x, y) \leqslant H(y).$$

Действительно, из (9.4) следует, что

$$I(x, y) - I(y, y) = I(x, y) - H(y) =$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{m} p(x_i) \Phi\left[\frac{p(x_i, y_k)}{p(x_i)}\right],$$

а так как $\Phi(0)=0$ и $\Phi(t)<0$ при $0<t<1$, то отсюда и вытекает свойство IV.

V. При любом обратимом преобразовании случайной величины количество информации, содержащееся в ней относительно другой величины, не изменяется. Пусть $z=z(x)$ — произвольная функция случайной величины x, но такая, что существует однозначная обратная функция $x=\varphi(z)$. Тогда

$$I(z, y)=I(x, y). \tag{9.6}$$

Это, в частности, означает, что никакое усиление сигнала не может увеличить количества информации в нем. С другой стороны, в сложных системах усиление применяется именно для улучшения информационных качеств системы в целом. Это связано с рядом физических причин, которые будут рассмотрены в следующих главах, и которые, конечно, не противоречат соотношению (9.6). Свойство V доказывается непосредственной проверкой соотношения (9.6).

VI. Необратимые преобразования случайной величины x в общем случае разрушают информацию, содержащуюся в ней. Пусть $f(x)$ — некоторое необратимое преобразование величины x. Тогда

$$I(f(x), y)\leqslant I(x, y). \tag{9.7}$$

Несмотря на ясность и «почти очевидность» этого свойства, его строгое доказательство достаточно сложно (см., например, [67]) и здесь приводиться не будет. К числу необратимых преобразований многомерной случайной величины (случайного вектора) относятся вырожденные линейные преобразования, например, отбрасывание одной из компонент вектора. Таким образом, частным случаем формулы (9.7) будет соотношение

$$I((x_1, x_2,\ldots,x_n), y)\geqslant I((x_1, x_2,\ldots,x_{n-1}), y). \tag{9.8}$$

По поводу этого соотношения полезно заметить [20], что из условий $I(x_1, y)=0$ и $I(x_2, y)=0$ еще не следует, что $I((x_1, x_2), y)=0$. Укажем также, что если $(x_1, x_2,\ldots,x_n, y)$ образуют марковскую цепь, то

$$I((x_1, x_2,\ldots,x_n), y)=I(x_n, y), \tag{9.9}$$

что легко доказывается с помощью условия $p(y|x_1,x_2,\ldots,x_n)=p(y|x_n)$.

VII. Свойства V и VI можно подытожить следующим утверждением: никакое преобразование случай-

ной величины не может увеличить содержание в ней информации относительно другой, связанной с ней величины. В частности, это означает, что какие бы сложные преобразования принятого сигнала ни осуществлялись бы приемным устройством, нельзя получить большего количества информации об отправленном сигнале, чем то, которое содержится во входном сигнале приемника. Самое лучшее, что можно сделать — это преобразовывать сигнал так, чтобы полезная информация не терялась. Этот вопрос требует особого рассмотрения; в частности, можно показать [64 и др.], что иногда могут быть найдены такие необратимые преобразования, которые не разрушают полезной информации.

VIII. Для случая трех величин x, y, z, связанных между собой, может быть доказано [67] следующее тождество:

$$I((x, y), z) + I(x, y) = I(x, (y, z)) + I(y, z), \qquad (9.10)$$

которое можно проверить непосредственно.

IX. Пусть x, y, z — три случайные величины. Тогда выполняется важное соотношение, указанное А. Н. Колмогоровым (полное доказательство этого соотношения имеется в работе Р. Л. Добрушина [67]):

$$I((x, y), z) = I(x, z) + MI(y, z|x_i). \qquad (9.11)$$

Здесь через $I(y, z|x_i)$ обозначено количество информации в y относительно z, вычисленное при условии, что x приняло конкретное значение x_i:

$$I(y, z|x_i) = \iint p(y, z|x_i) \log \frac{p(y, z|x_i)}{p(y|x_i) \cdot p(z|x_i)} \, dy \, dz.$$

Формула (9.11) также может быть проверена непосредственно.

X. Если имеются два случайных вектора $x = (x_1, x_2, \ldots, x_n)$ и $y = (y_1, y_2, \ldots, y_n)$ таких, что пары (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n) независимы между собой, то

$$I(x, y) = I(x_1, y_1) + I(x_2, y_2) + \ldots + I(x_n, y_n).$$

Для случая, когда соответствующие плотности вероятностей существуют, это свойство доказывается непосредственно; для общего случая доказательство разработано Р. Л. Добрушиным [67].

§ 10. ЕДИНИЦЫ ИЗМЕРЕНИЯ ЭНТРОПИИ И КОЛИЧЕСТВА ИНФОРМАЦИИ

Для сравнения любых количественных величин необходимо установить единицы их измерения. Это, конечно, относится и к энтропии и количеству информации. В силу специфической связи количества информации и энтропии (количество информации есть изменение энтропии) единицы измерения для них одинаковы, поэтому достаточно ввести единицы измерения для одной из них, например, для H. Укажем также, что из определений I и H следует их безразмерность.

Рассмотрим сначала дискретный случай. За единицу измерения энтропии принимается неопределенность случайного объекта, специально заданного так, чтобы

$$H(x) = \sum_{i=1}^{n} p(x_i) \log p(x_i) \quad 1 \qquad (10.1)$$

Легко видеть, что для однозначного определения единицы измерения энтропии необходимо конкретизировать: 1) число возможных состояний случайного объекта, 2) основание логарифма в формуле (10.1); тогда определится распределение $p(x_i)$, неопределенность которого равна единице (либо определить n и распределение $p(x_i)$, тогда из (10.1) найдется основание логарифма; но первый путь более естественен). Для определенности возьмем наименьшее число возможных состояний, при котором объект еще остается «случайным», $n=2$; далее, выберем в качестве основания логарифма число 2. Тогда из

$$p(x_1) \log_2 p(x_1) \quad p(x_2) \log_2 p(x_2) = 1 \qquad (10.2)$$

следует, что $p(x_1) = p(x_2) = 0{,}5$. Следовательно, единицей неопределенности служит энтропия случайного объекта с двумя равновероятными возможными состояниями. Эта единица называется двоичной единицей энтропии (или количества информации), либо просто «битой» (от английского bit, происходящего от binary digit).

Можно ввести и другие единицы измерения неопределенности. Например, если в качестве основания логарифма взять 10, получим десятичную единицу неопределенности (иногда ее называют «дитой», от dit); если использовать натуральные логарифмы, будем иметь натуральные единицы неопределенности («ниты»); и т. д. Неопределенностью в одну десятичную единицу обладает, например, объект с десятью равновероятными состояниями.

Аналогично определяются единицы измерения в непрерывном случае. Например, считается, что в среднем в $\{x\}$ содержится одна нита информации об $\{y\}$, если

$$\int\int p(x, y) \ln \frac{p(x, y)}{p(x)p(y)} dx\, dy = 1. \tag{10.3}$$

Однако если единица количества информации и в непрерывном случае имеет абсолютный характер, то единица дифференциальной энтропии относительна: то, что

$$H_\varepsilon(x) = -\int p(x) \log_2 p(x)\, dx = 1, \tag{10.4}$$

означает, что неопределенность распределения $p(x)$ на одну биту больше неопределенности стандартного распределения с $\varepsilon=1$.

§ 11. КОЛИЧЕСТВО ИНФОРМАЦИИ В ИНДИВИДУАЛЬНОМ СОБЫТИИ

До сих пор речь шла лишь о среднем количестве информации в одном объекте относительно другого, о характеристике соответствия объектов в целом. Естественно, что эта характеристика не зависела в конечном счете от того, какие именно состояния примут случайные объекты в конкретном случае:

$$I(x, y) = \sum_{i, k} p(x_i, y_k) \log \frac{p(y_k|x_i)}{p(y_k)}. \tag{11.1}$$

Это соотношение можно толковать так: в среднем на любую реализацию пары состояний (x_i, y_κ) случайных объектов x и y приходится количество информации $I(x, y)$.

Однако в ряде практически важных случаев оказывается необходимым оценить, какое количество информации содержится в конкретной паре (x_l, y_κ). То, что оно действительно зависит от индивидуальных событий, можно проиллюстрировать следующими примерами. Вам необходимо задать собеседнику вопрос, ответ на который будет «почти наверняка» отрицательный, например, является ли он левшой. Его отрицательный ответ почти ничего не добавит к тому, что вы знаете; если же ответом будет «да», вы получите значительно большее количество информации. Другой пример: вы наблюдаете за экраном радиолокатора, установленного на полигоне вдалеке от обычных воздушных трасс. Если на экране видна лишь обычная шумовая «травка», это не вызывает у вас инте-

реса: вы и так знаете, что самолеты здесь появляются редко. Совсем другое отношение вызывает событие, заключающееся в появлении импульса, хотя присутствие и отсутствие цели могут рассматриваться просто как возможные состояния одного случайного объекта.

Допуская существование количественной меры информации $i(x_l, y_\kappa)$ для конкретной пары (x_i, y_κ) состояний связанных случайных объектов x и y, мы, очевидно, должны потребовать, чтобы между индивидуальным и средним количествами информации существовало соотношение

$$I(x, y) = Mi(x_l, y_k) = \sum_{i,k} p(x_i, y_k) \cdot i(x_l, y_k). \qquad (11.2)$$

Хотя, конечно, равенство средних значений не является доказательством совпадения усредняемых функций, сравнение (11.1) и (11.2) наводит на мысль о том, что мерой количества информации в индивидуальном случае может служить величина

$$i(x_l, y_k) = \log \frac{p(x_l|y_k)}{p(x_i)} = \log \frac{p(y_k|x_l)}{p(y_k)}. \qquad (11.3)$$

Для того, чтобы окончательно принять это определение индивидуального количества информации, необходимо, с одной стороны, показать, что так определяемая величина удовлетворяет всем требованиям, которые разумно предъявить к мере количества информации в конкретном событии, а с другой стороны — доказать, что совокупности этих требований удовлетворяет только функция (11.3).

Рассмотрим свойства функции $i(x_l, y_k)$.

1. Пусть, наблюдая объект y, мы хотим получить информацию о состоянии объекта x. Нас интересует, какое количество информации о том, что x находится в состоянии x_l получено, если в результате наблюдения установлено, что y находится в состоянии y_k. Априорная вероятность состояния x_i определяется числом $p(x_l)$; апостериорная — числом $p(x_l|y_k)$. Таким образом, весь результат наблюдения y_k сводится к изменению вероятности состояния x_l с $p(x_i)$ до $p(x_i|y_k)$. Разумно потребовать, чтобы количество информации в y_k о x_i однозначно определялось этими вероятностями, т. е.

$$i(x_l, y_k) = F(p(x_l), p(x_i|y_k)). \qquad (11.4)$$

Функция $i(x_i, y_k)$, определяемая в соответствии с (11.3), удовлетворяет этому требованию.

2. Функция $i(x_l, y_k)$ симметрична относительно своих аргументов:

$$i(x_i, y_k) = \log \frac{p(x_i|y_k)}{p(x_i)} = \log \frac{p(x_i, y_k)}{p(x_i)\,p(y_k)} =$$

$$= \log \frac{p(y_k|x_i)}{p(y_k)} = i(y_k, x_i). \qquad (11.5)$$

Это свойство также представляется естественным, поскольку количество информации должно характеризовать взаимное соответствие состояний x_i и y_k.

3. Характерно, что индивидуальное количество информации $i(x_i, y_k)$ может быть не только положительным, но и отрицательным и нулем. Это хорошо согласуется с интуитивными представлениями: если вероятность x_i после наблюдения y_k возросла — количество информации об x_i увеличилось, если же эта вероятность уменьшилась, то и количество информации уменьшается; если наблюдение не изменило вероятности x_i, оно не принесло дополнительной информации. Напомним, что несмотря на это, среднее количество информации $I(x, y)$ всегда неотрицательно. Более того, можно показать, что и среднее количество информации, приносимое конкретным состоянием y_k, — тоже всегда положительно:

$$I(x, y_k) = \sum_i p(x_i|y_k) \log \frac{p(x_i|y_k)}{p(x_i)} \geqslant 0. \qquad (11.6)$$

Это тоже воспринимается как естественное отражение того, что если x и y статистически связаны, то наблюдение конкретного состояния y_k должно в среднем приносить положительное количество информации об x. $I(x, y_k) = 0$ только при равенстве распределений $p(x_i)$ и $p(x_i|y_k)$, в этом случае наблюдение y_k не изменяет состояние знаний наблюдателя о случайном объекте x.

4. При фиксированной априорной вероятности $p(x_i)$ индивидуальное количество информации $i(x_i, y_k)$ имеет максимальное значение при $p(x_i|y_k) = 1$, т. е. однозначной связи между x_i и y_k. Это максимальное значение равно

$$i(x_i, x_i) = -\log p(x_i). \qquad (11.7)$$

Таким образом, максимальное количество информации, которое можно получить о состоянии x_i с вероятностью $p(x_i)$ равно $-\log p(x_i)$ (иногда эту величину называют собственной информацией события x_i). В общем случае

$$i(x_i, y_k) \leqslant \begin{cases} i(x_i, x_i), \\ i(y_k, y_k). \end{cases} \qquad (11.8)$$

Если под x и y понимать многомерные случайные объекты (т. е., по существу, совокупности одномерных объектов) то, кроме свойств 1—4, можно установить еще ряд полезных свойств.

5. Пусть, например, рассматривается количество информации в паре (y_k, z_e) относительно x_i. Естественно потребовать, чтобы это количество не зависело от того, рассматриваются объекты y и z в совокупности или в определенной последовательности. Обозначим количество информации z_e относительно x_i при фиксированном y_k как $i(x_i, z_e | y_k)$. Тогда это требование можно записать следующим образом:

$$i(x_i, (y_k, z_e)) = i(x_i, y_k) + i(x_i, z_e | y_k) = \tag{11.9}$$

$$= i(x_i, z_e) + i(x_i, y_k | z_e). \tag{11.9a}$$

Легко убедиться, что функция (11.3) удовлетворяет и этому требованию, если под $i(x_i, z_e | y_k)$ понимать

$$i(x_i, z_e | y_k) = \log \frac{p(x_i | z_e, y_k)}{p(x_i | y_k)}. \tag{11.10}$$

Свойство, выражаемое формулой (11.9), можно истолковать и в несколько ином свете [64]: если последовательно наблюдаются два объекта y и z, отражающих один и тот же объект x, и наблюдатель рассматривает апостериорную вероятность состояния x_i после наблюдения первого объекта как априорную вероятность перед наблюдением второго, то полное увеличение количества информации относительно x равно сумме количеств информации, полученных при каждом наблюдении.

6. Рассмотрим теперь два n-мерных случайных объекта, $(x, \xi, \ldots)$ и $(y, \eta, \ldots)$, таких, что пары (x, y), $(\xi, \eta), \ldots$ статистически не зависят друг от друга. Вследствие такой независимости естественным является требование аддитивности:

$$i((x_i, \xi_e, \ldots), (y_k, \eta_m, \ldots)) = i(x_i, y_k) + i(\xi_e, \eta_m) + \ldots. \tag{11.11}$$

Легко показать, что функция (11.3) удовлетворяет условию (11.11).

Таким образом, функция (11.3), во-первых, не противоречит требованию равенства ее среднего значения среднему количеству информации; во-вторых, хорошо согласуется со всеми интуитивно предъявленными требованиями к мере количества информации в одном индивидуальном событии относительно другого индивидуального события. Чтобы логически завершить введение функции (11.3) в качестве такой меры, остается доказать, что она является единственной совместимой с рассмотренными требованиями 1—6 функцией.

Можно показать, что не все свойства 1—6 являются независимыми; часть из них оказываются следствиями других. Во всяком случае, для доказательства единственности (11.3) достаточно рассмотреть лишь свойства, выражаемые формулами (11.9) и (11.11). Приведем это доказательство [64].

Найдем вид функции $F\ (a,b)$,

$$i(x_i, y_k) = F(p(x_i),\ p(x_i|y_k)) = F(a, b),$$

удовлетворяющей условиям (11.9) и (11.11). В новых обозначениях тождество (11.9) может быть переписано как

$$F(\alpha, \beta) + F(\beta, \gamma) = F(\alpha, \gamma), \tag{11.12}$$

где α, β и γ — конкретные значения переменных a и b. Предполагая дифференцируемость $F(a, b)$ по a, дифференцируя (11.12) по α при фиксированном β, получаем:

$$\left(\frac{\partial F(a, b)}{\partial a}\right)_{(\alpha,\beta)} = \left(\frac{\partial F(a, b)}{\partial a}\right)_{(\alpha,\gamma)}. \tag{11.13}$$

Так как это тождество справедливо при любых β и γ, то $\partial F/\partial a$ не зависит от b. Поэтому функция $F(a, b)$ имеет вид:

$$F(a, b) = f(a) + \varphi(b). \tag{11.14}$$

Подставляя (11.14) в (11.12), получаем, что

$$f(\beta) = -\varphi(\beta),$$

следовательно,

$$F(a, b) = f(a) - f(b). \tag{11.15}$$

Ограничившись в (11.11) двумерными случайными объектами, с учетом (11.15), получаем:

$$f(\alpha\beta) - f(\gamma\delta) = [f(\alpha) - f(\gamma)] + [f(\beta) - f(\delta)]. \tag{11.16}$$

Предполагая дифференцируемость функции $f(x)$ и фиксируя γ и δ в (11.16), имеем:

$$\beta \left[\frac{df(x)}{dx}\right]_{(\alpha\beta)} = \left[\frac{df(x)}{dx}\right]_{(\alpha)}.$$

Это соотношение при $\alpha = 1$ обращается в

$$\left[\frac{df(x)}{dx}\right]_{(\beta)} = \frac{A}{\beta}, \tag{11.17}$$

где $A = f'(x = 1) = \text{const}$. Так как (11.17) выполняется при любых β, то

$$f'(x) = \frac{A}{x},$$

и, следовательно,

$$f(x) = A \log x + B, \tag{11.18}$$

где A и B произвольные константы. Теперь из (11.4), (11.15) и (11.18) следует, что

$$i(x_i, y_k) = -A \log \frac{p(x_i | y_k)}{p(x_i)}. \tag{11.19}$$

Привлекая свойство 3, требующее, чтобы при увеличении вероятности x_i после наблюдения y_κ количество информации возрастало, видим, что A должна быть отрицательной константой, а поскольку основание логарифма не фиксировано, ее можно принять равной — 1. Тем самым и завершается обоснование принятия (11.3) в качестве меры индивидуального количества информации.

В порядке обобщения можно ввести меру индивидуального количества информации для непрерывных объектов $\{x\}$ и $\{y\}$. Квантуя случайные величины x и y, применяя дискретную теорию, а затем переходя к пределу при стремлении интервалов квантования к нулю, получим, что количество информации о состоянии x объекта $\{x\}$ при наблюдении состояния y объекта $\{y\}$ будет исчисляться величиной

$$i(x, y) = \log \frac{p(x | y)}{p(x)}, \tag{11.20}$$

где $p(x)$ и $p(x | y)$ — соответствующие плотности вероятностей. И. М. Гельфанд и А. М. Яглом [65] и Р. Л. Добрушин [67] предложили для величины $i(x, y)$ удачное наименование информационной плотности и провели строгий анализ свойств этой величины.

Следует указать, что может быть построено логически непротиворечивое изложение теории информации, если принять понятие индивидуального количества информации за исходное понятие всей теории. Так, в частности, излагают теорию С. Гольдман [66], Ф. Вудворд и И. Дэвис [64], Р. Фэно [68]. Относительные достоинства и недостатки такого изложения были обсуждены нами ранее (см. § 7).

Часть III

ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ИХ ХАРАКТЕРИСТИКИ

ГЛАВА VI

ИНФОРМАЦИОННЫЕ СИСТЕМЫ

§ 1. ОБЩАЯ МОДЕЛЬ ИНФОРМАЦИОННОЙ СИСТЕМЫ

Всякая (искусственная или естественная) система взаимодействующих объектов может рассматриваться как информационная система. Любая часть совокупности взаимодействующих объектов (в частности, и один из объектов) может изучаться с целью извлечения информации о другой части этой совокупности (в частности, о другом отдельном объекте), так как взаимодействие обеспечивает соответствие состояний, т. е. отражение, содержание информации. Объекты, образующие информационную систему, могут иметь совершенно произвольную природу.

Из этого, конечно, не следует, что теория информации призвана заменить или объять другие науки, изучающие специфические взаимодействия между объектами определенного класса. Но из этого следует, что среди бесконечного множества свойств, которые присущи любой системе взаимодействующих объектов, неотъемлемым свойством является свойство объектов отражать друг друга, содержать информацию друг о друге. В некоторых явлениях информационные отношения не играют существенной роли или замаскированы — тогда наука, изучающая эти явления, может достичь определенных успехов без привлечения теории информации; в других случаях информационный подход неизбежен. Никто всерьез не примет попытку изложить, например, классическую механику в терминах теории информации, но тот факт, что всякий механический объект содержит информацию о взаимодействующем с ним объекте, прекрасно иллюстрируется примером открытия Нептуна

по возмущениям движения Урана. Стоит обратить внимание на тот факт, что релятивистская механика уже по существу информационна; например, в ней понятия одновременности и протяженности связываются со свойствами световых сигналов. Термодинамика, наука, казалось бы, весьма далекая от рассмотрения информационных аспектов энергетических превращений, не смогла обойтись без введения понятия энтропии, которую уже Больцман, следуя своей блестящей интуиции, называл «мерой недостающей информации»*).

Эти примеры призваны подчеркнуть еще раз ту мысль, что информационные свойства присущи любым взаимодействиям.

Пусть мы имеем систему объектов произвольной природы, взаимосвязанных между собой. Из множества связей конкретного объекта с другими обычно можно выделить лишь несколько наиболее существенных, опустив из рассмотрения остальные. В этом случае некоторая сложная система объектов упрощенно может быть изображена подобно рис. 20. Суще-

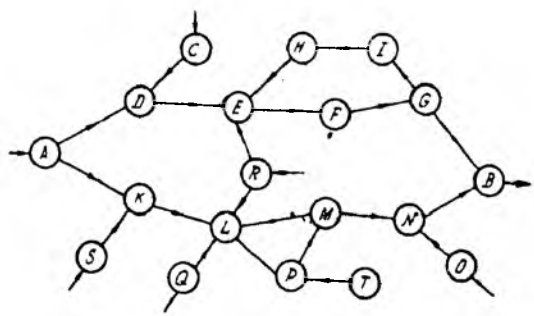

Рис. 20.

ственные связи между объектами изображены стрелками, направление которых соответствует переходу от причины к следствию. Благодаря наличию непосредственных связей, объект *B*, например, содержит информацию об объектах *I*, *G*, *N*; связи через посредство других объектов обеспечивают содержание в объекте *B* информации об объектах *A*, *R*, *S*, O и др.

Обычно получателя интересует информация о каком-нибудь одном объекте, например, *A*, и объект *B* наблюдается с целью извлечения именно этой информации. Информация об интересующем получателя объекте рассматривается как полезная, информация о других объектах предстает как ненужная, бесполезная и даже как вредная, поскольку ее наличие может

*) Позднее Л. Бриллюэн разъяснил точный смысл этих слов (см., например, [51]).

затруднить извлечение полезной информации*). Если получатель не располагает исчерпывающей информацией об остальных объектах, их влияние должно рассматриваться как «помехи», или «шум». Таким образом, всякий раз, как влияние не интересующего нас объекта нарушает однозначность соответствия состояний объектов A и B, говорят, что имеют место помехи. Соответствующие объекты (например, C, R, S, T, O, Q) считаются источниками помех.

Если объекты A и B не взаимодействуют непосредственно, то соответствие их состояний устанавливается благодаря наличию цепочек из промежуточных объектов. Таких связующих последовательностей объектов иногда может быть несколько ($Д$, E, F, G и K, L, M, N на рис. 20); иногда лишь часть последовательности мультиплетна (L, M и L, P, M). В этих случаях говорят о многоканальных, многолучевых или многопутевых системах.

Наконец, в структуре информационной системы могут существовать замкнутые последовательности объектов, несущие полезную информацию (например, E, F, G, I, H, E на рис. 20). Такие системы обычно называют системами с петлями обратной связи. Петли обратных связей могут охватывать как несколько промежуточных объектов, («внутренние» петли), так и целиком всю систему, соединяя оконечный и начальный объекты («внешние» петли).

Итак, ко всякой информационной системе могут быть различены объекты следующих четырех типов:

1. Начальный объект. Вся остальная система используется для получения информации именно об этом объекте. Начальный объект часто называют источником информации.

2. Конечный объект. Зная закон соответствия состояний начального и конечного объектов и непосредственно наблюдая последний, получатель извлекает информацию о состоянии первого.

3. Промежуточные, вспомогательные объекты. С помощью этих объектов устанавливается соответствие между начальным и конечным объектами.

4. Объекты, взаимодействие с которыми разрушает однозначность соответствия состояний начального и конечного объектов; источники помех.

Следует указать, что иногда разделение указанных типов объектов может быть осуществлено лишь условно. Простейший пример — реальный усилитель; по существу являясь объектом третьего типа, одновременно является источником тепловых шумов. Другой пример — линия связи на тропосфер-

*) Вопросы отсеивания ненужной и сохранения полезной информации будут обсуждены отдельно.

ном рассеянии. С одной стороны, наличие неоднородностей тропосферы обеспечивает само существование связи на расстоянии, с другой — хаотические движения тех же неоднородностей вызывают неконтролируемые замирания сигнала, затрудняющие связь. Однако для удобства рассмотрения даже такие системы искусственно изображаются в виде эквивалентной комбинации объектов указанных четырех типов.

Подчеркнем еще раз, что одна и та же реально существующая информационная система может быть качественно различной для двух наблюдателей, обладающих различной информацией об этой системе. Для того, чтобы извлечь информацию об объекте A, наблюдая объект B, необходимо з н а т ь закон соответствия их состояний. Если наблюдатель не знает этого закона, наблюдение объекта не может непосредственно дать ему нужной информации, вся система оказывается для него в качественно ином состоянии, нежели для наблюдателя, знающего этот закон. Чтобы убедиться в реальности такой ситуации, достаточно представить себе терпящую бедствие радиофицированную яхту, на которой после гибели радиста не оказалось людей, знакомых с азбукой Морзе. Другим примером может служить читатель, разглядывающий книгу, написанную на незнакомом языке.

Итак, качественное различие информационных систем может иметь причиной не только объективные отличия самих систем, но и (тоже объективное!) отличие в состоянии знаний наблюдателей, использующих эти системы. Естественно, для прикладной теории больший интерес представляют различные типы информационных систем, прежде всего отличающиеся по функциональному назначению. При этом, конечно, следует рассматривать эти системы с точки зрения лица, располагающего квалификацией, необходимой для нормальной эксплуатации системы. Примем это положение и в дальнейшем не будем делать специальной оговорки об этом.

Вопрос о классификации информационных систем по их функциональному назначению или использованию нельзя считать достаточно полно рассмотренным: существует множество тонких различий между такими системами. Однако несколько типов информационных систем различаются вполне четко: 1) системы связи, или системы передачи информации, 2) системы хранения информации, 3) системы обработки (преобразования) информации, 4) системы измерения, 5) системы наблюдения, или исследования. Перейдем к более детальному рассмотрению этих систем.

§ 2. СИСТЕМЫ СВЯЗИ

Системами связи называются информационные системы, основной функцией которых является перенос информации в пространстве.

Существует много разновидностей систем связи, к ним относятся почта и радиовещание, телефон и телеграф, гелиограф и сигнализация флажками, акустические системы связи и сигнальная веревка водолазов и т. д. и т. п. Особого рассмотрения заслуживают технические системы связи, в которых для переноса информации из одного пункта в другой используются динамические сигналы. Для возбуждения динамического сигнала создается специальная передающая система, а для регистрации сигнала в пункте назначения — приемная система. Совокупность объектов, связывающих передающую и приемную системы, называется линией связи. Например, в телефонной связи линия представляет собой пару проводов, в радиосвязи линией связи является пространство, в котором распространяются радиоволны.

При необходимости описания потенциальных возможностей системы связи можно условиться не учитывать конкретных особенностей отправителя и получателя*). В этом случае начальным объектом в системе связи считается входное устройство передающей системы, а конечным объектом — выходной сигнал приемной системы.

Рассмотрим некоторые разновидности технических систем связи, работающих с динамическими сигналами.

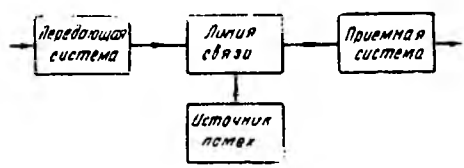

Рис. 21.

1. Одноканальная система связи. Систему связи, предназначенную для односторонней передачи информации между двумя заданными пунктами, будем называть одноканальной системой, или просто каналом связи. Простейшая блок-схема канала связи приведена на рис. 21

*) Отметим, однако, что полное отвлечение от свойств человека не всегда целесообразно: учет этих свойств ставит новые проблемы в технике связи. Например, в телевидении существует колоссальный разрыв между количеством информации, которое способен принять в единицу времени человек (десятки бит в секунду), и количеством информации, которое способна переносить система телевизионного вещания (миллионы бит в секунду).

и содержит, кроме передающей и приемной систем и линии связи, источники помех, действующих в общем случае на все элементы системы. Помехи, возникающие (или создаваемые искусственно) в линии связи, называются внешними (в радиосвязи к их числу относятся, например, промышленные помехи, атмосферные и космические шумы, помехи от посторонних радиосредств и т. п.). Под внутренними помехами обычно понимают шумы, возникающие в передающей и приемной системах (тепловые шумы сопротивлений, шумы электронных ламп и пр.). При некоторых условиях иногда оказывается возможным пренебречь влиянием помех; такая система связи называется каналом без помех. В других случаях исключить из рассмотрения помехи нельзя. Обычно, однако, считается, что внутренними шумами можно либо пренебречь по сравнению с внешними, либо что система допускает «пересчет» всех источников в один эквивалентный с известными характеристиками, и этот источник помех выносится в линию связи.

Целесообразно различать каналы связи, работающие на непрерывных и дискретных сигналах. Имеется в виду дискретность канала в том смысле, что множества элементарных символов на входе и выходе линии связи дискретны и конечны. Проблемы, возникающие при рассмотрении дискретных каналов связи, можно проиллюстрировать [68] с помощью схемы на рис. 22. От общей схемы (рис. 21) она отличается

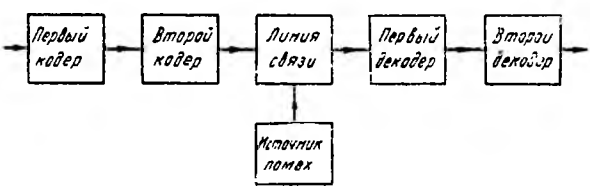

Рис. 22.

детализацией и передающей, и приемной систем: передающая система представлена двумя кодирующими, а приемная — двумя декодирующими устройствами.

Входным сигналом системы связи может служить печатный текст или графическое изображение, звуковая волна, показания прибора, и т. п. Назначение первого кодирующего устройства состоит в том, чтобы представить входной сигнал в некоторой стандартной форме, например, в виде последовательности двоичных символов. Основная проблема такого перекодирования заключается в том, чтобы стандартное представление было наиболее экономичным, т. е. требова-

ло бы (в среднем) наименьшего возможного числа двоичных символов.

Благодаря наличию помех в линии связи, соответствие между отправленным и принимаемым символами перестает быть однозначным; поэтому в общем случае попытки определить по принятому символу, какой из возможных символов был отправлен, неизбежно связаны с ошибками. Однако имеется возможность ослабления влияния помех с помощью подходящего перекодирования сигнала. В простейшем случае такую возможность можно реализовать с помощью многократного повторения передачи и последующего сличения полученных текстов. Такой метод, однако, применим лишь при малых вероятностях ошибок и, кроме того, резко увеличивает необходимое время передачи. Существуют способы более эффективного кодирования. Назначение *второго кодирующего устройства* и *первого декодирующего устройства* и заключается в реализации избранного метода помехоустойчивого кодирования и декодирования сигнала. Главная проблема при этом состоит в том, чтобы максимально снизить вероятность ошибок, хотя помехи случайным образом искажают полезный сигнал*). В идеальном случае выходной сигнал первого кодирующего устройства совпадает с входным сигналом второго декодирующего устройства. Итак, функция второго кодирующего устройства состоит в представлении стандартной последовательности символов в избранном новом коде, а первое декодирующее устройство по принятой последовательности восстанавливает сигнал снова в стандартной форме.

Наконец, функция *второго декодирующего устройства* сводится к восстановлению входного сигнала всей системы; при этом считается, что стандартное представление сигнала было безошибочно определено первым декодирующим устройством.

2. *Многоканальная система связи.* Довольно часто возникает необходимость передачи информации от группы близких источников к группе получателей, сосредоточенных в другом пункте. Наглядным примером может служить необходимость передачи данных от различных приборов, установленных на искусственном спутнике, на группу наземных регистрирующих устройств. Другим примером может служить телефонная связь между двумя крупными городами. С одной стороны, необходимо создать для каждой пары отправитель—получатель отдельный канал связи. С другой стороны, экономические соображения (например, высокая стоимость

*) Впоследствии мы убедимся, что при определенных условиях вероятности ошибок могут быть сделаны сколь угодно малыми.

сооружения проводной линии связи) или технические трудности (которые возникли бы, например, при создании отдельной линии связи для каждого прибора на спутнике) препятствуют увеличению числа отдельных линий связи для каждой пары связываемых объектов. Выход состоит в том, чтобы объединить каналы, направив всю информацию по одной линии связи (если, конечно, она допускает это). Такие комбинированные системы связи и называются многоканальными.

Для того, чтобы сигнал, адресуемый конкретному получателю, поступал только к нему, необходимо, очевидно, снабдить сигналы различных каналов некоторым дополнительным физическим признаком, параметром отбора, по которому на приемном конце и производилась бы фильтрация. Поэтому в многоканальной системе связи появляются дополнительно устройства для разделения сигналов, принадлежащих разным каналам (см. рис. 23). Для многоканальных систем специфи-

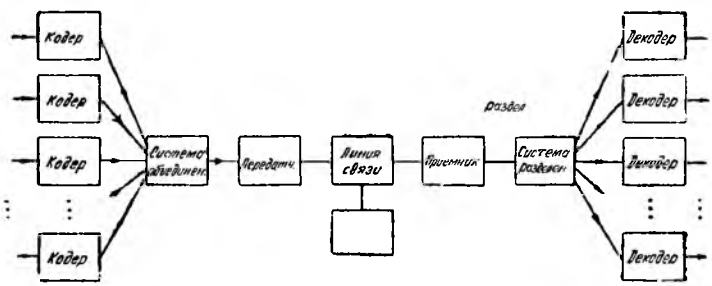

Рис. 23.

ческой особенностью является то, что в результате неидеальности разделения сигналы соседних каналов несколько искажают сигнал данного канала; эти так называемые перекрестные помехи обычно являются основным видом помех в таких системах.

3. Многопутевые (многолучевые) каналы связи. В ряде случаев приходится пользоваться такими каналами связи, сигнал в которых по некоторым физическим причинам расщепляется на несколько компонент. При этом каждая из компонент сигнала следует по отдельному пути, претерпевая специфические для этого пути преобразования (изменения, задержку во времени, а иногда и нелинейные искажения). Если бы удавалось на приемном конце разделить сигналы, пришедшие по разным путям, то мы имели бы одну из рассмотренных выше систем связи. Однако часто такой возможности не имеется, и приемник фиксирует сигнал, являющийся результатом некоторого суммарного воздействия ком-

понент, пришедших по разным путям. Такие системы связи принято называть многопутевыми (многолучевыми).

Примером многолучевой системы связи может служить радиосвязь на коротких волнах, которые распространяются путем земной волны и путем отражения от ионосферы (иногда претерпевая многократное отражение от ионосферы и поверхности земли). Другой пример — радиосвязь на рассеянии, при которой неоднородности тропосферы или ионосферы могут рассматриваться как отдельные рассеивающие центры, излучающие отдельные компоненты сигнала.

4. Системы связи со случайными параметрами. Основные трудности при рассмотрении многопутевых каналов возникают не столько в связи с необходимостью учета соотношений между компонентами сигнала, сколько в связи с тем, что необходимо учитывать временные изменения условий прохождения компонент по различным путям. Эти изменения обычно носят случайный характер, что и вызывает ряд затруднений. Для преодоления этих трудностей удобным средством оказалось построение модели системы связи со случайными параметрами.

В некоторых случаях к рассмотрению многопутевых каналов со случайным изменением параметров компонент сигнала можно подойти чисто феноменологически: рассматривать всю систему как одноканальную, однопутевую, а все вероятностные свойства отнести к случайному изменению некоторых воображаемых параметров такого канала. Например, если нас интересуют только флуктуации амплитуды принимаемого сигнала, их можно отобразить, смоделировать введением случайно изменяющегося (с соответствующим распределением) затухания линии связи. Флуктуации фазы можно моделировать введением случайно меняющейся «задержки».

Подобная модель оказывается полезной не только при рассмотрении многолучевых систем связи; существуют системы, которые непосредственно отображаются такой моделью (например, связь на метеорных следах), что делает ее изучение еще более важным.

5. Сложные системы связи. Для некоторых целей приходится сооружать комплексные системы связи: дублировать каналы для повышения надежности связи; передавать информацию по последовательности каналов с различными свойствами; сооружать сети со сложным переплетением каналов и т. п. Такие составные системы связи будем называть сложными. Сложные системы связи обладают рядом специфических особенностей, которые должны учитываться при их построении и использовании.

§ 3. СИСТЕМЫ ХРАНЕНИЯ ИНФОРМАЦИИ

В подавляющем большинстве практических ситуаций информация, имеющаяся к моменту времени t, не может или не должна использоваться немедленно; но обычно есть уверенность, что эта информация потребуется в дальнейшем. Для обеспечения переноса информации во времени и создаются разнообразные системы хранения информации*). Примерами таких систем могут служить магнитофон и записная книжка, библиотека и запоминающие устройства ЭЦВМ, атлас географических карт и картинная галерея, таблицы функций и т. д.

Для технических систем хранения информации (которые и представляют основной интерес при теоретико-информационном подходе) главными характеристиками являются информационная емкость (т. е. максимальное количество информации, которое способна хранить система) и долговременность хранения информации без потерь (или с потерями, не превышающими допустимых пределов). При рассмотрении быстродействующих систем, в которые входят запоминающие устройства, важным параметром является время выборки, т. е. промежуток времени между моментом обращения к памяти и моментом получения нужной информации.

Разнообразие систем хранения информации очень велико, даже если не вдаваться в технические отличия между конкретными системами. Различают (по длительности хранения) долговременные и оперативные запоминающие устройства; имеются ЗУ, к которым можно обращаться сколько угодно раз, но есть ЗУ, хранящие информацию лишь до первого обращения к ним. Существуют ЗУ, допускающие обращение к ним лишь в фиксированные моменты времени (обычно периодически повторяющиеся); имеются ЗУ, к которым можно обращаться в произвольные моменты времени.

При всем многообразии систем хранения информации в них можно выделить несколько основных подсистем с различным функциональным назначением. Кроме собственно хранилища информации, имеется адресная система, обеспечивающая отыскание по ряду признаков нужной ячейки хранилища; входные и выходные блоки обеспечивают возможность запроса и выдачи данных, а также занесения новой информации как путем вытеснения ненужной, так и путем заполнения свободных ячеек.

Введение запоминающих устройств в сложные информационные системы значительно расширяет возможности послед-

*) Системы хранения информации, используемые в электронной вычислительной и управляющей технике, называются также запоминающими устройствами, или памятью.

них и, конечно, значительно увеличивает трудности их исследования. Поэтому обычно системы с памятью и системы без памяти рассматриваются обособленно.

§ 4. ПРЕОБРАЗОВАТЕЛИ ИНФОРМАЦИИ

Всевозможные операции, происходящие в информационных системах, не сводятся лишь к передаче и хранению информации. Во многих системах самым существенным является переработка информации, иногда простая, иногда весьма сложная. Для такой переработки служат специальные устройства, *преобразователи информации*. Преобразование информации ставит ряд сложных проблем, лишь часть из которых разрешена теорией информации в достаточной степени полно.

К числу относительно простых, и в то же время наиболее часто встречающихся преобразований относится *перекодирование*. Мы уже имели возможность обсудить функционирование кодирующих и декодирующих устройств, входящих в состав систем связи (см. § 2). По существу перекодирование является операцией перехода от представления некоторой информации в одном коде к представлению ее в другом коде (под кодом понимается вся совокупность правил образования сигнала). Это высказывание, будучи верным, носит слишком общий характер (например, под такое определение подходит как переписывание текста, так и конспектирование его) и требует уточнения. Будем называть перекодированием такое преобразование одного сигнала в другой, при котором количество информации, несомое вторым сигналом, равно*) количеству информации, несомому первым. Это, конечно, не означает, что и каждый элемент вторичного сигнала несет ту же нагрузку, что и элемент первичного сигнала: перекодирование может осуществляться так, что один сигнал будет содержать большее число элементов, чем другой.

По принципу действия все кодирующие устройства можно разбить на два класса. К первому относятся *кодирующие устройства без памяти*, которые осуществляют перекодирование сигнала поэлементно, или мгновенно. Примерами таких устройств могут быть: первое кодирующее и второе декодирующее устройства в схеме системы связи, рассмотренной в § 2; текущее (мгновенное) перекодирование осуществляется при записи или воспроизведении звука на магнитофоне, при усилении напряжения безынерционным усилителем,

*) Более общо: «сколь угодно мало отличается», или даже «отличается на величину, не превышающую заданной», если при перекодировании допускаются ошибки.

при фотографировании, при передаче телеграмм простой азбукой Морзе, при телефонном разговоре и т. д. и т. п. Более сложным для изучения объектом являются кодирующие устройства с памятью. Их особенность состоит в том, что каждый элемент выходного сигнала устройства определяется в общем случае не одним элементом входного сигнала, а некоторым множеством таких элементов. Примерами таких устройств служат устройства, осуществляющие оптимальное кодирование для передачи сигнала по каналу с шумами; устройства для передачи секретными кодами (исключая простейшие коды); а также фильтры с конечной полосой пропускания и т. п.

К числу преобразователей информации, тоже сравнительно подробно изученных, относятся накопители информации. Неотъемлемой частью накопления информации является ее запоминание; поэтому иногда запоминающие устройства рассматривают как накопители. Однако накопление может (или должно) иногда производиться не путем запоминания всех входных сигналов, а путем запоминания результата некоторой обработки этих сигналов. Накопителем является, например, сумматор, по выходному сигналу которого нельзя однозначно определить слагаемые. Накопителем является прибор, выдающий по входной реализации случайного процесса его гистограмму. Цель, с которой создаются накопители, — накопление нужной информации. Если вся поступающая информация нужна, то накопитель является просто хранителем информации, запоминающим устройством. Если же входные сигналы несут не только полезную информацию, но и ненужную, то накопитель играет роль фильтра, отбирающего и накапливающего только то, что необходимо для дальнейшего использования; в этом случае накопление не сводится к простому запоминанию. Примером накопителей такого типа являются системы накопления сигналов в радиолокационных станциях, участвующие в процессе обнаружения полезных сигналов в шумах.

Перечисляя типичные системы преобразования информации, следует упомянуть сравнивающие устройства, устанавливающие степень сходства сравниваемых сигналов; решающие устройства, т. е. системы, отображающие пространство входных сигналов на пространство решений (управляющих сигналов); квантующие устройства, ставящие в соответствие непрерывным сигналам их дискретные отображения; фильтры, осуществляющие отбор сигналов по некоторым признакам, и целый ряд других систем обработки информации. С каждой из таких систем связан ряд информационных проблем, в первую очередь—вопрос об оптимальности работы таких систем в смысле минимальных по-

терь полезной информации (или в некотором близком к этому смысле).

Важно подчеркнуть, что наиболее сложные преобразования информации по существу еще не исследованы теорией информации. Сюда, например, относятся: составление справочников, реферирование и рецензирование статей; конспектирование текстов и пр. Такие преобразования можно назвать смысловой фильтрацией; и до тех пор, пока в теорию не будет введена характеристика смысла, рассмотрение этих преобразований невозможно.

§ 5. ДРУГИЕ ТИПЫ ИНФОРМАЦИОННЫХ СИСТЕМ

Многообразие информационных систем весьма велико, и, видимо, вопросы их классификации еще будут обсуждаться в научной литературе. Кроме рассмотренных выше трех типов информационных систем, можно различить еще несколько характерных групп систем, из которых кратко обсудим системы измерения и системы исследования.

Основные особенности систем измерения связаны с особенностями измерительных сигналов, которые были обсуждены в гл. I. В состав системы измерения входят: генераторы зондирующих и эталонных сигналов; линии связи, по которым сигналы подаются и отводятся от объекта измерения; система сравнения сигнала от измеряемого объекта с эталоном. Основными проблемами при проектировании систем измерения являются: выбор наиболее информативных сигналов; уменьшение шумов (погрешностей измерения), т. е. доведение до возможного минимума влияния неконтролируемых изменений не интересующих нас объектов; отыскание таких способов обработки данных, которые позволили бы извлекать максимум наличной информации.

Системы исследования создаются для приема и расшифровки сигналов, код которых неполностью известен. К числу таких систем относятся устройства для приема и извлечения информации из естественных сигналов (см § 5 гл. I), системы, создаваемые для перехвата и расшифровки радиограмм противника, закодированных секретным «разгадоустойчивым» кодом, и т. п. Одним из основных моментов в работе таких систем является выдвижение и проверка гипотез относительно неизвестного кода. Эта процедура ставит весьма сложные проблемы. Ведь если относительно неизвестного кода может быть выдвинуто очень большое число гипотез, то простой перебор их становится бесперспективным делом. Возникает задача отбора «наиболее правдоподобных» гипотез, решение которой опирается на сравнительную оценку правдоподобности гипотез. Введение этой оценки — далеко не тривиальное

дело. Естественно, что чем больше информации учитывает та или иная оценка, тем эффективнее становится процедура исследования. Приложение теории информации к исследованию таких систем представляется вполне оправданным.

Другой важный момент заключается в том, что даже в том случае, если процедура разгадывания резко выделила одну из гипотез, необходимо иметь в виду относительность вывода о том, что именно эта гипотеза верна. Мы можем (и должны) действовать в соответствии с этим выводом; успешность действий, опирающихся на предположение о верности данной гипотезы, будет всякий раз повышать правдоподобность этого предположения. Но если однажды мы потерпим неудачу, — это не должно подрывать нашей веры в познаваемость мира. Яркие примеры «крушения законов» дает нам история науки; и всегда «подрыв принципов науки» на основе фактов был шагом вперед в познании мира. Возможно, что сейчас физика пошла вновь к этому этапу: обнаружены такие галактики, строение которых нельзя объяснить, если считать, что закон всемирного тяготения справедлив всегда и везде.

Заканчивая обсуждение типов информационных систем, отметим еще раз, что весьма часто в практике встречаются системы, которые нельзя однозначно отнести к какому-либо из указанных типов. Такие системы представляют собой сложный комплекс, в который входят каналы связи, системы переработки и хранения информации и другие устройства. Примером может служить универсальная цифровая вычислительная машина. Еще более сложный комплекс представляет собой, например, система управления запуском спутника земли. Такие сложные системы могут рассматриваться и по частям; но ряд вопросов (особенно некоторые вопросы оптимальности) должен иногда решаться с учетом взаимодействия всех частей.

Укажем также на весьма мало исследованный и очень перспективный класс информационных систем, характеристики которых изменяются в ходе работы систем таким образом, что свойства системы в целом улучшаются (в некотором смысле); это так называемые самонастраивающиеся, или самоорганизующиеся системы.

ГЛАВА VII

ИСТОЧНИКИ ИНФОРМАЦИИ. СКОРОСТЬ СОЗДАНИЯ ИНФОРМАЦИИ; ε-ЭНТРОПИЯ НЕПРЕРЫВНЫХ И ДИСКРЕТНЫХ ИСТОЧНИКОВ

§ 1. ОПРЕДЕЛЕНИЕ СКОРОСТИ СОЗДАНИЯ ИНФОРМАЦИИ

При рассмотрении процесса передачи информации по некоторой системе связи вполне четко разделяются следующие вопросы: какое количество информации передается по данной системе в единицу времени; каково максимальное количество информации, которое данная система может передать в единицу времени; каково количество информации, поступающее на вход канала в единицу времени. Эти количества информации естественно называть соответственно *скоростью передачи информации*, *пропускной способностью* и *скоростью создания информации*. Первые две величины определяются свойствами канала и способом кодирования на его входе; вопросы, связанные с этими величинами, будут обсуждены в гл. VIII. В данной главе будут рассмотрены вопросы описания источников информации, питающих каналы, прежде всего — описание того свойства источников, которое проявляется в конечности количества информации, выдаваемого любым источником за конечный интервал времени.

При передаче дискретных (по информативному параметру) сигналов проблема определения скорости создания информации представляется довольно простой: скорость создания информации можно определить как энтропию источника на единицу времени*). При рассмотрении непрерывных источников вопрос не решается столь просто: абсолютная мера неоп-

*) Как будет видно из дальнейшего, такое определение скорости создания информации дискретным источником требует уточнения.

ределенности в этом случае бесконечна, а дифференциальная энтропия определяется с точностью до произвольной постоянной. Однако требование бесконечной точности воспроизведения непрерывной реализации любым источником является нереальным, практически неосуществимым, а поэтому и не имеющим смысла. Две реализации, отличающиеся (в определенном смысле) на величину, не большую некоторой заданной, воспринимаются как сигналы, несущие совершенно идентичную информацию. Это соображение переводит вопрос в иную плоскость: необходимо дать определение скорости создания информации, если известно, с какой точностью источник воспроизводит заданный непрерывный сигнал.

Конкретизируем сначала, что понимается под заданной точностью воспроизведения. Пусть $x(t)$ некоторая заданная реализация, которую необходимо передать, а $x'(t)$ — реализация, которая в действительности передается. Будем считать, что можно указать количественно, насколько x отличается от x'; другими словами, задается некоторая разумная мера, отличия x от x', $\rho(x, x')$. С помощью этой величины и определяется параметр ε точности воспроизведения. Сделать это можно различным образом; примерами могут служить следующие требования:

$$P\{\rho(x, x') \leqslant \varepsilon\} \geqslant 1 - \delta, \tag{1.1}$$

$$M\rho^2(x, x') \leqslant \varepsilon^2, \tag{1.2}$$

$$\rho(x, x') \leqslant \varepsilon \quad \text{и т. д.} \tag{1.3}$$

Простым случаем будет задание $\rho(x, x')$ в виде разности $x(t) - x'(t)$, тогда (1.2) означает ограничение на дисперсию ошибки воспроизведения, а (1.3)—на максимальное значение разности. Более сложным будет частотно-взвешенный критерий и т. д.

Рассматривая теперь процессы $X = \{x(t)\}$ и $X' = \{x'(t)\}$, мы можем утверждать, что они содержат информацию друг о друге. Будем оперировать с количеством информации $I(x, x')$, приходящимся в среднем на единицу времени. Это количество информации зависит не только от параметра точности ε, но и от характера статистической связи x и x'. Определим теперь **скорость создания информации** как **минимальное количество информации,** которое необходимо (в единицу времени) для того, чтобы реализация $x'(t)$ с заданной точностью ε воспроизводила реализацию $x(t)$ (при заданном распределении $p(x)$):

$$H_\varepsilon^a(x) = \min_{\{p(x'|x)\}} I(x, x'). \tag{1.4}$$

Отметим, что численное значение величины $H_\varepsilon^\alpha(x)$ в общем случае будет различным для разных определений параметра точности ε, в связи с чем и введен индекс α.

По ряду причин, в первую очередь в связи с явной аналогией между $H_\varepsilon^\alpha(x)$ для непрерывного источника и энтропией дискретного источника, А. Н. Колмогоров предложил [20] называть величину $H_\varepsilon^\alpha(x)$ ε-энтропией. Колмогоров подчеркнул, что понятие ε-энтропии представляет и более широкий интерес. В частности, можно указать на интересную интерпретацию ε-энтропии с помощью пространства сигналов. Каждая точка этого пространства ставится в соответствие некоторой определенной реализации непрерывного процесса; функция $\rho(x, x')$, количественно характеризующая различие двух реализаций, рассматривается как расстояние между соответствующими точками. На континуальном множестве возможных сигналов размещается „ε-сеть", т. е. дискретное подмножество точек, удовлетворяющее условию, чтобы ε-окрестности узлов сети, не прекрываясь, в совокупности полностью охватывали все пространство сигналов. Мера ε-окрестности принимается за вероятность соответствующего узла; это позволяет вычислить энтропию ε-сети. Варьируя размещение узлов сети, можно найти такую ε-сеть, которая будет обладать минимальной энтропией; это и будет ε-энтропия рассматриваемого множества непрерывных сигналов.

§ 2. ε-ЭНТРОПИЯ ГАУССОВЫХ ИСТОЧНИКОВ

Вычисление ε-энтропии в общем случае является сложной задачей. Однако в случае гауссовых величин и процессов особых сложностей не возникает; получающиеся при этом результаты имеют важное значение в связи с экстремальностью нормального распределения для ε-энтропии при среднеквадратичном определении точности воспроизведения.

Начнем с рассмотрения гауссовой случайной величины; пусть величина x распределена нормально, $N_x(0, \sigma^2)$. Пусть, далее, случайная величина x' с некоторой погрешностью η воспроизводит величину x: $x = x' + \eta$. Наложим требование, чтобы дисперсия погрешности была равна заданной величине ε^2, $\varepsilon^2 = D\eta$; припишем среднеквадратичному критерию точности индекс $\alpha = 0$. Тогда

$$H_\varepsilon^0(x) = \min I(x, x') = \min [H_\varepsilon(x) - H_\varepsilon(x \mid x')] =$$

$$= H_\varepsilon(x) - \max H_\varepsilon(x \mid x'). \qquad (2.1)$$

Здесь $H_\varepsilon(\cdot)$ — дифференциальные энтропии соответствующих

распределений. По предположению, $p(x) = N_x(0, \sigma^2)$, следовательно, $H_\varepsilon(x) = \frac{1}{2} \log (2 \pi e \sigma^2)$. Далее, так как $x = x' + \eta$, то максимум $H_\varepsilon(x \mid x') = H_\varepsilon(x' + \eta \mid x')$ при условии $D_\eta = \varepsilon^2$ достигается, если а) η и x' независимы и б) η нормально. При этом

$$\max H_\varepsilon (x \mid x') = \frac{1}{2} \log (2 \pi e \varepsilon^2). \tag{2.2}$$

Таким образом, ε-энтропия нормальной величины равна

$$H_\varepsilon^0(x) = \frac{1}{2} \log \frac{\sigma^2}{\varepsilon^2}. \tag{2.3}$$

Обратимся теперь к простейшему гауссову процессу: пусть спектр процесса ограничен верхней частотой F и равномерен в этой полосе частот. Мощность источника (дисперсию процесса) обозначим через P. Взяв снова среднеквадратичный критерий различия, пользуясь теоремой Котельникова и некоррелированностью отсчетов (спектр равномерен), приводим задачу к только что рассмотренной. Так как на единицу времени приходится $2F$ отсчетов, то, сразу воспользовавшись формулой (2.3), получаем, что ε-энтропия рассматриваемого случайного процесса на единицу времени выразится как

$$H_\varepsilon^0(x) = F \log \frac{P}{N_\varepsilon}, \tag{2.4}$$

где через $N_\varepsilon = \varepsilon^2$ обозначена мощность „шумов“ источника.

Сделаем одно важное замечание в связи с этой формулой Шэннона [7], а именно, подчеркнем, что мощность „шума“ источника N_ε не может быть больше мощности сигнала P. Из независимости x' и η следует, что $D(x) = D'(x) + D(\eta)$, откуда

$$N_\varepsilon = D(\eta) = D(x) - D(x') \leqslant D(x) = P. \tag{2.5}$$

При $N_\varepsilon = P$ $H_\varepsilon^0 = 0$; это означает, что при точности воспроизведения, равной дисперсии сигнала, достаточно задавать одну-единственную (а именно—нулевую) реализацию; ясно, что при этом информация не передается.

Формула (2,4), выведенная для белого гауссова процесса позволяет перейти к важным обобщениям. Отношение

$P/N_\varepsilon = (P/F)/(N_\varepsilon/F)$ можно рассматривать как отношение спектральных плотностей. Пусть теперь источник по-прежнему является гауссовым, но с произвольной спектральной плотностью $P(f)$, не обязательно ограниченной по частоте. Разбив всю область частот на достаточно малые интервалы df, в которых $P(f)$ можно считать постоянной, мы получаем возможность применить (2,4) к элементарным интервалам:

$$dH_\varepsilon^0 = df \log \frac{P(f)}{N_\varepsilon(f)}. \tag{2.6}$$

Теперь, чтобы найти ε-энтропию рассматриваемого процесса, необходимо найти минимальное значение функционала

$$H_\varepsilon^0 = \int \log \frac{P(f)}{N_\varepsilon(f)}\, df \tag{2.7}$$

при условиях

$$\int N_\varepsilon(f)\, df = \varepsilon^2 \tag{2.8}$$

и (см. 2.5)

$$N_\varepsilon(f) \leqslant P(f). \tag{2.9}$$

Решая вариационную задачу (2.7)—(2.8) обычным методом (с тем, чтобы условие (2.9) привлечь позже), получим, что спектр $N_\varepsilon(f)$ удовлетворяет условию

$$N(f) = \lambda^2 = \text{const}. \tag{2.10}$$

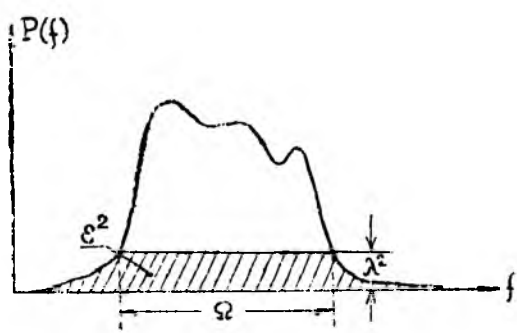

Рис. 24.

Так как должно выполняться условие (2.9), окончательно получаем, что вид спектра $N_\varepsilon(f)$ определяется (см. рис. 24) соотношением

$$N_\varepsilon(f) = \min(\lambda^2,\ P(f)). \tag{2.11}$$

Таким образом, для вычисления ε-энтропии гауссового процесса с произвольным спектром необходимо найти параметр λ^2 из уравнения

$$\int_0^\infty \min(\lambda^2,\ P(f))\, df = \varepsilon^2 \tag{2.12}$$

и область Ω частот, где $P(f) > \lambda^2$, а затем вычислить интеграл

$$H_\varepsilon^0(x) = \int_\Omega \log \frac{P(f)}{\lambda^2}\, df. \tag{2.13}$$

Этот результат позволяет [20] указать, в каком именно смысле сигнал с неограниченным спектром может считаться имеющим ограниченный спектр: 1) реальный сигнал должен иметь спектр, обладающий достаточно быстро спадающими „хвостами" и концентрированной основной частью (см. рис. 24); 2) параметр точности ε должен быть не слишком малым; при этих условиях скорость создания информации приближенно такова, как если бы вне области Ω вообще не было бы частотных компонент.

§ 3. ε-ЭНТРОПИЯ ДИСКРЕТНОГО СЛУЧАЙНОГО ОБЪЕКТА

Принимая энтропию дискретных источников информации за скорость создания информации (см. § 1), мы тем самым неявно выдвигаем требование точного воспроизведения источником любой реализации. Однако вопрос о конечной точности воспроизведения может быть поставлен и по отношению к дискретному источнику.

Пусть задан дискретный случайный объект X:

$$X = \begin{Bmatrix} x_1,\ x_2, \dots,\ x_K, \dots \\ p_1,\ p_2, \dots,\ p_K, \dots \end{Bmatrix}. \tag{3.1}$$

Определим теперь объект X', воспроизводящий X с некоторой заданной точностью ε. Поскольку возможные состояния $\{x_k\}$ объекта X могут иметь различную природу (в том числе и быть качествами), не имеет смысла говорить о количественном различии между возможными символами x_k и x'_k. Пусть поэтому объект X' имеет те же возможные состояния, что и X. Условие близости X и X', определяющее

„точность„ воспроизведения, задается на множестве совместных распределений X и X'. Тогда ε-энтропию дискретного объекта можно определить как

$$H_\varepsilon^\partial(X) = \min_{\{P_{XX'}\}_\varepsilon} I(X, X'), \tag{3.2}$$

где $\{P_{XX'}\}_\varepsilon$ — множество совместных распределений, удовлетворяющих условию ε-близости. Это условие естественно задать в одном из видов:

$$P(X \neq X') \leqslant \varepsilon, \tag{3.3}$$

$$P(X \neq X' \mid X') \leqslant \varepsilon. \tag{3.4}$$

Как будет видно из дальнейшего, ε-энтропии, соответствующие (3.3) и (3.4), оказываются равными.

Поскольку преобразование $X \to X'$ можно рассматривать как кодирование источником, то распределение, при котором обеспечивается минимальность $I(X, X')$, будем кратко называть „экономичным кодом“. В. Ерохиным, рассуждения которого [61[здесь воспроизводятся, указан соответствующий экономичный код и дана явная формула для $H_\varepsilon^\partial(X)$.

Будем в дальнейшем предполагать, что для всех P_k выполняется условие

$$P_k < 1 - \varepsilon, \tag{3.5}$$

так как в противном случае $H_\varepsilon^\partial(X) = 0$: достаточно с вероятностью 1 воспроизводить то x_r, для которого $P_r \geqslant 1 - \varepsilon$, чтобы условие ε-близости было выполнено.

Обозначим через r_{kl} совместные вероятности того, что $X = x_k$, $X' = x_l$. Тогда

$$\sum_l r_{kl} = P_k, \tag{3.6}$$

$$\sum_k r_{kl} = q_l, \tag{3.7}$$

где q_l — вероятность состояния x_l для объекта X'.

Так как

$$I(X, X') = H(X) - MH(X|X'),$$

то для нахождения минимума $I(X, X')$ нужно максимизировать по r_{kl} величину

$$MH(X \mid X') = \sum q_l \log q_l - \sum_{k,l} r_{kl} \log r_{kl}, \tag{3.8}$$

при условиях (3.5), (3.6). Решение задачи, проводимое методом множителей Лагранжа с привлечением неравенства Иенсена для выпуклых функций, получается довольно громоздким, и в деталях здесь приводиться не будет. Укажем лишь конечный результат.

Введем вспомогательную величину θ. Перенумеруем, если надо, возможные состояния x_k так, чтобы вероятности p_k не возрастали с увеличением k. Определим для всякого θ, $0 < \theta < p_1$ величины $s(\theta)$ и $n(\theta)$:

$$s(\theta) = \sum_{p_k > \theta} p_k, \tag{3.9}$$

$$n(\theta) = \sum_{p_k > \theta} 1. \tag{3.10}$$

В качестве θ будем брать решение уравнения

$$\theta [n(\theta) - 1] = s(\theta) + \varepsilon - 1. \tag{3.11}$$

Можно показать, что $\theta(\varepsilon)$ является непрерывной, монотонно возрастающей однозначной функцией ε.

Основной результат состоит в следующем. Экономичный код достигается, если вероятности q_k определить как

$$q_k = \frac{p_k^*}{\sum_k p_k^*} \tag{3.12}$$

$$p_k^* = \begin{cases} p_k - \theta & \text{при } p_k > \theta, \\ 0 & \text{при } p_k \leqslant \theta, \end{cases}$$

а совместные вероятности r_{kl} как

$$r_{kl} = p_k^* \delta_{kl} + (p_k - p_k^*) q_l. \tag{3.13}$$

(При этом из (3.11) — (3.13) следует, что

$$P(X = x_k,\ X' = x_k) = (1 - \varepsilon) q_k,$$

т. е. одновременно

$$P(X \neq X') \leqslant \varepsilon \text{ и } P(X = X' \mid X') \leqslant \varepsilon,$$

откуда и следует упомянутое выше равенство соответствующих ε-энтропий).

Таким образом, объект X', воспроизводящий объект X с точностью ε, имеет возможными только те состояния x_κ, вероятности p_κ которых превышают величину $\theta(\varepsilon)$.

Явная формула для $H_\varepsilon^\partial(X)$ получается в виде

$$H_\varepsilon^\partial(X) = \sum_{p_\kappa > \theta} p_\kappa \log \frac{1}{p_\kappa} - [n(\theta) - 1]\,\theta \log \frac{1}{\theta} - \\ - (1-\varepsilon) \log \frac{1}{1-\varepsilon}. \tag{3.14}$$

Полезно, наконец, неравенство

$$H_\varepsilon^\partial(X) \geqslant H(X) - \varepsilon \log \frac{n-1}{\varepsilon} - (1-\varepsilon) \log \frac{1}{1-\varepsilon}, \tag{3.15}$$

которое переходит в равенство, когда $\varepsilon \leqslant (n-1)\,p_n$.

ГЛАВА VIII

ИНФОРМАЦИОННЫЕ ХАРАКТЕРИСТИКИ СИГНАЛОВ. ПАРАМЕТРЫ ИНФОРМАЦИОННЫХ СИСТЕМ И ИХ ЭЛЕМЕНТОВ

§ 1. ИЗБЫТОЧНОСТЬ

Одной из важнейших характеристик сигнала является количество информации, которое он содержит. Однако по ряду причин количество информации, несомое сигналом, обычно меньше, чем то количество информации, которое сигнал м о г б ы нести по своей физической природе; информационная нагрузка на каждый элемент сигнала меньше той, которую элемент способен нести. Для описания этого свойства сигналов введено понятие и з б ы т о ч н о с т и и соответствующие количественные меры избыточности.

Пусть сигнал длиной в n символов содержит количество информации I. Если такое представление информации обладает избыточностью, то же самое количество информации I может быть представлено с помощью меньшего числа символов. Обозначим через n_0 наименьшее число символов, представляющее количество информации I без потерь. Удельное количество информации на один символ (которое иногда называют с о д е р ж а т е л ь н о с т ь ю) в первом случае равно $I_1=I/n$, во втором $I_{1\max}=I/n_0$. Очевидно,

$$nI_1 = n_0 I_{1\max}. \tag{1.1}$$

В качестве меры избыточности R можно принять относительное удлинение сигнала, соответствующее данной избыточности:

$$R = \frac{n-n_0}{n} = 1 - \frac{n_0}{n} = 1 - \frac{I_1}{I_{1\max}}. \tag{1.2}$$

В зависимости от того, при каких условиях достигается максимум содержательности сигнала, различаются частные виды избыточности [79]. Пусть, например, мы хотим устранить избыточность, вызванную наличием статистической связи между символами, сохранив при этом вероятности отдельных символов. Тогда $I_{1\max} = I_1' = -\Sigma p_i \log p_i$, и избыточность, обусловленная взаимосвязью символов, выразится числом

$$R_p = 1 - \frac{I_1}{I_1'}. \tag{1.3}$$

Если избыточность из-за взаимосвязи устранена, то в общем случае остается избыточность, вызванная неэкстремальностью распределения вероятностей. При конечном числе m символов в алфавите максимальная содержательность, равная $I_0' = \log m$, достигается при равномерном распределении вероятностей, и избыточность, обусловленная только неэкстремальностью распределения, определится числом

$$R_\varphi = 1 - \frac{I_1'}{I_0'}. \tag{1.4}$$

Полная избыточность определится соотношением

$$R = 1 - \frac{I_1}{I_0'}. \tag{1.5}$$

Указанные характеристики избыточности связаны соотношением

$$R = R_p + R_\varphi - R_p \cdot R_\varphi, \tag{1.6}$$

из которого следует, что при малых R_p и R_φ полная избыточность приближенно равна сумме частных избыточностей:

$$R \approx R_p + R_\varphi. \tag{1.7}$$

Хотя с точки зрения наиболее экономичного и эффективного использования информационных систем естественно сводить избыточность сигналов до минимума, не следует думать, что избыточность — явление, играющее лишь отрицательную роль. Наоборот, именно избыточность обеспечивает информационную устойчивость сигналов при воздействии помех. Одна из основных задач теории информации состоит в том, чтобы определить минимальную избыточность, обеспечивающую заданную надежность при заданных свойствах помех. Именно

в ходе решения этой проблемы были получены наиболее важные результаты теории информации.

§ 2. СКОРОСТЬ ПЕРЕДАЧИ ИНФОРМАЦИИ. ПРОПУСКНАЯ СПОСОБНОСТЬ

Введение понятий энтропии, количества информации и избыточности позволяют характеризовать свойства информационных систем. Так, информационная емкость («объем памяти») запоминающего устройства исчисляется количеством информации, которое способно хранить данное устройство; величина избыточности сигналов в системе дает представление об эффективности использования данной системы и т. д. Однако для сравнения информационных систем только такого описания недостаточно. Обычно нас интересует не только передача данного количества информации, но передача его в возможно более короткий срок; не только хранение определенного количества информации, но хранение с помощью минимальной по объему аппаратуры и т. п. Для целей такого описания информационных систем вводятся дополнительные параметры. Рассмотрим — не только в качестве примера, но и потому, что это наиболее важный случай — введение таких параметров для систем передачи информации.

Пусть до приема средняя неопределенность того, какой из возможных сигналов будет передан, исчислялась априорной энтропией $H_T(x)$. Пусть, далее средняя апостериорная неопределенность того, какой сигнал был передан, равна $H_T(x\,|\,y)$. Следовательно, среднее количество информации, получаемое в результате всей передачи, равно $I_T = H_T(x) - H_T(x\,|\,y)$. Если передача одного сигнала длится T единиц времени, то естественно определить скорость передачи информации R как

$$R = \frac{I_T}{T} = \frac{1}{T}\,|H_T(x) - H_T(x\,|\,y)\,| = H(x) - H(x\,|\,y), \qquad (2.1)$$

где $H(x)$ и $H(x\,|\,y)$ — априорная и апостериорная энтропии, приходящиеся на единицу времени. Если, например, речь идет о дискретном канале связи, то единицей времени удобно считать время передачи одного символа; тогда $H(x)$ и $H(x\,|\,y)$ — энтропии на символ. Для непрерывных каналов единицей времени может служить либо обычная единица (секунда, например), либо интервал времени между отсчетами $\left(\frac{1}{2F}\text{ сек.}\right)$, в последнем случае $H(x)$ и $H(x\,|\,y)$ имеют смысл дифференциальных энтропий на степень свободы.

Скорость передачи информации R является характеристикой фиксированной системы связи, т. е. параметром, зависящим от свойств шумов в линии связи и от свойств передающего и приемного устройств. Желая охарактеризовать потенциальные возможности канала связи, естественно рассмотреть случай, когда отыскивается такой способ передачи сигналов, при котором скорость передачи информации по данному каналу максимальна. Так возникает понятие пропускной способности канала, которая определяется как

$$C = \max_{\{A\}} R = \max_{\{A\}} |H_A(x) - H_A(x|y)|, \qquad (2.2)$$

где $\{A\}$—множество возможных приемно-передающих систем, среди которых и отыскивается система с максимальной R при заданных свойствах линии, связывающей приемник и передатчик*).

Чтобы понятие пропускной способности получило точный смысл, необходимо конкретизировать, по какому множеству $\{A\}$ максимизируется R. Так как это множество можно определить по-разному, имеет смысл говорить о нескольких типах пропускных способностей. Условимся прежде всего, что будем рассматривать только такие системы передачи информации, сигналы в которых являются реализациями стационарных эргодических случайных процессов**).

Наиболее важным является случай, когда множество $\{A\}$ образовано приемно-передающими системами, отличающимися только способом кодирования при фиксированных таких параметрах, как число символов для дискретных каналов, средняя мощность передаваемого сигнала в непрерывном случае и т. п. Именно таким образом определил К. Шэннон [7] пропускную способность канала связи.

С другой стороны, В. И. Сифоров показал [77], что в ряде случаев целесообразно рассмотреть предел C, к которому стремится пропускная способность C при стремлении мощности полезного сигнала к бесконечности (таким образом, в $\{A\}$ включаются передатчики с различными мощностями). Оказалось, что все каналы связи разбиваются на два класса: каналы первого рода (терминология В. И. Сифорова), для которых указанный предел бесконечен, и каналы второго рода с конечной пропускной способностью при бесконечной мощно-

*) При более тщательном отношении к строгости формулировок и определений, математики определяют C как верхнюю грань R по $\{A\}$ [20].

**) В современном развитии теории имеется тенденция к ослаблению этих ограничений.

сти передатчика. Предел C В. И. Сифоров предложил называть собственной пропускной способностью канала.

Иногда на множество $\{A\}$ накладываются дополнительные ограничения, кроме ограничений, наложенных К. Шэнноном. В таких случаях говорят об условной пропускной способности.

Говоря о пропускной способности каналов с шумами, важно подчеркнуть, что при этом имеется в виду передача информации без потерь, т. е. при сколь угодно малой вероятности ошибок. Это может показаться неожиданным, так как при наличии шумов сигнал на входе приемника всегда отличается от сигнала на выходе передатчика, и потери информации кажутся неизбежными. Однако введение избыточности уменьшает вероятность ошибки. Основываясь на интуитивных соображениях (например, на опыте введения избыточности путем многократного повторения), легко придти к выводу, что при повышении требований к малости вероятности ошибки избыточность и в общем случае должна неограниченно возрастать, а скорость передачи стремиться к нулю. Здесь мы имеем блестящий пример того, как сильно интуиция может привести в заблуждение. К. Шэннон показал, что существуют такие способы введения избыточности, при которых обеспечиваются одновременно и сколь угодно малая вероятность ошибки и конечная (отличная от нуля) скорость передачи информации. Это замечательное открытие и было одной из главных причин столь большого внимания к новой теории связи. Теоремы, связанные с этим свойством систем связи, будут рассмотрены в главе X.

§ 3. ПОМЕХОУСТОЙЧИВОСТЬ, ЭФФЕКТИВНОСТЬ, НАДЕЖНОСТЬ

При разработке любой информационной системы возникает целый ряд проблем принципиального и технического характера, различной важности и различной сложности. Для разрешения этих проблем необходимо введение ряда дополнительных количественных параметров, описывающих различные свойства информационных систем. В данном параграфе мы сосредоточим внимание на наиболее важных требованиях принципиального значения, предъявляемых к любой информационной системе. Таких требований три:

1) система должна быть достаточно помехоустойчивой.

2) система должна быть достаточно эффективной.

3) система должна быть достаточно надежной.

Обсудим, что понимается под каждым из этих требований,

и как можно численно охарактеризовать соответствующие свойства в рамках теории информации. Рассмотрение будем вести на примере систем связи, но все рассуждения легко обобщить на любую информационную систему.

Помехоустойчивостью системы связи называется способность системы осуществлять передачу информации при наличии помех. Это качество системы может быть выражено в большей или меньшей степени, поэтому имеет смысл говорить о различной помехоустойчивости систем. Если при одинаковых помехах и одинаковых входных сигналах выходные сигналы одной системы более близки (в определенном смысле) к оригиналу, чем выходные сигналы другой системы, то говорят, что вторая система менее помехоустойчива. Чтобы иметь возможность сравнивать системы по помехоустойчивости, необходимо ввести параметры, численные значения которых будут характеризовать сопротивляемость системы помехам.

В зависимости от конкретных нужд исследователя или проектировщика может потребоваться более или менее детальное описание помехоустойчивости, поэтому имеется несколько способов введения количественных характеристик помехоустойчивости. Рассмотрим сначала способы описания помехоустойчивости дискретных систем. Эти системы характерны тем, что все возможные сигналы конечной длительности образуют дискретное конечное множество; пусть общее число возможных сигналов равно N. Действие шумов сводится к тому, что некоторые символы в сигнале подменяются другими, в результате чего вместо переданного (например, i-го) сигнала принимается другой (например, k-й) сигнал. Помехоустойчивость системы связи наиболее полно может быть охарактеризована набором вероятностей $\{P_{ik}\}$ того, что при передаче i-го сигнала будет принят k-й (i, k=1,2,..., N); и если мы хотим задать требования к помехоустойчивости системы с учетом ценности каждого из сообщений в отдельности, то задание всей матрицы $\{P_{ik}\}$ необходимо.

Однако сравнение систем по их матрицам $\{P_{ik}\}$ (которые можно назвать «стохастическими матрицами трансформации сообщений» [71]) связано с рядом затруднений, а часто и не необходимо: достаточно ввести более простые характеристики помехоустойчивости. К таким более простым параметрам относится, например, средняя вероятность ошибочного приема, $P_{\text{ош. ср.}}$:

$$P_{\text{ош.ср.}} = \sum_{i=1}^{N} p_i \cdot (1 - P_{ii}), \tag{3.1}$$

где P_i — вероятность передачи i-го сигнала.

Другим собирательным параметром, характеризующим помехоустойчивость системы, может служить остаточная средняя неопределенность относительно переданного сообщения, т. е. энтропия

$$H_0 = -(1 - P_{\text{ош.ср.}}) \log (1 - P_{\text{ош.ср.}}) - P_{\text{ош.ср.}} \log P_{\text{ош.ср.}}. \quad (3.2)$$

Можно ввести и другие параметры помехоустойчивости дискретной системы связи (см., например, [21]).

Для непрерывных систем связи описание помехоустойчивости требует специфического подхода, так как множество возможных сигналов даже конечной длительности несчетно. Действие шумов в линии связи сводится к тому, что вместо отправленного сигнала $x(t)$ на выходе приемника наблюдается другая функция времени, $y(t)$. Чем ближе $y(t)$ к $x(t)$ при заданном шуме, тем более устойчива система по отношению к данной помехе. Для количественного описания помехоустойчивости необходимо ввести меру различия двух функций $x(t)$ и $y(t)$. Чаще всего в качестве такой меры принимается средний квадрат разности сравниваемых функций (см., например, [21], [7]):

$$R_1 = \overline{(x(t) - y(t))^2} = \frac{1}{T} \int_0^T |x(t) - y(t)|^2 dt. \quad (3.3)$$

Это, конечно, не единственно возможная мера различия двух функций времени. «Расстояние» между функциями $x(t)$ и $y(t)$ может быть также определено с помощью так называемой абсолютной ошибки

$$R_2 = \overline{|x(t) - y(t)|} = \frac{1}{T} \int_0^T |x(t) - y(t)| dt. \quad (3.4)$$

Другим способом является «частотно-взвешенный эффективный критерий» [7]. Идея этого критерия состоит в том, чтобы придавать различным частотным компонентам разности x и y разные веса. Это эквивалентно пропусканию разности $x(t) - y(t)$ через фильтр с определенной переходной функцией $h(t)$; выходной сигнал такого фильтра выразится как

$$f(t) = \int_{-\infty}^{\infty} |x(\tau) - y(\tau)| h(t - \tau) d\tau. \quad (3.5)$$

«Расстояние» между функциями $x(t)$ и $y(t)$ определится как средняя мощность сигнала на выходе рассматриваемого гипотетического фильтра:

$$R_3 = \frac{1}{T}\int_0^T f^2(t)\,dt. \qquad (3.6)$$

Введенные выше меры различия отправляемого и принимаемого сигналов могут служить основой для характеристики помехоустойчивости систем. Например, система может считаться достаточно помехоустойчивой, если «расстояние» между отправленным сигналом и сигналом на выходе системы не превышает заданной величины. Важной задачей является определение скорости передачи информации и пропускной способности каналов при заданной точности воспроизведения непрерывных сигналов.

Следует указать, что в качестве меры помехоустойчивости могут быть приняты и другие числовые характеристики, например, логарифм обратной величины среднеквадратичной ошибки в непрерывном случае [21], минус логарифм вероятности ошибки в дискретном случае [79], различным способом введенные понятия эквивалентного отношения сигнала к шуму [80] и пр. Однако общие объективные свойства информационных систем (например, наличие порогового эффекта) не зависят от того, как определяются числа, отражающие эти свойства, и проявляются в любой правильной теории помехоустойчивости.

Повышение помехоустойчивости связано с увеличением избыточности и с соответствующим уменьшением скорости передачи информации, что в общем случае является нежелательным. В связи с этим вводится понятие эффективности информационной системы: из двух систем связи с одинаковыми по ширине полосами частот более эффективной является та, которая передаст заданное количество информации за меньший промежуток времени; из двух запоминающих устройств с одинаковым числом ячеек более эффективным является то, которое может хранить большее количество информации, и т. д. Стремление увеличить эффективность приводит к стремлению уменьшить избыточность; однако увеличение скорости передачи информации возможно лишь до некоторого предела (до R, равного пропускной способности), после чего уменьшение избыточности приведет лишь к потере помехоустойчивости. Более того, если кодирование осуществляется неоптимальным способом, то и при $R < C$ уменьшение избыточности уменьшает помехоустойчивость, хотя и увеличивает эффективность. Именно в этой связи и говорят [79], что требование эффективности противоречит требованию помехоустойчивости.

Как и помехоустойчивость, эффективность системы может оцениваться с различных точек зрения, т. е. можно ввести

несколько разных параметров, характеризующих эффективность количественно. Рассмотрим некоторые из них.

Во-первых, чем больше скорость передачи информации, тем более эффективно используется канал связи. Имея в виду, что максимальная скорость передачи есть пропускная способность, вводится [78] коэффициент использования канала.

$$\eta_1 = \frac{R}{C}, \tag{3.7}$$

который показывает, насколько близка скорость передачи к пропускной способности.

Далее, вследствие неоптимального кодирования при наличии шума скорость передачи информации может оказаться меньше, чем величина энтропии источника на единицу времени (последнюю удобно называть скоростью создания информации источником). Выбор более эффективного кода — один из источников повышения эффективности всей системы, поэтому для оценки эффективности целесообразно пользоваться [78] коэффициентом передачи информации, который определяется как отношение скорости передачи информации к скорости создания информации источником:

$$\eta_2 = \frac{R}{H}. \tag{3.8}$$

(Естественно, такой коэффициент целесообразно рассматривать лишь при $H < C$).

Перейдем теперь к третьему требованию к любой системе — надежности. Надежностью системы называется ее способность безотказно (т. е. не выходя из строя) работать в течение достаточно продолжительного времени.

На первый взгляд, может показаться, что это требование носит чисто технический характер и зря поставлено в один ряд с принципиально информационными требованиями эффективности и помехоустойчивости. На самом деле это не так. Стремление повысить помехоустойчивость и эффективность системы обычно приводит к необходимости усложнения структуры системы. Пока это касалось сравнительно простых устройств, на такие усложнения, обычно небольшие, соглашались. Бурный рост информационной техники привел к созданию сложных многокомпонентных систем (радиолокационных станций, радиорелейных систем связи, электронных вычислительных и управляющих машин и пр.), содержащих многие сотни и тысячи элементов и деталей. При этом оказалось, что даже если срок службы каждой детали достаточно велик, при доста-

точно большом числе деталей система в целом будет часто выходить из строя. При увеличении сложности системы средняя частота отказов возрастает, и это может служить причиной предпочтения менее совершенной, но более надежной системы перед системой более совершенной и сложной. Так проблема надежности из технической превратилась в принципиальную. Собственно, с проблемой помехоустойчивости происходила та же эволюция: пока работа систем велась в условиях «сильных» сигналов, вопросы шумов в системе были чисто техническими вопросами; при переходе к условиям «слабых» сигналов стало ясно, что для правильного решения вопросов приема сигналов в шумах необходим глубокий теоретический анализ.

Технический аспект в проблеме надежности играет, несомненно, весьма важную роль: повышение срока службы отдельных деталей является сильным средством повышения надежности системы. Для нас сейчас, однако, более важно подчеркнуть одну принципиально информационную сторону проблемы надежности, а именно — возможность создания систем с достаточно высокой надежностью из элементов с низкой надежностью. Один из путей приложения теории информации для решения этой проблемы можно усмотреть, если рассматривать структуру системы как сигнал, несущий информацию о функциональных свойствах системы. Выход из строя элемента схемы можно сопоставить с искажением символа в дискретном сигнале; при такой интерпретации требование функциональной устойчивости при выходе из строя элемента системы отождествляется с требованием сохранения полезной информации при наличии помех. Становится очевидным смыкание с теорией помехоустойчивого кодирования: способы введения избыточности в сигнал естественно обобщаются на введение избыточности в структуру системы. Первые результаты на этом пути уже получены ([75, 76, 72, 73] и др.), однако общая теория надежности еще далеко не сформировалась. Следует также указать, что далеко не всякая система допускает столь прямое информационное описание ее структуры.

§ 4. ДРУГИЕ ПАРАМЕТРЫ ИНФОРМАЦИОННЫХ СИСТЕМ

В некоторых случаях параметры, рассмотренные в предыдущих параграфах данной главы, недостаточно полно характеризуют изучаемую систему. В связи с этим вводятся дополнительные параметры информационных систем.

1. Добротность системы. Как уже отмечалось, важнейшими показателями работы информационной системы являются эффективность и помехоустойчивость. Какому бы из этих качеств ни отдавалось предпочтение, лучшей системой будет та, которая при заданной эффективности обладает наи-

большей помехоустойчивостью (или наоборот, при заданной помехоустойчивости обладает наибольшей эффективностью). Поэтому целесообразно [74] ввести некоторый обобщающий параметр качества системы — добротность. Если помехоустойчивость исчисляется параметром S, а эффективность — параметром M, то добротность разумно определить как

$$Q=S\cdot M. \tag{4.1}$$

Наиболее совершенной системой следует считать систему, обладающую при заданных условиях наибольшей добротностью.

2. Отношение сигнала к шуму. Описание относительных достоинств одних систем перед другими требует конкретизации условий, в которых системы сравниваются. Может, например, оказаться, что из двух систем одна предпочтительна при работе в условиях сильных помех, другая работает лучше первой при слабых шумах. Для конкретизации соотношения полезного сигнала и помех определим отношение сигнала к шуму

$$a=\frac{v}{\sigma}, \tag{4.2}$$

где v — амплитуда полезного сигнала, σ — среднеквадратичное отклонение шума. Иногда целесообразно пользоваться «отношением сигнала к шуму по мощности»

$$b=a^2=\frac{v^2}{\sigma^2}=\frac{P_S}{P_N}, \tag{4.3}$$

где P_S и P_N — соответственно мощности сигнала и шума. Если помеха не является шумом в собственном смысле этого слова (например, перекрестные помехи в многоканальных линиях связи; помехи от импульсов считывания в ферритно-матричных запоминающих устройствах и т. п.), то под P_N следует понимать мощность мешающих сигналов.

В ряде случаев удобно ввести параметр

$$h=\log b=\log\frac{P_S}{P_N}, \tag{4.4}$$

который будем называть превышением сигнала над помехой, или просто превышением [79].

3. Вероятности трансформации символов. Для дискретных систем условия работы удобнее формулировать феноменологически, указывая вероятности подмены

одних символов другими. Свойства канала определятся полностью, если вместе с вероятностями $\{p_i\}$ отдельных символов задать стохастическую матрицу трансформации символов $\{p_{i\kappa}\}$, где i и κ — номера символов, $i, \kappa = 1, 2, \ldots, m$, m — число символов в алфавите, $p_{i\kappa}$ — вероятность перехода i-го символа в κ-й при воздействии шумов. Иногда достаточно задать лишь среднюю вероятность трансформации символа,

$$p_{\text{тр}} = \sum_{i=1}^{m} p_i (1 - p_{ii}). \tag{4.5}$$

4. Задержка передачи. Задержкой передачи называется [71] интервал времени между началом передачи, т. е. моментом поступления сигнала на первое кодирующее устройство (см. рис. 22), и моментом окончания приема, т. е. моментом прекращения сигнала на выходе второго декодирующего устройства. Задержка передачи складывается из нескольких интервалов:

а) времени кодирования τ_1, в течение которого второе кодирующее устройство вводит необходимую избыточность для передачи сигнала по каналу с шумами;

б) времени передачи τ_2 (длительность передаваемого сигнала);

в) времени распространения τ_p;

г) времени декодирования τ_3 принятого сигнала первым декодирующим устройством.

Таким образом,

$$\tau = \tau_1 + \tau_2 + \tau_p + \tau_3. \tag{4.6}$$

ГЛАВА IX

ДИСКРЕТНЫЕ СИСТЕМЫ БЕЗ ШУМОВ

§ 1. ПРОБЛЕМА ОПТИМАЛЬНОГО ПРЕДСТАВЛЕНИЯ ИНФОРМАЦИИ В ДИСКРЕТНЫХ СИСТЕМАХ БЕЗ ШУМОВ

Если в дискретной информационной системе (т. е. системе, информационные символы которой образуют дискретное множество) отсутствуют шумы, то, как бы ни кодировались сигналы, потерь информации не будет. Однако это не означает, что при этом не возникает никаких проблем. В большинстве случаев имеется заинтересованность не только в безошибочном представлении информации, но и в экономном ее представлении.

Если, например, производится запись нужной информации в запоминающее устройство, то важно, чтобы занималось минимально необходимо число ячеек; при передаче информации желательно занимать канал связи на максимально короткий срок и т. п. Здесь легко узнать требование наибольшей эффективности системы (см. § 3 предыдущей главы); удовлетворение этого требования не является тривиальным делом даже в отсутствии шумов.

Рассмотрим проблему представления информации в дискретной системе без шумов в следующей постановке. Пусть алфавит системы состоит из конечного числа m символов. Пусть, далее, требуется представить в этом алфавите любой сигнал из множества возможных сигналов $\{u_\kappa\}$, $\kappa = 1, 2, \ldots, M$; вероятности $\{P_\kappa\}$ каждого из сигналов известны. Обычно $m < M$, поэтому каждому из возможных входных сигналов ставится в соответствие некоторая последовательность символов алфавита, называемая кодовым словом. Кодовое слово должно находиться во взаимно однозначном соответствии с кодируемым сигналом. А так как возможна последовательная передача кодовых слов, то, кроме очевидного

требования о недопустимости одинаковых кодовых слов для разных сигналов, накладывается требование, чтобы ни одно кодовое слово нельзя было получить из другого, более короткого, путем добавления дополнительных символов. Например, кодовые слова 001 и 0010 не могут использоваться в одном множестве. Это необходимо для однозначного отделения кодовых слов друг от друга при их последовательной передаче. Можно потребовать, чтобы различение слов происходило с помощью пространственного или временного интервала (как при письме или телеграфии), но это равносильно введению в алфавит системы еще одного дополнительного символа специального назначения, что не способствует повышению эффективности. Можно, далее, предложить, чтобы все кодовые слова были одинаковой длины, тогда их разделение осуществлялось бы простым подсчетом числа символов; однако такое кодирование (называемое р а в н о м е р н ы м) также в общем случае не оптимально, как будет ясно из дальнейшего.

При любом представлении информации с каждым из M сигналов $\{u_к\}$ сопоставляется некоторое кодовое слово; обозначим через $n_к$ длину кодового слова, соответствующего сигналу $u_к$. Обозначим через L среднюю длину кодового слова при выбранном способе кодирования,

$$L = \sum_{к=1}^{M} n_к \cdot P(u_к). \qquad (1.1)$$

Ясно, что как $n_к$, так и L зависят от того, каким именно способом строятся кодовые слова и как они сопоставляются с входными сигналами $u_к$. Возникает ряд вопросов принципиального значения:

В каких пределах находится минимальная средняя длина L кодовых слов?

Каковы при этом индивидуальные длины $n_к$ кодовых слов?

В каком соотношении находятся величины L и $n_к$ с вероятностными характеристиками ансамбля возможных сигналов $\{u_к\}$?

Ответить на эти вопросы можно, исходя из общих положений теории информации, но для того, чтобы результаты носили точный характер, необходимо строгое доказательство соответствующих утверждений. В данном параграфе мы рассмотрим эвристическое решение проблемы эффективного кодирования при отсутствии шумов; количественные доказательства будут даны в следующем параграфе.

Итак, требуется наиболее эффективным образом представить в алфавите из m символов информацию, содержащуюся в ансамбле M сигналов с заданными вероятностями $P(u_к)$.

Индивидуальные количества информации равны соответственно $-\log P(u_к)$, а среднее количество информации на сигнал равно энтропии ансамбля сигналов, $H(u)=-\sum_{к=1}^{M} P(u_к)\log P(u_к)$.

Очевидно, что кодирование будет тем более эффективным, чем большее количество информации будет приходиться на каждый символ в кодовом слове. Так как в алфавите всего m символов, то максимальное количество информации на символ равно $\log m$. Среднее количество информации на один входной сигнал $u_к$, а следовательно, и на одно кодовое слово, равно $H(u)$. Так как информация не должна теряться при кодировании, то минимальная средняя длина кодового слова не может быть меньше, чем частное от деления представляемого среднего количества информации на максимальное количество информации на символ:

$$L \geqslant \frac{H(u)}{\log m}. \tag{1.2}$$

Так мы из общих соображений получаем нижнюю грань L_{min}. Аналогично рассуждая, легко придти к выводу, что для оптимального (в смысле эффективности) кодирования длина индивидуального кодового слова должна приближаться к частному от деления индивидуального количества представляемой информации на $\log m$; длина кодового слова не должна быть короче этого частного:

$$n_к \geqslant \frac{i(u_к)}{\log m} = \frac{-\log P(u_к)}{\log m} = z. \tag{1.3}$$

Легко видеть, что, усредняя (1. 3), мы получим (1. 2).

Общие положения теории информации позволяют найти и верхнюю грань минимально необходимой средней длины кодового слова L_{min}. Действительно, если z в (1. 3) не является целочисленным, то равенство в соотношении (1.3), а следовательно, и в (1.2), не может быть достигнуто. Однако ясно, что минимальная избыточность будет достигаться, если брать $n_к=(z)$, где (z) означает целое число, ближайшее к z сверху. Отсюда

$$\frac{i(u_к)}{\log m} \leqslant n_к \leqslant \frac{i(u_к)}{\log m} + 1, \tag{1.4}$$

а усредняя (1.4), получаем

$$\frac{H(u)}{\log m} \leqslant L_{min} \leqslant \frac{H(u)}{\log m} + 1. \tag{1.5}$$

Соотношения (1.4) и (1.5) полностью отвечают на вопросы о принципиальных возможностях оптимального кодирования при отсутствии шума и являются выражением основной теоремы для дискретных систем без шума. В следующем параграфе справедливость этой теоремы будет доказана строго; тот факт, что здесь она выглядит как следствие общих положений теории информации, указывает на эвристическую ценность теории.

§ 2. ФУНДАМЕНТАЛЬНАЯ ТЕОРЕМА О КОДИРОВАНИИ ПРИ ОТСУТСТВИИ ШУМА

Дадим строгое обоснование теореме о кодировании для бесшумных систем, рассмотренной в предыдущем параграфе.

Начнем с условий существования множества M кодовых слов, построенных из символов алфавита в m букв, причем ни одно кодовое слово не является началом другого, более длинного (необходимость этого была обсуждена в § 1).

Теорема (Л. С. Крафта [81]). Пусть $n_1, n_2,\ldots,n_\kappa,\ldots,n_M$ — последовательность M произвольных целых чисел. Необходимым и достаточным условием существования множества кодовых слов, удовлетворяющих вышеуказанным условиям и обладающих длинами, равными заданным числам $\{n_\kappa\}$, является выполнение неравенства

$$\sum_{\kappa=1}^{M} m^{-n_\kappa} \leqslant 1. \tag{2.1}$$

Доказательство [81]. Пусть l_j — число кодовых слов длины j рассматриваемого множества, если оно существует. Очевидно, что

$$\begin{aligned} l_1 &\leqslant m, \\ l_2 &\leqslant (m - l_1)\cdot m = m^2 - l_1 m, \\ l_3 &\leqslant [(m - l_1)\cdot m - l_2]\, m = m^3 - l_1 m^2 - l_2 m, \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ l_n &\leqslant m^n - l_1 m^{n-1} - l_2 m^{n-2} - \cdots - l_{n-1} m. \end{aligned} \tag{2.2}$$

Эти условия являются и необходимыми, и достаточными. Деля последнее неравенство на m^n, получаем:

$$\sum_{j=1}^{n} l_j m^{-j} \leqslant 1. \tag{2.3}$$

Так как j — длина кодовых слов рассматриваемого множества (по предположению существующего), то

$$\sum_{j=1}^{n} l_j m^{-j} = \sum_{n_\kappa \leqslant n} m^{-n_\kappa}, \tag{2.4}$$

причем число членов в правой части (2.4) равно числу кодовых слов с длиной, не превосходящей n, т. е. равно $\sum_{j=1}^{n} l_j$. Возьмем в качестве n длину N наиболее длинного кодового слова; тогда $n_\kappa \leqslant N$ при всех κ, и

$$\sum_{n_\kappa \leqslant N} m^{-n_\kappa} = \sum_{\kappa=1}^{M} m^{-n_\kappa}, \tag{2.5}$$

откуда с учетом (2.3) и (2.4) следует необходимость условия (2.1).

Для доказательства достаточности условия (2.1) нужно показать, что всегда можно найти множество кодовых слов рассматриваемого типа, таких, что $\sum_{j=1}^{N} l_j m^{-j} \leqslant 1$. Но это неравенство через (2.3) сводится к системе неравенств (2.2), которая и выражает возможность данного построения.

Перейдем теперь к доказательству основной теоремы. Так как для любых двух распределений $P(u_\kappa)$ и $Q(u_\kappa)$ выполняется неравенство*)

$$\sum_{\kappa} P(u_\kappa) \log \frac{P(u_\kappa)}{Q(u_\kappa)} \geqslant 0, \tag{2.6}$$

то, взяв в качестве $Q(u_\kappa)$

$$Q(u_\kappa) = \frac{m^{-n_\kappa}}{\sum_{r=1}^{M} m^{-n_r}}, \tag{2.7}$$

получаем:

$$H(u) = -\sum_{\kappa=1}^{M} P(u_\kappa) \log P(u_\kappa) \leqslant -\sum_{\kappa=1}^{M} P(u_\kappa) \log \frac{m^{-n_\kappa}}{\sum_{r=1}^{M} m^{-n_r}} =$$

$$= \log \sum_{r=1}^{M} m^{-n_r} + \sum_{\kappa=1}^{M} P(u_\kappa)\, n_\kappa \log m. \tag{2.8}$$

*) Неравенство (2.6) является дискретным аналогом соотношения (5.4), доказанного в главе IV.

Из (2.1) следует, что $\log \sum_{r=1}^{M} m^{-n_r} \leqslant \log 1 = 0$, следовательно,

$$H(u) \leqslant \sum_{\kappa=1}^{M} P(u_\kappa) \cdot n_\kappa \cdot \log m, \tag{2.9}$$

откуда следует

$$L = \sum_{\kappa=1}^{M} P(u_\kappa) \cdot n_\kappa \geqslant \frac{H(u)}{\log m}. \tag{2.10}$$

Знак равенства в (2.10) достигается, если $P(u_\kappa) = m^{-n_\kappa}$, так как при этом $P(u_\kappa) = Q(u_\kappa)$ в (2.6) и $\sum_{\kappa=1}^{M} m^{-n_\kappa} = 1$. Если же вероятности $P(u_\kappa)$ не являются целочисленными степенями m, то достичь указанной нижней границы в общем случае невозможно. Однако можно показать, что при кодировании более крупными блоками (т. е. если отдельные кодовые слова ставить в соответствие не каждому из u_κ, а некоторой достаточно длинной последовательности таких сигналов) — можно сколь угодно приблизиться к нижней грани.

Пусть кодирование осуществляется указанными крупными блоками; кодируются последовательности сразу в K входных сигналов u_κ. Таких различных блоков будет, очевидно, M^K. Предположив, что сигналы u_κ независимы между собой, имеем, что энтропия множества указанных блоков равна $K \cdot H(u)$. Определим теперь целые числа n_κ неравенствами

$$-\frac{\log p(y_\kappa)}{\log m} \leqslant n_\kappa \leqslant -\frac{\log p(y_\kappa)}{\log m} + 1, \tag{2.11}$$

где $p(y_\kappa)$ – вероятность осуществления κ-го из M^K блоков. При таком задании n_κ

$$\sum_{\kappa=1}^{M^K} m^{-n_\kappa} \leqslant \sum_{\kappa} p(y_\kappa) = 1,$$

и, по теореме Крафта, можно блокам y_κ сопоставить кодовые слова длиной n_κ. Средняя длина таких кодовых слов L_K удовлетворяет неравенствам

$$\frac{H(y)}{\log m} \leqslant L_K \leqslant \frac{H(y)}{\log m} + 1,$$

или

$$\frac{KH(u)}{\log m} \leqslant L_K \leqslant \frac{KH(u)}{\log m} + 1. \qquad (2.12)$$

Деля (2.12) на K, получаем:

$$\frac{H(u)}{\log m} \leqslant \frac{L_K}{K} \leqslant \frac{H(u)}{\log m} + \frac{1}{K}. \qquad (2.13)$$

При $K \to \infty$ $\frac{L_K}{K} \to \frac{H(u)}{\log m}$. Таким образом доказана

Фундаментальная теорема о кодировании при отсутствии шума (К. Шэннон):

При кодировании множества сигналов с энтропией $H(u)$ в алфавите, насчитывающем m символов, при условии отсутствия шумов, средняя длина кодового слова не может быть меньше чем $H(u)/\log m$. Если вероятности сигналов не являются целочисленными отрицательными степенями числа m, то точное достижение указанной нижней границы невозможно; но при кодировании достаточно длинными блоками к этой границе можно сколь угодно приблизиться.

§ 3. О СВОЙСТВАХ ОПТИМАЛЬНЫХ И БЛИЗКИХ К ОПТИМАЛЬНЫМ КОДОВ

Фундаментальная теорема о кодировании при отсутствии шума является теоремой существования: она доказывает, что оптимальные коды существуют, но не дает никаких указаний на то, как построить такой код. Хотя вопросы построения кодов являются, по существу, прикладными вопросами, обсудим здесь — в целях дополнительного разъяснения фундаментальной теоремы — некоторые из этих вопросов.

1. При рассмотрении доказательства фундаментальной теоремы может сложиться впечатление, что предположение о специальном характере кодовых слов (ни одно из них не должно быть расширением другого, более короткого) слишком сильно ограничивает класс рассматриваемых кодов и лишает общности самую теорему. А. Фейнстейн [81] привел даже пример, который, на первый взгляд, подтверждает это впечатление. Пусть множество $\{u_\kappa\}$ состоит из трех элементов с вероятностями $P(u_1) = P(u_2) = 2P(u_3) = \frac{2}{5}$. Построим двоичный код из символов a_1 и a_2, отказавшись от рассма-

триваемого ограничения: $u_1 \to a_1$, $u_2 \to a_2$, $u_3 \to a_1 a_2$. Легко убедиться, что средняя длина кодового слова, подсчитываемая по формуле

$$L = 1 \cdot P(u_1) + 1 \cdot P(u_2) + 2P(u_3) = \frac{6}{5}, \qquad (3.1)$$

тогда как энтропия $H(u)$ больше (!) этой величины:

$$H(u) = \log_2 5 - \frac{4}{5} > \frac{6}{5}. \qquad (3.2)$$

Однако в этом примере замаскирован тот факт, что при выбранном коде невозможна многократная безошибочная передача. Действительно, приняв последовательность $a_1\ a_2\ a_1$, мы можем ее декодировать либо как $u_1\ u_2\ u_1$, либо как $u_3\ u_1$. Чтобы информация не терялась, необходимо введение дополнительного разделяющего символа, и легко убедиться, что как бы этот символ ни вводился (перед каждым ли кодовым словом, или только перед двумя из них), средняя длина кодового слова не может быть меньше энтропии кодируемого ансамбля сигналов.

Общность фундаментальной теоремы не должна подвергаться сомнению, конечно, не только потому, что данный пример, направленный против этой общности, оказался несостоятельным. Шэннон [7] дал доказательство теоремы, не опирающееся на специальные предположения о свойствах кодовых слов и исходящее из свойства E (рассмотренного в § 9 гл. V) эргодических процессов. Подробный вариант этого доказательства имеется у А. М. Яглома и И. М. Яглома ([63], стр. 178—184).

2. Хотя фундаментальная теорема не указывает способа построения оптимального кода (или кодов, если оптимальный код не единственный), из нее вытекает, какими свойствами обладает оптимальный код. Для обеспечения минимальности средней длины кодового слова избыточность в кодовых словах должна быть сведена к минимуму (желательно — к нулю). Это означает, что оптимальный код (с конечным числом возможных символов) должен состоять из кодовых слов, каждый символ в которых статистически не зависит от других симзолов, и все символы равновероятны. Если требования независимости и равновероятности символов почему-либо невыполнимы, то, чем лучше они выполняются, тем ближе к оптимальному код.

Эти общие соображения приводят к идее построения оптимального или близкого к нему кода, получившего название кода Шэннона-Фэно. Процедура построения кода

Шэннона-Фэно направлена на максимальное удовлетворение требований равновероятности и независимости символов. Рассмотрим эту процедуру на примерах построения кода в двоичном алфавите. Множество кодируемых сигналов разбивается на две группы так, чтобы вероятности принадлежать к каждой из этих групп были возможно более близки друг к другу. Если кодируемый сигнал относится к первой группе, то в качестве первого символа кодового слова берется О, а для кодовых слов второй группы используется 1. Затем каждая из групп разбивается на две по возможности равновероятных подгруппы, и символ О или 1 берется вторым символом кодового слова в зависимости от того, к какой из подгрупп относится кодируемый сигнал. Такое разбиение на подгруппы производится до тех пор, пока в подгруппе не останется один-единственный из кодируемых сигналов. Для удобства разбиения кодируемые сигналы обычно располагаются в порядке убывания их вероятностей.

Рассмотрим сначала примеры, в которых код Шэннона-Фэно позволяет достичь теоретического предела эффективности.

Пример 1. Пусть все кодируемые сигналы равновероятны и вероятности их являются целочисленными степенями основания кода (в данном случае — 2); сигналов в ансамбле 8. Тогда производя разбиения так, как это показано на табл. 1,

Таблица 1

Кодируемые сигналы	Вероятности	Разбиения	Кодовые слова
u_1	0,125		000
u_2	0,125		001
u_3	0,125	Второе	010
u_4	0,125		011
u_5	0,125	Первое	100
u_6	0,125	Третье	101
u_7	0,125		110
u_8	0,125		111

получаем оптимальный код. Действительно, средняя длина кодового слова равна энтропии ансамбля кодируемых сигналов: $\log_2 8 = 3$. По самому построению кодовых слов каждый символ в них является независимым и равновероятным.

Пример 2. Пусть кодируемые сигналы неравновероятны, но их вероятности по-прежнему являются целочисленными

степенями двойки. Расположив сигналы по убывающим вероятностям и произведя разбиения в соответствии с процедурой Шэннона-Фэно, снова получаем оптимальный код (см. таблицу 2). Действительно, средняя длина кодового слова в точности равна энтропии ансамбля сигналов в битах. Символы кодовых слов равновероятны и независимы; ни одно кодовое слово не является расширением более короткого; число символов в каждом кодовом слове равно индивидуальному количеству информации, несомому соответствующим сигналом.

Таблица 2

Кодируемые сигналы	Вероятности	Разбиения	Кодовые слова
u_1	0,25		00
u_2	0,25	Первое	01
u_3	0,125		100
u_4	0,125	Второе	101
u_5	0,0625		1100
u_6	0,0625	Третье	1101
u_7	0,0625	Четвертое	1110
u_8	0,0625		1111

Как следует из условий фундаментальной теоремы, при невыполнении условия $P(u_к) = m^{-n_к}$, где $n_к$ — целое для всех $к$, получение оптимального кода невозможно; однако путем укрупнения блоков можно сколь угодно приблизиться к оптимальному коду. Продемонстрируем это примером.

Пример 3. Пусть необходимо кодировать в двоичном коде сообщение, состоящее из двух неравновероятных букв, u_1 и u_2; $P(u_1) = 0{,}89$, $P(u_2) = 0{,}11$. Проследим, как приближается код Шэннона-Фэно к оптимальному при укрупнении кодируемых блоков.

а) Кодирование по одной букве

Буквы	Вероятности	Разбиения	Кодовые слова
u_1	0,89	Первое	0
u_2	0,11		1

В полученном коде для передачи каждой буквы требуется один знак, тогда как энтропия равна $0{,}4999 \approx 0{,}5$ бит; средняя длина кодового слова вдвое больше теоретически необходимой!

б) Кодирование по две буквы

Кодируемые блоки	Вероятности	Разбиения	Кодовые слова
u_1u_1	0,792	Первое	0
u_1u_2	0,098	Второе	10
u_2u_1	0,098	Третье	110
u_2u_2	0,012		111

Средняя длина кодового слова 1,318, следовательно, на одну букву в среднем приходится 0,66 символов — на 32% больше теоретического минимума (0,5 символов).

в) Кодирование по три буквы

Кодируемые блоки	Вероятности	Разбиения	Кодовые слова
$u_1u_1u_1$	0,705	Первое	0
$u_1u_1u_2$	0,087		100
$u_1u_2u_1$	0,087	Второе	101
$u_2u_1u_1$	0,087	Третье	110
$u_1u_2u_2$	0,011		11100
$u_2u_1u_2$	0,011	Четвертое	11101
$u_2u_2u_1$	0,011	Пятое	11110
$u_2u_2u_2$	0,001		11111

Средняя длина кодового слова 1,658; на букву в среднем приходится 0,552 символа — уже лишь на 10% больше возможного минимума. При кодировании по 4 буквы получим в среднем 0,52 символа на букву, т. е. всего на 4% больше $H(u)$, и т. д.

Важно еще раз подчеркнуть, что уменьшение среднего числа символов на букву достигается благодаря приписыванию наиболее коротких кодовых слов наиболее вероятным кодируемым сигналам. При кодировании очень длинных блоков вступает в силу свойство эргодических случайных процессов; наиболее вероятные блоки образуют очень малую долю всех возможных блоков, и логарифм их числа очень близок к энтропии всего ансамбля. Все эти высоковероятные блоки почти равновероятны, поэтому длина кодовых слов определяется числом разбиений именно в этой группе, и для этой группы блоков код будет равномерным, причем среднее число символов на букву будет весьма близко к энтропии $H(u)$. Остальные блоки хотя и будут давать длинные кодовые слова, но их влияние на среднюю длину будет ничтожно, так как суммарная вероятность таких слов весьма мала.

ГЛАВА X

ДИСКРЕТНЫЕ СИСТЕМЫ С ШУМАМИ

§ 1. ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ ПРИ НАЛИЧИИ ШУМА

Наличие шумов в информационной системе приводит к тому, что соответствие между входным и выходным сигналами системы перестает быть однозначным. Сигнал, отправленный по системе связи, в которой присутствуют шумы, не обязательно совпадает с принятым сигналом; сигнал, запрошенный из запоминающего устройства с шумами, может отличаться от того, который был заслан в это устройство, и т. д. В дискретных системах влияние шумов проявляется в случайной подмене одних символов другими. Однако, несмотря на такие случайные искажения, соответствие обычно не разрушается полностью, что и обеспечивает возможность функционирования информационных систем даже при наличии шумов.

Изучение возможностей работы в условиях помех имеет тем большее значение, что на практике помехи присутствуют в любых системах, а в ряде случаев функционирование при наличии сильных шумов является нормальным режимом работы системы. Именно при рассмотрении вопросов передачи информации при наличии шумов были получены наиболее важные результаты теории информации, которые вслед за А. Н. Колмогоровым [20] можно назвать крупными научными открытиями.

Рассмотрим несколько более подробно проблемы, возникающие при передаче информации при наличии шума по дискретной системе связи. На вход второго кодирующего устройства (см. рис. 25) поступает одна из N_0 последовательностей n_0 символов, которую необходимо передать по каналу с шумами, по возможности, не совершив ошибки. Идея состоит в том, чтобы внести достаточную избыточность, обеспечив тем

самым необходимую помехоустойчивость. Внесение избыточности связано с увеличением длительности сигнала на некоторое число n_c символов. Число возможных последовательностей сразу резко увеличивается, но первое кодирующее устройство работает только с N_0 из них, отмеченными на рис. 25 черными кружочками, остальные сигналы не использу-

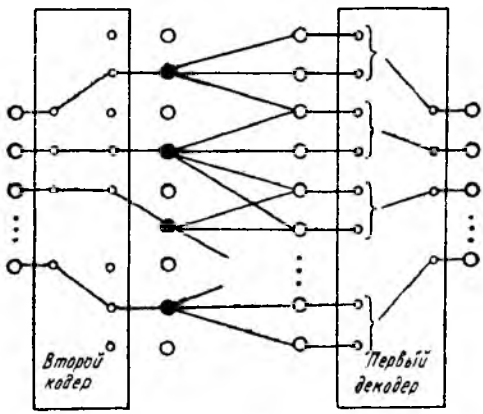

Рис. 25.

ются. Затем, при передаче последовательности в n_0+n_c символов по каналу с шумами, некоторые из символов искажаются (т. е. подменяются другими символами); а так как это происходит случайным образом, то одной и той же отправленной последовательности может отвечать несколько различных принятых последовательностей. Обратно, каждой из принятых последовательностей соответствует некоторое подмножество множества возможных последовательностей на входе канала. В таких условиях требуется принять однозначное решение о том, какой же сигнал был отправлен. Это достигается тем, что все множество принимаемых сигналов разбивается на N подмножеств, сопоставляемых с N возможными отправляемыми сигналами; если, например, получен сигнал из i-й группы, то считается, что был послан i-й сигнал, который и выдается «в чистом виде» получателю. Проблема состоит в том, чтобы выяснить, возможны ли такое размещение N передаваемых последовательностей среди возможных входных канальных последовательностей и такое соответствующее разбиение на подгруппы множества принимаемых последовательностей, чтобы вероятность ошибки была не больше заданной сколь угодно малой величины. Другая проблема заключается в том, чтобы указать, существует ли некоторая минимальная необ-

ходимая избыточность, или при уменьшении вероятности ошибки соответственно должна увеличиваться и избыточность.

Результаты исследований, сформулированные К. Шэнноном в виде теорем, показали, что во-первых, как бы ни были сильны шумы, можно создать условия, при которых возможна передача информации при сколь угодно малой вероятности ошибки (Первая и Обратная теоремы); и, во-вторых, при этом не потребуется прогрессивно понижать среднюю скорость передачи информации при повышении требований к малости вероятности ошибки (Вторая теорема). Правда, условия «безошибочной» передачи возникают лишь при сложном кодировании очень длинных входных сигналов, однако, принципиальное значение указанных теорем от этого не уменьшается.

§ 2. ПЕРВАЯ ТЕОРЕМА ШЭННОНА О КОДИРОВАНИИ В ПРИСУТСТВИИ ШУМОВ

Прежде чем переходить к формулировке и доказательству фундаментальной теоремы для дискретных каналов с шумами, уточним еще раз условия работы системы связи, о которых будет идти речь.

Как будет видно в дальнейшем, теорема опирается на определенное предположение о соотношении между скоростью создания информации H и пропускной способностью C канала. Напомним, что имеются в виду стационарные системы связи, так что возможность записи быстрых процессов с последующей замедленной передачей не должна казаться противоречащей условию теоремы ($H \leqslant C$): при таком методе передачи в стационарном случае потребовалась бы неограниченная память кодирующих устройств и, кроме того, неограниченная задержка. Таким образом, предполагается, что потоки информации стационарны и что емкости запоминающих устройств кодирующей и декодирующей систем ограничены. Стационарность сигналов на входе системы связи позволяет рассматривать кодирование блоками произвольной длины, а это, в свою очередь, обеспечивает условие применимости свойства E случайных процессов, на основе которого и строится доказательство теоремы.

Далее, для простоты будем считать, что между последовательными ошибками не существует корреляции, вероятность искажения очередного символа не зависит от того, был ли искажен предыдущий символ. Такие каналы принято называть «каналами без памяти».

Теорема. Если скорость создания информации H источником на входе шумящего канала без памяти с пропускной способностью

C меньше пропускной способности, то существует такой код, при котором вероятность ошибок на приемном конце сколь угодно мала.

Идея доказательства, предложенная Шэнноном [7] и оставшаяся в основе более поздних, более подробных и строгих доказательств ([43, 48, 67, 81]), состоит не в том, чтобы указать, как именно нужно кодировать сигналы для обеспечения малости вероятности ошибки, а в том, чтобы показать, что такой код вообще существует. Для этого определяется некоторый класс кодов и показывается, что средняя (по множеству указанных кодов) вероятность ошибки может быть сделана меньше сколь угодно малой величины ε. В совокупности вероятностей ошибки должна существовать по крайней мере одна вероятность, меньшая средней, чем и доказывается существование искомого кода*).

Пусть $H(x)$ и $H(x|y)$ — априорная и апостериорная энтропии на символ (со стороны приемного конца) для системы, реализующей пропускную способность C канала. В силу свойства E (см. § 9 гл. V) при достаточно большой длительности (n символов) передачи все возможные последовательности любого ансамбля распадаются на высоковероятную и маловероятную группы; при этом о количестве сигналов в соответствующих группах можно сделать следующие утверждения:

а) Группа высоковероятных передаваемых сигналов содержит около $2^{nH(x)}$ последовательностей**).

б) Группа высоковероятных принимаемых сигналов содержит около $2^{nH(y)}$ последовательностей.

в) Каждый высоковероятный принимаемый сигнал мог (с приблизительно одинаковыми вероятностями) произойти от примерно $2^{nH(x|y)}$ передаваемых сигналов высоковероятной группы.

г) Каждому отправляемому сигналу из высоковероятной группы может (с приблизительно одинаковыми вероятностями) соответствовать примерно $2^{nH(y|x)}$ принимаемых высоковероятных сигналов.

В силу свойства E энтропии дискретных процессов, при увеличении n все соответствующие ε и δ будут стремиться к нулю.

Пусть теперь по тому же каналу передается информация со скоростью на входе, равной $H < C$. При этом число вы-

*) При доказательстве теоремы мы будем следовать К. Шэннону [7]; читатель, интересующийся более строгими доказательствами, найдет их в работах [43, 48, 67, 81]; строгие доказательства простых частных случаев даны в [82].

**) Предполагается, что энтропии исчисляются в битах.

соковероятных отправляемых сигналов длиной в n символов будет равно $2^{nH} < 2^{nH(x)}$. Как уже отмечалось в предыдущем параграфе, проблема выбора определенного кода состоит в указании того, какие именно из $2^{nH(x)}$ возможных последовательностей выбираются в качестве 2^{nH} разрешенных к отправке, и как разбиваются на 2^{nH} подгрупп $2^{nH(y)}$ выходных последовательностей. Рассмотрим класс всевозможных кодов, которые получатся, если 2^{nH} разрешенных последовательностей размещать случайным образом среди $2^{nH(x)}$ возможных сигналов высоковероятной группы; найдем среднее значение вероятности ошибки для этих кодов.

Пусть принят некоторый сигнал y_k. Вероятность ошибки при этом равна вероятности того, что данный сигнал может происходить более чем от одного из 2^{nH} разрешенных сигналов. Поскольку код получается случайным (равновероятным) выбором 2^{nH} последовательностей из $2^{nH(x)}$, то вероятность того, что заданный сигнал на входе канала попадет в число разрешенных, равна

$$\frac{2^{nH}}{2^{nH(x)}} = 2^{n(H-H(x))}. \tag{2.1}$$

Принятому сигналу y_k соответствует $2^{nH(x|y)}$ возможно отправленных сигналов. Отсюда средняя вероятность того, что ни один из $2^{nH(x|y)}$ сигналов (кроме одного действительно отправленного) не является разрешенным, равна (пренебрегаем единицей по сравнению с $nH(x|y)$)

$$P = (1 - 2^{n(H-H(x))})^{2^{nH(x|y)}}. \tag{2.2}$$

Это есть средняя вероятность безошибочного приема. Далее, так как $H < C = H(x) - H(x|y)$, то

$$H \quad H(x) = -H(x|y) - \tau_1, \tag{2.3}$$

где $\tau_1 > 0$. Подставляя (2.3) в (2.2), получаем

$$P = (1 - 2^{-nH(x|y) - n\tau_1})^{2^{nH(x|y)}}. \tag{2.4}$$

Легко показать, что $\lim\limits_{n\to\infty} P = 1$. Действительно,

$$\lim_{n\to\infty} \log P = \lim_{n\to\infty} [2^{nH(x|y)} \cdot \log (1 - 2^{-n[H(x|y)+\tau_1]})] =$$

$$= \lim_{n\to\infty} \frac{\log (1 - 2^{-n[H(x|y)+\tau_1]})}{2^{-nH(x|y)}}.$$

Применяя правило Лопиталя, имеем

$$\lim_{n\to\infty} \log P = \lim_{n\to\infty} \left(- \frac{H(x|y)+\gamma_1}{H(x|y)} \cdot \frac{2^{-n\gamma_1}}{1-2^{-n[H(x|y)+\gamma_1]}} \right) = 0, \quad (2.5)$$

откуда и следует, что $\lim_{n\to\infty} P = 1$, т. е. что при случайном кодировании достаточно длинными блоками средняя вероятность ошибки может быть сделана сколь угодно малой. Утверждение о существовании по крайней мере одного кода, дающего вероятность ошибки меньше средней, завершает доказательство.

Как подчеркнул К. Шэннон, соображения, связанные с законом больших чисел, позволяют сделать утверждение более сильное, нежели утверждение о существовании по крайней мере одного кода. Если среднее значение множества положительных чисел отличается от нуля меньше чем на ε, то доля всех чисел, превышающих $\sqrt{\varepsilon}$, не больше $\sqrt{\varepsilon}$ от общего числа элементов множества. Так как ε может быть сделано (при $n \to \infty$) сколь угодно малым, то можно сказать, что почти любой код, выбранный наугад, будет близок к оптимальному при кодировании достаточно длинными блоками.

Отметим далее, что равенство (2.5) справедливо при любом, сколь угодно малом положительном γ_1. Это означает, что теорема допускает условие $H \leqslant C$. Это и придает особый смысл понятию пропускной способности: пропускная способность оказывается не просто максимально возможной скоростью передачи информации, но максимальной скоростью, при которой еще возможна передача со сколь угодно малой вероятностью ошибки.

§ 3. ВТОРАЯ ТЕОРЕМА ШЭННОНА

Для обеспечения достаточной помехоустойчивости приходится вводить в передаваемый сигнал избыточность, уменьшая тем самым скорость передачи информации. Вполне естественно опасение, что при усилении ограничений на малость вероятности ошибки необходимая избыточность будет возрастать, прогрессивно снижая скорость передачи информации, возможно, до нуля. Однако все сомнения снимаются Второй теоремой Шэннона о кодировании для каналов с шумами, которая может быть сформулирована следующим образом:

При условии $H \leqslant C$ среди кодов, обеспечивающих (согласно Первой теореме) сколь угодно малую вероятность ошибки, сущест-

вует код, при котором скорость передачи информации R сколь угодно близка к скорости создания информации H.

Скорость передачи информации (на символ) определяется как

$$R = H - H(x|y), \tag{3.1}$$

где $H(x|y)$ — апостериорная энтропия отправленного сигнала на символ, или рассеяние информации в канале. Доказательство теоремы начинается с утверждения о том, что минимальная необходимая избыточность на символ равна $H(x|y)$ добавочных символов. Далее мы должны показать, что код можно выбрать так, чтобы $H(x|y)$ была сколь угодно малой величиной.

Для этого нам понадобится неравенство Фэно [81], которое можно получить из общих положений теории информации. Чтобы снять неопределенность $H(x|y)$, оставшуюся после приема благодаря возможным ошибкам, необходимо дополнительное количество информации, равное $H(x|y)$ на каждый символ. Для того, чтобы точно установить, какой именно символ x был передан, если принят символ y, необходимо: во-первых, установить, была ли вообще совершена ошибка при передаче данного символа. Если вероятность ошибки равна $p(e)$, то количество информации, нужное для обнаружения ошибки, равно $H(p(e), 1-p(e))$.

Во-вторых, если обнаружилось, что ошибка совершена, то нужно установить, какой именно из $m-1$ остальных символов был действительно передан. Очевидно, (см. § 2 гл. V) количество информации, необходимое в этом случае, не будет превышать величины $\log(m-1)$, а так как это будет происходить с вероятностью $p(e)$, то среднее добавочное количество информации будет не более $p(e)\log(m-1)$.

Таким образом, имеем:

$$H(x|y) \leqslant -p(e)\log p(e) - (1-p(e))\log(1-p(e)) + \\ + p(e)\log(m-1). \tag{3.2}$$

Это неравенство, конечно, может быть отнесено не только к отдельным символам, но и к сигналам любой длины; при этом под $H^*(x|y)$ понимается апостериорная неопределенность отправленной последовательности, $p^*(e)$ — вероятность ошибочного отождествления сигнала, m нужно заменить на число N возможных сигналов (высоковероятной группы). В условиях первой теоремы можно записать следующую последовательность неравенств:

$$H^*(x|y) \leqslant -\varepsilon \log \varepsilon - (1-\varepsilon)\log(1-\varepsilon) + \varepsilon \log(N-1) <$$

$$< 1 + \varepsilon \log(N-1) < 1 + \varepsilon nC, \qquad (3.3)$$

где n — число символов в сигнале. Теперь видно, что рассеяние информации на символ (так как ε может быть сделано сколь угодно малым, а n сколь угодно большим) может быть сколь угодно малым:

$$H(x|y) = \frac{H^*(x|y)}{n} < \frac{1}{n} + \varepsilon C, \qquad (3.4)$$

а следовательно, скорость передачи R может быть сколь угодно близкой к скорости создания информации H, что и требовалось доказать.

4. ОБРАТНАЯ ТЕОРЕМА ШЭННОНА ДЛЯ КАНАЛОВ С ШУМАМИ

Обратная теорема указывает условия, которые возникают при передаче информации по каналу с шумами со скоростью, превышающей пропускную способность.

Теорема. Если скорость создания информации H больше пропускной способности канала C, то никакой код не может сделать вероятность ошибки сколь угодно малой. Минимальное рассеяние информации на символ, достижимое при $H > C$, равно $H - C$; никакой код не может обеспечить меньшего рассеяния информации.

Для доказательства первой части теоремы воспользуемся неравенством Фэно (3.2). Имеем:

$$H^*(x|y) \leqslant 1 + p^*(e)\log(N-1). \qquad (4.1)$$

Далее, в силу свойства E, при больших n

$$H^*(x) = \log N. \qquad (4.2)$$

Следовательно,

$$nR = H^*(x) - H^*(x|y) \geqslant \log N - p^*(e)\log(N-1) - 1. \qquad (4.3)$$

Но, по определению пропускной способности, $R \leqslant C$, следовательно

$$p^*(e) \geqslant \frac{\log N}{\log(N-1)} - \frac{1}{\log(N-1)} - \frac{nC}{\log(N-1)} \approx 1 - \frac{nC+1}{\log N}. \qquad (4.4)$$

Так как $N = 2^{nH}$, то

$$p^*(e) \geqslant 1 - \frac{nC+1}{nH}, \tag{4.5}$$

откуда видно, что $p^*(e)$ не может быть сколь угодно малой, как бы велико n ни было. Более того, можно показать, что $p^*(e)$ стремится к 1 при увеличении n. Действительно, при $H > C$ величина τ_1 в (2.3) становится отрицательной. При этом $\lim_{n\to\infty} \log P = -\infty$ (см. (2.5)), и вероятность безошибочного приема стремится к нулю.

Второе утверждение теоремы является прямым следствием определения пропускной способности как максимальной скорости передачи: так как

$$R = H - H(x|y) \leqslant C,$$

то

$$H(x|y) \geqslant C - H, \tag{4.6}$$

что и требовалось доказать.

§ 5. ОБСУЖДЕНИЕ ТЕОРЕМ О КОДИРОВАНИИ ДЛЯ КАНАЛОВ С ШУМАМИ

В связи с большим принципиальным значением теорем, рассмотренных в данной главе, необходимо сделать ряд замечаний.

1. Первая и вторая теоремы Шэннона являются теоремами, которые лишь указывают на существование кодов, обеспечивающих произвольную малость вероятности ошибок и не уменьшающих скорость передачи информации; при этом вопрос о том, как построить такие коды, остается в стороне. Нужно сказать, что до сих пор не найдено общего метода построения кодов, реализующих теоретические пределы для $p^*(e)$ и R. Существует, однако, большое количество работ, посвященных построению специальных помехоустойчивых кодов в различных важных частных случаях (см., например, сборник [84]). Хотя при этом широко привлекаются понятия и методы теории информации, данное научное направление носит прикладной характер; поэтому рассмотрение результатов, полученных в теории кодирования, здесь проводиться не будет*).

2. Обратная теорема Шэннона утверждает, что при $H > C$ безошибочная передача невозможна; при этом чем больше

*) Предполагается, что теория кодирования будет рассмотрена во II части данного курса, посвященной приложениям теории информации.

отношение H/C, тем больше остаточная неопределенность $H(x|y)$. Последняя связана с вероятностью ошибки при приеме. Естественно возникает вопрос о том, как связана минимальная вероятность ошибки, достигаемая при наилучшем кодировании, с отношением H/C. Кроме почти очевидного утверждения о монотонном возрастании вероятности ошибки при увеличении H/C желательно дать количественное решение этой задачи. Для бинарного канала решение получается просто; приведем его здесь.

Пусть источник выдает H двоичных символов в секунду, которые передаются по каналу с пропускной способностью $C < H$. Пусть после приема и декодирования вероятность ошибки на символ равна ε. Рассеяние информации при этом будет равно $H(x|y) = \varepsilon \log \varepsilon + (1-\varepsilon)\log(1-\varepsilon)$. Следовательно, количество информации на символ, передаваемое по каналу, равно $I = 1 + \varepsilon \log \varepsilon + (1-\varepsilon)\log(1-\varepsilon)$, а за единицу времени передается количество информации, равное $H[1 + \varepsilon \log \varepsilon + (1-\varepsilon)\log(1-\varepsilon)]$. Если $\varepsilon(\kappa)$ — минимальная вероятность ошибки, то реализуется максимальная скорость передачи, т. е.

$$H[1 + \varepsilon(\kappa)\log\varepsilon(\kappa) + (1-\varepsilon(\kappa))\log(1-\varepsilon(\kappa))] = C. \qquad (5.1)$$

Отсюда и определится минимальная достижимая вероятность ошибки при $\kappa = H/C > 1$:

$$\frac{1}{\kappa} = 1 + \varepsilon(\kappa)\log\varepsilon(\kappa) + (1-\varepsilon(\kappa))\log(1-\varepsilon(\kappa)). \qquad (5.2)$$

Решение этого уравнения представлено графически на рис. 26 ($\varepsilon(\kappa) = 0$ при $\kappa < 1$ согласно первой теореме). Легко видеть, что при $\kappa \to \infty$ $\varepsilon(\kappa) \to 0{,}5$, что означает, что доля передаваемой информации из всей поступающей на вход канала стремится к нулю при $\kappa \to \infty$; чем быстрее ведется передача, тем меньшее количество информации передается.

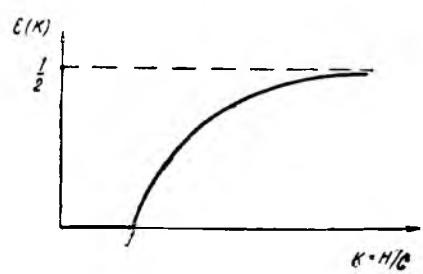

Рис. 26.

3. Все три теоремы о кодировании для каналов с шумами дают результаты, носящие асимптотический характер: утверждения теорем выполняются тем лучше, чем более длинными блоками осуществляется кодирование. Это говорит

о большой силе сложных методов кодирования; однако при практическом осуществлении этих методов возникли бы значительные затруднения (сложность кодирующих и декодирующих устройств, необходимость большой емкости запоминающих устройств, наличие большой задержки передачи и т. д.). Поэтому очень важно для практики дать теоретические пределы вероятности ошибки при кодировании блоками не очень большой длины n_0. Этому вопросу посвящены работы П. Элиаса ([86] и др.), К. Шэннона [85], А. Фейнстейна [81] и других ученых. Оказалось, что для минимальной вероятности ошибки, которую можно в принципе достичь при кодировании блоками длины n_0, могут быть указаны верхняя и нижняя грани:

$$\kappa_2 \cdot 2^{-n_0 E_2} \leqslant p^*(e) \leqslant \kappa_1 \cdot 2^{-n_0 E_1}. \qquad (5.3)$$

В этих неравенствах κ_1 и κ_2 слабо зависят от n_0; (как $\sqrt{n_0}$); E_1 и E_2 являются функциями C и H и не зависят от n_0. E_1 и E_2 положительны при $H < C$ и равны нулю при $H = C$. E_1 и E_2 уменьшаются при увеличении H, что вполне согласуется с интуитивными представлениями: вероятность ошибки при увеличении H возрастает. Легко видеть, что полученные оценки хорошо согласуются с Первой и Второй теоремами: при $n \to \infty$ $p^*(e) \to 0$.

4. При изложении теорем о кодировании для каналов с шумами имелись в виду каналы «без памяти», т. е. предполагалась статистическая независимость искажений последовательно поступающих символов. Хотя для ряда практических систем это предположение хорошо согласуется с действительностью, имеется целый ряд систем, для которых оно заведомо неверно. А. Фейнстейн [81] и А. Я. Хинчин [48] показали, что утверждения Первой и Второй теорем могут быть обобщены на случай каналов «с памятью» и полностью остаются в силе. Однако доказать для каналов с памятью Обратную теорему не удалось до сих пор. Исследование каналов с памятью требует более тонких методов, чем рассмотрение каналов без памяти. В частности, вводится две пропускных способности, стационарная C_s и эргодическая C_e, причем $C_s \geqslant C_e$. Соответственно усложняются и формулировки теорем.

5. Доказательство Первой теоремы, данное Шэнноном (воспроизведено в § 2), не является строгим и содержит некоторые пробелы. Например, С. К. Заремба [83] обратил внимание на то, что при вычислении вероятности ошибки Шэннон не учитывает того, что множество входных последовательностей, отвечающих принятой последовательности, не может выбираться произвольно (или случайным образом) из общего множества входных последовательностей, но определяется принятой последовательностью. Поэтому принадлежность

данной входной последовательности к одному из разбиений всего множества не является независимым случайным событием. Правда, можно показать, что при $n \to \infty$ отношение числа элементов любого из подмножеств, соответствующих принимаемым сигналам, к общему числу возможных сигналов стремится к нулю. Это, однако, не снимает необходимости учета всех особенностей при строгом доказательстве.

ГЛАВА XI

СИСТЕМЫ, РАБОТАЮЩИЕ С НЕПРЕРЫВНЫМИ СИГНАЛАМИ

§ 1. СКОРОСТЬ ПЕРЕДАЧИ ИНФОРМАЦИИ, ПРОПУСКНАЯ СПОСОБНОСТЬ И СКОРОСТЬ СОЗДАНИЯ ИНФОРМАЦИИ В СЛУЧАЕ НЕПРЕРЫВНЫХ СИГНАЛОВ

Основой для описания свойств информационных систем, работающих с непрерывными сигналами, служат понятие количества информации, содержащегося в одной непрерывной случайной величине относительно другой величины (см. § 8 гл. VII), понятие дифференциальной энтропии (§ 4 гл. V) и ε-энтропии (§ 9 гл. VII). Все эти понятия имеют корректное количественное определение для случайных величин. При переходе к процессам, какими и являются реальные сигналы, возникает ряд трудностей, связанных, в основном, с необходимостью учета связей между разнесенными во времени значениями реализаций случайного процесса. В некоторых частных случаях эти трудности преодолеваются, но создание более общих построений остается актуальной проблемой.

Изложение результатов, касающихся систем с непрерывными сигналами (в дальнейшем для краткости будем называть их просто непрерывными системами), начнем с наиболее простого случая — когда в высокой степени точности выполняется предположение об ограниченности частотного спектра рассматриваемых сигналов. В этом случае, согласно теореме Котельникова (§ 4 гл. II), любой сигнал полностью определится на всей оси времени, если задать его отсчеты через $1/2F$ секунд, где F — граничная частота спектра. Поэтому для вычисления информационных характеристик системы можно считать, что сигнал как бы дискретен во времени и непрерывен лишь по информационному параметру.

Если при этом считать, что отдельные отсчеты независимы, то полученные характеристики будут предельными для данной непрерывной системы, так как при этом достигается максимальное содержание информации на каждый отсчет.

Опираясь на меру количества информации в одной непрерывной случайной величине (x) относительно другой (y), естественно определить среднюю (на один отсчет) скорость передачи информации по каналу, (в котором $x(t)$ — входной а $y(t)$ — выходной сигналы) как

$$R = \lim_{n\to\infty} \frac{1}{n} \underbrace{\int \cdots \int}_{2n} p(\vec{x}, \vec{y}) \log \frac{p(\vec{x}, \vec{y})}{p(\vec{x})\, p(\vec{y})}\, d\vec{x}\, d\vec{y}, \qquad (1.1)$$

где $\vec{x} = (x_1, x_2, \ldots, x_n)$, $\vec{y} = (y_1, y_2, \ldots, y_n)$ — совокупности n последовательных отсчетов входного и выходного сигналов системы. Если отсчеты независимы, то

$$R = \iint p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)}\, dxdy. \qquad (1.2)$$

Очевидно, если мы хотим скорость передачи исчислять в единицах информации на единицу времени (секунду), то интеграл в правой части (1.2) нужно умножить на число отсчетов в единице времени ($2F$ в секунду).

Перейдем теперь к определению пропускной способности. Поскольку апостериорная неопределенность определяется прежде всего тем, как именно шум взаимодействует с полезным сигналом (что обычно не зависит от устройства передатчика и приемника), то пропускная способность непрерывного канала выражается как

$$C = \max_{p(x)} \iint p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)}\, dxdy, \qquad (1.3)$$

где максимум находится по всевозможным распределениям входной величины x при ограничениях, которые накладывает данный канал. В силу свойства VII количества информации (см. § 10 гл. VII) как бы принятый сигнал y ни обрабатывался впоследствии, это не может увеличить пропускную способность канала C.

Докажем теперь полезную для всего дальнейшего изложения теорему.

Теорема. Если 1) сигнал $x(t)$ и шум $n(t)$ независимы и 2) принимаемый сигнал $y(t)$ является их суммой, $y(t) = x(t) + n(t)$, то количество инфор-

мации на один отсчет может быть вычислено по формуле

$$I(x, y) = H(y) - H(n), \tag{1.4}$$

где $H(y)$ и $H(n)$ — соответственно дифференциальные энтропии принимаемого сигнала и шума (вычисленные, конечно, при одинаковом стандарте сравнения).

Доказательство следует из того, что

$$I(x, y) = H(x) - H(x|y) = H(y) - H(y|x),$$

а так как $y = x + n$, и x статистически не зависит от n, то

$$H(y|x) = H(x + n|x) = H(n), \tag{1.5}$$

откуда и следует (1.4).

В качестве одного из следствий этой теоремы можно сформулировать следующее утверждение: при аддитивном независимом от сигнала шуме максимизация скорости передачи достигается при $\max\limits_{p(x)} H(y)$. Этим можно пользоваться при вычислении пропускной способности соответствующих каналов.

Необходимо, далее, определить скорость создания информации непрерывным источником. Предположим, что нам необходимо передать по системе связи некоторую непрерывную функцию времени $x(t)$. Если бы передатчик мог воспроизводить эту функцию с бесконечной точностью, то на вход линии связи поступало бы бесконечное количество информации на каждый отсчет. Однако в действительности любой передатчик генерирует сигнал $x'(t)$, воспроизводящий $x(t)$ с конечной точностью. Следовательно, на вход линии связи поступает конечное количество информации на каждый отсчет,

$$I(x'x) = \int\int p(x, x') \log \frac{p(x, x')}{p(x)\,p(x')}\, dx dx'. \tag{1.6}$$

Между x и x' имеется связь, определяемая одной из характеристик точности, например, $\overline{(x - x')^2} = \varepsilon^2$, или $|x - x'| \leqslant \varepsilon/2$, или каким-либо другим параметром верности (см. § 3 гл. XI). Количество информации $I(x, x')$ при фиксированной мере точности ε зависит от вида распределения $p(x'|x)$, поэтому скорость создания информации источником определяется как

$$\min_{p(x'|x)} \int\int p(x, x') \log \frac{p(x, x')}{p(x)\,p(x')}\, dx dx'. \tag{1.6}$$

Предполагая, что функция ошибки $n'(t)$ $x(t) - x'(t)$ не зависит от x, пользуясь теоремой (1.4) и вспоминая результаты § 6 гл. V, получаем, что n' при среднеквадратичном критерии точности распределено нормально, а при требовании, чтобы разность $|x-x'|$ не превосходила $\varepsilon/2$ — равномерно. Легко видеть, что такое определение скорости создания информации тождественно определению ε-энтропии (см. § 9 гл. VII). Таким образом, в непрерывном случае скорость создания информации на входе канала определяется ε-энтропией $H_{\varepsilon_0}(x)$ (на один отсчет или на единицу времени) входного сигнала. Это еще один, возможно, решающий довод в пользу обособления понятия ε-энтропии, осуществленного А. Н. Колмогоровым [20].

§ 2. ПРОПУСКНАЯ СПОСОБНОСТЬ ГАУССОВЫХ КАНАЛОВ СВЯЗИ

Назовем гауссовым каналом связи канал, для которого выполняются следующие условия:

1) ширина полосы частот канала ограничена и равна F герц;

2) шум в канале нормален («гауссов шум»);

3) спектр мощности шума равномерен в полосе частот канала («белый» шум) и равен N единиц мощности на единицу полосы частот;

4) средняя мощность полезного сигнала фиксирована и равна P^*;

5) сигнал и шум статистически независимы;

6) выходной сигнал равен сумме полезного сигнала и шума.

Определим пропускную способность C такого канала связи. Согласно теореме сформулированной в предыдущем параграфе, в соответствии с условиями 5) и 6) C определится как

$$C = \max_{p(x)} [H(y) - H(n)]. \tag{2.1}$$

Согласно условиям 2) и 3)

$$H(n) = F \log 2\pi eNF. \tag{2.2}$$

Далее идет следующая цепь рассуждений. Так как сигнал и шум независимы, то они и некоррелированы, и, следовательно, средняя мощность их суммы равна сумме их средних мощностей:

$$y^2 = P^* + NF. \tag{2.3}$$

Необходимо найти максимум дифференциальной энтропии величины $y(t)$ на отсчет при фиксированной средней мощности (2.3). Согласно § 6 гл. V это означает, что $y(t)$ дол-

жен быть распределен нормально; а согласно свойству I § 11 гл. V, спектр мощности сигнала $y(t)$ должен быть равномерным в полосе частот F. Так как $y(t) = x(t) + n(t)$ (условие 6)), то из нормальности y и n следует, что полезный сигнал тоже должен быть нормально распределенным, а из равномерности спектров $y(t)$ и $n(t)$ — следует равномерность спектра $x(t)$. Обозначим, через P величину P^*/F. Тогда

$$\max_{p(x)} H(y) = F \log 2\pi e (PF + NF). \qquad (2.4)$$

Окончательно, вычтя (2.2) из (2.4), имеем:

$$C = F \log \left(1 + \frac{P}{N}\right) \qquad \text{(бит в секунду)} \qquad (2.5)$$

— известная формула Шэннона-Таллера. Таким образом, мы не только нашли выражение пропускной способности, но и показали, что пропускная способность гауссова канала связи реализуется, если закодировать полезный сигнал так, чтобы его спектр был равномерным в предоставленной полосе частот, а распределение мгновенных значений — нормальным.

Формула (2.5) и связанные с ней вопросы интенсивно исследовались различными авторами. В частности, была показана возможность получения формулы (2.5), исходя из геометрической модели сигнала. Кратко воспроизведем этот вывод. Если сигнал обладает энергией E, то (см. § 6 гл II) расстояние точки, изображающей сигнал, от начала координат равно $\sqrt{E}$. Каждый отдельный сигнал, искажаемый шумом, изображается как сферическая область неопределенности около точки сигнала; радиус этой сферы равен $\sqrt{E_n}$, E_n — энергия шума. Сферичность области неопределенности следует из нормальности шума (так как $p(n)$ зависит от Σx_n^2, x_n — отсчеты шума). Ставится задача: найти максимальное число различимых сигналов в условиях (1—6). Для обеспечения максимальности объема, охватывающего пространство возможных принимаемых сигналов с энергией $E = E_s + E_n$, это пространство должно быть сферичным. Будем считать, что сигналы различаются с высокой степенью надежности, если их сферы неопределенности не пересекаются. Тогда верхним пределом числа различимых сигналов является отношение объемов сферы принимаемых сигналов и сферы неопределенности отдельного сигнала:

$$N_0 = \frac{(\sqrt{E_s + E_n})^n}{(\sqrt{E_n})^n} = \left(1 + \frac{E_s}{E_n}\right)^{\frac{n}{2}} = \left(1 + \frac{P}{N}\right)^{\frac{n}{2}}. \qquad (2.6)$$

(n — число отсчетов, принимаемых во внимание). Так как

$$C = \lim_{n\to\infty} \frac{\log N_0}{n}, \qquad (2.7)$$

то C, исчисляемое в битах в секунду, оказывается равным в точности (2.5).

На недостаточность таких рассуждений неоднократно указывал А. А. Харкевич. Если делать геометрический вывод формулы Шэннона-Таллера на строгом уровне, то необходимо учитывать два упущенных фактора. Первый состоит в том, что даже при наиплотнейшей укладке сфер малого радиуса внутри большой сферы максимальное число вмещающихся сфер не будет совпадать с отношением объемов. Более того, при $n\to\infty$ отношение объема всех укладывающихся сфер к объему охватывающей сферы стремится к нулю. Другой фактор состоит в том, что для надежного различия принимаемых сигналов вовсе не необходимо, чтобы сферы неопределенности не пересекались, так как при возрастании числа измерений объем сферы сосредоточивается около ее поверхности; в результате различение со сколь угодно малой вероятностью ошибки оказывается возможным, даже если данная сфера лежит всеми своими частями внутри других сфер. Однако, как показали исследования А. А. Харкевича и Э. Л. Блоха, хотя эти соображения и существенно меняют ход решения, оба фактора действуют в разных направлениях и в конечном счете компенсируют друг друга; так что и корректное решение приводит к формуле Шэннона-Таллера.

§ 3. СКОРОСТЬ ПЕРЕДАЧИ ИНФОРМАЦИИ ПО ГАУССОВЫМ КАНАЛАМ С ПРОИЗВОЛЬНЫМИ СПЕКТРАМИ СИГНАЛА И ШУМА. ОПТИМАЛЬНЫЕ СПЕКТРЫ

Пользуясь формулой Шэннона-Таллера, можно вычислить скорость передачи информации по гауссову каналу связи, если спектры сигнала и шума не являются равномерными. Пусть спектры мощности заданы в интервале частот канала функциями $P(f)$ и $N(f)$. Разбив полосу частот F на малые участки, в которых $P(f)$ и $N(f)$ можно считать постоянными, получаем возможность применить формулу Шэннона-Таллера к элементарным каналам. В пределе, очевидно, получаем:

$$R = \int_F df \cdot \log\left(1 + \frac{P(F)}{N(f)}\right). \qquad (3.1)$$

Это соотношение позволяет решать разнообразные задачи об отыскании оптимальных спектров сигнала или шума при

различных конкретных предположениях. Например, Шэннон [38] рассмотрел следующую задачу. Пусть задан некоторый спектр $N(f)$ гауссова шума в канале с полосой частот F. Как распределить заданную мощность полезного сигнала по частотам, чтобы обеспечить максимальную скорость передачи информации по каналу? Из интуитивных соображений ясно, что основная часть мощности полезного сигнала должна быть сосредоточена в той части спектра, где мощность шумов мала; однако важно дать количественный ответ.

Задача, очевидно, сводится к вариационной задаче нахождения максимума функционала

$$R = \int_{-\infty}^{\infty} A(f) \ln\left(1 + \frac{P(f)}{N(f)}\right) df \tag{3.2}$$

при дополнительном условии

$$\int_{-\infty}^{\infty} P(f)\, df = P^*. \tag{3.3}$$

$A(f)$ — вспомогательная функция, равная единице там, где $P(f) > 0$, и равная нулю для остальных значений f. Необходимость во введении функции $A(f)$ возникает в связи с тем, что возможны случаи, когда имеющуюся мощность полезного сигнала для достижения максимума R придется распределять не по всем частотам данного канала связи, а лишь в некоторых участках его полосы частот. В частности, $A(f) = 0$ вне полосы пропускания канала. $\frac{\partial A}{\partial P}$ есть совокупность δ-функций, отнесенных к точкам, где $P(f)$ обращается в нуль, терпя разрыв первой производной. Представив, например, δ-функцию в виде предела, к которому стремится гауссова кривая при $\sigma \to 0$, получим, что $P(f) \cdot \frac{\partial A}{\partial P} = 0$. С учетом этого обычная вариационная методика приводит к следующему результату

$$P(f) + N(f) = \frac{P^* + N_A^*}{F_A} = \text{const}, \tag{3.4}$$

где $F_A = \int A(f)\, df$ — ширина спектра полезного сигнала, которая в общем случае может быть меньше или равна F, $N_A^* = \int A(f) N(f)\, df$ — мощность шума, приходящегося на частоты, для которых $P(f) \neq 0$.

Введение функции $A(f)$ позволяет исключить решения, не имеющие физического смысла (а именно, такие $P(f)$, которые на некотором интервале частот принимают отрицательные значения), но придают соотношению (3.4) специфический смысл. Для пояснения полученного результата рассмотрим на произвольном примере два различных случая (см. рис. 27). Если величины P^*, N^* (полная мощность шу-

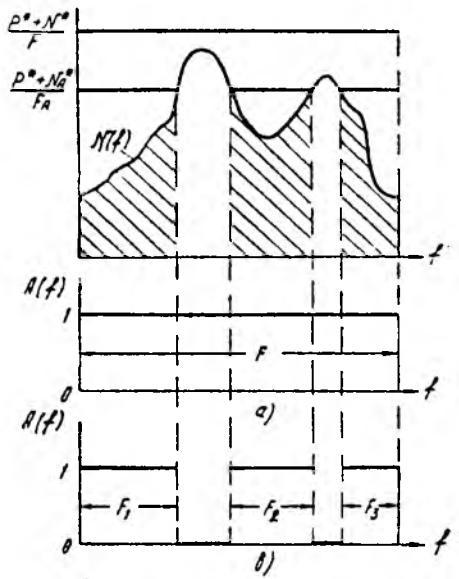

Рис. 27.

ма в полосе F) и функция $N(f)$ таковы, что $N(f)$ нигде не превышает значения $(P^*+N^*)/F$, то для достижения максимальной скорости передачи информации по каналу связи имеющуюся мощность полезного сигнала нужно распределить по частотам так, чтобы сумма спектров мощностей сигнала и шума была постоянной во всем интервале частот канала. В этом случае $A(f)=1$ для всех $f \in F$; $F_A = F$; $N_A^* = N^*$ (см. рис. 27-а). Однако, если распределение $N(f)$ резко неравномерно, и P^* недостаточно велико, чтобы для всех $f \in F$ удовлетворить условию $N(f) \leqslant (P^*+N^*)/F$, то соотношение (3.4) предписывает использовать для передачи не всю полосу частот F, а отфильтровывать области с большими плотностями шумов, распределяя в оставшейся полосе частот мощность сигнала по-прежнему так, чтобы $P(f)+N(f)$ было постоянной величиной. В этом случае $A(f)$ не равна тождественно единице во всем интервале частот F,

что означает, что в таких условиях выгоднее использовать не весь предоставленный участок спектра, а лишь те его части, в которых допускается выполнение равенства (3.4). При этом $F_A = F_1 + F_2 + F_3 < F$ (см. рис. 27, случай *b*) и N_A^* изображено на чертеже заштрихованной площадью. Для наглядности эту ситуацию можно сравнить с поведением жидкости, объем которой пропорционален имеющейся мощности полезного сигнала P^*, наливаемой в сосуд с рельефом дна $N(f)$ при условии сообщаемости всех частей сосуда.

Рассмотрим теперь обратную задачу. Пусть задан спектр полезного гауссова сигнала $P(f)$. Пусть, далее, мы имеем возможность варьировать спектр гауссова шума $N(f)$, сохраняя постоянной его полную мощность $N^* = \int N(f)\,df$. Требуется найти вид функции $N(f)$, обеспечивающей в этих условиях минимальную скорость передачи информации по каналу связи. Для этого необходимо минимизировать

$$R = \int_F \log\left(1 + \frac{P(f)}{N(f)}\right) df \tag{3.5}$$

при условии фиксированности $P(f)$ и

$$\int_F N(f)\,df = N^* = \text{const}. \tag{3.6}$$

Обычная вариационная методика приводит к следующему результату:

$$N(f) = \frac{1}{2} P(f)\left[\sqrt{1 + \frac{4(\lambda)}{P(f)}} - 1\right]. \tag{3.7}$$

Здесь λ — постоянный коэффициент, который определяется при заданной функции $P(f)$ и известной мощности шума N^* путем подстановки (3.7) в (3.6) и вычисления интеграла. Легко видеть, что чем больше N^*, тем больше λ; однако, на величину λ оказывает также влияние конкретный вид функции $P(f)$.

Как и следовало ожидать, $N(f)$ монотонно зависит от величины $P(f)$: туда, где сосредоточена бóльшая мощность полезного сигнала, должна направляться и бóльшая мощность мешающего шума. Обращает на себя внимание, что зависимость $N(f)$ от $P(f)$ носит довольно сложный характер: оптимальное распределение мощности шума по полосе частот даже при одном и том же $P(f)$ будет различным при разных запасах полной мощности N^* источника шума.

Соотношение (3.7) может быть выражено унифицированной кривой, если рассматривать зависимость между нормированными величинами $N(f)/\lambda$ и $P(f)/\lambda$.
Действительно,

$$\frac{N(f)}{\lambda} = \frac{P(f)}{2\lambda} \cdot \left[\sqrt{1 + \left(\frac{P(f)}{\lambda} \right)^{-1}} - 1 \right]. \quad (3.7')$$

Эта кривая приведена на рис. 28.

График показывает, что при $(P(f)/\lambda) > 100$ спектр шума перестает зависеть от формы спектра сигнала. Это означает, в частности, что если имеющаяся мощность шума мала по сравнению с полной мощностью полезного сигнала, то для максимального снижения скорости передачи необходимо равномерно распределить мощность шума по частотам, входящим в спектр полезного сигнала. В этом частном случае (3.7′) легко приводится к более обозримому виду. Действительно, при $(P(f)/\lambda) \gg 1$ для всех $f \in F$ после разложения в ряд корня в (3.7) и пренебрежения членами высших порядков, получим: $N(f) = \lambda = N^*/F$, т. е. мешающий шум должен быть белым. При этом, конечно, предполагается, что $P(f)$ нигде в F не обращается в нуль: нет никакого смысла создавать мешающий шум в той части полосы частот

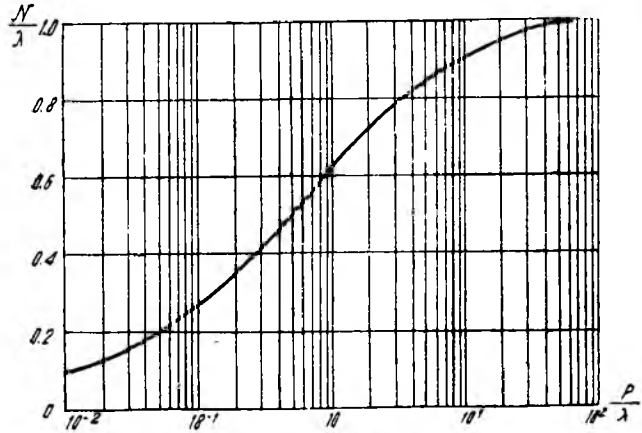

Рис. 28.

канала, где полезная мощность равна нулю. Таким образом, белый гауссов шум в условиях задачи является „наихудшим“ лишь при недостаточной мощности источника шума.

С другой стороны, при наличии источника шума весьма большой мощности (по сравнению с мощностью полезного

сигнала) оптимальный спектр мощности мешающего шума должен выбираться пропорциональным амплитудному спектру полезного сигнала, т. е. $N(f) = \sqrt{\lambda} \cdot \sqrt{P(f)}$. Этот результат легко получается из (3.7′) при условии $(P(f)/\lambda) \ll 1$.

В общем же случае оптимальную форму спектра шума можно найти, определив через (3.7) и (3.6) константу λ и используя график рис. 28.

§ 4. ВЫЧИСЛЕНИЕ КОЛИЧЕСТВА ИНФОРМАЦИИ, ПЕРЕДАВАЕМОГО ПО ГАУССОВЫМ КАНАЛАМ СВЯЗИ. ОЦЕНКИ ПРОПУСКНОЙ СПОСОБНОСТИ НЕПРЕРЫВНЫХ КАНАЛОВ СВЯЗИ

В связи с тем, что нормальный закон распределения является не просто одним из наиболее удобных и «приятных» при теоретическом рассмотрении, но действительно часто встречается в реальных задачах, значительный интерес представляют исследования М. С. Пинскера [88]—[90] и И. М. Гельфанда и А. М. Яглома [65], [87], посвященные информационным свойствам гауссовых случайных процессов. Изложим некоторые важные для приложений результаты этих исследований.

Поскольку определение характеристик системы связи базируется на количестве информации, содержащемся в выходном сигнале относительно входного сигнала, необходимо начать с вычисления интеграла (8.1) (гл. VII)

$$I(x, y) = \int\int p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)}\, dx\, dy, \tag{4.1}$$

где под x и y будем теперь понимать соответствующие сигналы. Пусть входной и выходной сигналы имеют ограниченные спектры одинаковой ширины, так что число отсчетов в разложении Котельникова на заданном интервале времени одинаково для обоих сигналов. Пусть, далее, оба сигнала являются гауссовскими и их совместное распределение также нормально; корреляционные матрицы объектов x, y и (x, y) задаются соответственно как $\|\rho_{xx}(i, \kappa)\|$, $\|\rho_{yy}(j, l)\|$, $\|\rho_{xy}(m, n)\|$; через ρ обозначены соответствующие коэффициенты корреляции, а числовыми аргументами функций ρ служат номера соответствующих отсчетов. Если нас интересует среднее количество информации, содержащееся в отрезке выходного сигнала длиной в n отсчетов, об отрезке входного сигнала такой же длины, то непосредственное вычисление интеграла (4.1) даст

$$I_n(x, y) = \frac{1}{2} \log \frac{D_{xx}^{(n)} \cdot D_{yy}^{(n)}}{D_{xy}^{(n)}}, \tag{4.2}$$

где $D^{(n)}$ — определители корреляционных матриц n-го ранга.

Таким образом, при гауссовых распределениях количество информации полностью определяется совокупностью вторых моментов соответствующих случайных объектов. Если предположить, далее, что отсчеты каждого из сигналов некоррелированы, то можно ограничиться рассмотрением одномерных случайных коррелированных величин x и y. При этом

$$D_{xy} = \begin{vmatrix} \rho_{xx} & \rho_{xy} \\ \rho_{yx} & \rho_{yy} \end{vmatrix} = 1 - \rho_{xy}^2; \quad D_{xx} = D_{yy} = 1, \tag{4.3}$$

и формула (4.2) дает количество информации на один отсчет:

$$I(x, y) = -\frac{1}{2} \log (1 - \rho_{xy}^2). \tag{4.4}$$

Формула (4.2) указывает, что количество информации в одном гауссовом случайном процессе относительно другого гауссова процесса полностью определяется соответствующими авто- и кросс-корреляционными функциями. С другой стороны, известно, что функция корреляции определяет статистический спектр соответствующего случайного процесса. Таким образом, появляется возможность найти связь количества информации со спектральными характеристиками рассматриваемых процессов, аналогично тому, как в § 12 гл. V было найдено выражение дифференциальной энтропии гауссова процесса через его спектр.

Так как при произвольной корреляции отсчетов среднее на один отсчет количество информации определится как

$$I(x, y) = \lim_{n \to \infty} \frac{1}{n} I_n(\vec{x}, \vec{y}), \tag{4.5}$$

то, учитывая (4.2) и (12.8 гл. V), получим:

$$I(x, y) = H(x) + H(y) - H(x, y) =$$

$$= \frac{1}{2F} \int_F \log [G_{xx}(f) \cdot G_{yy}(f)]\, df - \lim_{n \to \infty} \frac{1}{2n} \log D_{xy}^{(n)}. \tag{4.6}$$

Второй член в (4.6) выразится по аналогии с (12.8 гл. V) через взаимную спектральную плотность G_{xy} процессов $x(t)$ и $y(t)$ и спектральные плотности G_{xx} и G_{yy} этих процессов. М. С. Пинскер показал, что

$$\lim_{n \to \infty} \frac{1}{2n} \log D_{xy}^n = \frac{1}{2F} \int_F \log [G_{xx}(f) G_{yy}(f) - |G_{xy}(f)|^2]\, df. \tag{4.7}$$

Отсюда сразу следует, что

$$I(x,y) = \frac{1}{2F}\int_F \log \frac{G_{xx}(f)\cdot G_{yy}(f)}{G_{xx}(f)\,G_{yy}(f) - |G_{xy}(f)|^2}\,df. \qquad (4.8)$$

Переходя к безразмерной круговой частоте $\lambda = \pi\frac{f}{F}$ и рассматривая спектры на всей оси $-\infty < \lambda < \infty$, формулу (4.8) можно записать в виде

$$I(x,\ y) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log \frac{f_{xx}(\lambda)\cdot f_{yy}(\lambda)}{f_{xx}(\lambda)\,f_{yy}(\lambda) - |f_{xy}(\lambda)|^2}\,d\lambda. \qquad (4.9)$$

Соотношение (4.9) известно под названием формулы Пинскера.

Формула Пинскера может быть обобщена на случай неограниченных спектров [88]. Для этого строится последовательность случайных гауссовых процессов с ограниченными, но расширяющимися спектрами, которая сходится к заданным процессам с неограниченными спектрами. Можно ввести параметр h, обратно пропорциональный ширине спектра члена такой последовательности. Тогда количество информации, приходящееся на единицу времени для заданного h, определится в соответствии с (4.9) как

$$\frac{1}{h}I_{n,h}(x,y) =$$

$$= \frac{1}{4\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} \log \frac{\sum\limits_{\kappa=-\infty}^{\infty} f_{xx}\left(\lambda+\frac{2\pi\kappa}{h}\right)\sum\limits_{\kappa=-\infty}^{\infty} f_{yy}\left(\lambda+\frac{2\pi\kappa}{h}\right)}{\sum\limits_{\kappa=-\infty}^{\infty} f_{xx}\left(\lambda+\frac{2\pi\kappa}{h}\right)\sum\limits_{\kappa=-\infty}^{\infty} f_{yy}\left(\lambda+\frac{2\pi\kappa}{h}\right) - \left|\sum\limits_{\kappa=-\infty}^{\infty} f_{xy}\left(\lambda+\frac{2\pi\kappa}{h}\right)\right|^2} d\lambda. \qquad (4.10)$$

Устремив теперь h к нулю, получаем обобщенную формулу Пинскера для количества информации на единицу времени:

$$I(x,y) = \frac{1}{4\pi}\int_{-\infty}^{\infty} \log \frac{f_{xx}(\lambda)\,f_{yy}(\lambda)}{f_{xx}(\lambda)\,f_{yy}(\lambda) - |f_{xy}(\lambda)|^2}\,d\lambda. \qquad (4.11)$$

Основываясь на приведенных выше результатах и опираясь на свойство экстремальности энтропии нормального распределения, М. С. Пинскер получил ряд теорем о пропускной способности гауссовых и других непрерывных каналов связи. Приведем некоторые из этих теорем.

Теорема 1. Если условное распределение $p(\vec{y}|\vec{x})$ нормально, а входной сигнал x может иметь любое распределение с заданной матрицей вторых моментов, то

$$C = \frac{1}{2} \log \frac{D_{xx} D_{yy}}{D_{xy}}, \tag{4.12}$$

где D_{xy} и D_{yy} — определители матриц центральных вторых моментов распределений $p(\vec{x}, \vec{y})$ и $p(\vec{y})$.

Теорема 2. Если x может иметь любое распределение с заданной матрицей центральных вторых моментов, то

$$C \geqslant \frac{1}{2} \log \frac{D_{xx} D_{yy}}{D_{xy}}, \tag{4.13}$$

где D_{xy} и D_{yy} — определители матриц центральных вторых моментов распределений $p(\vec{x}', \vec{y}')$ и $p(\vec{y}')$, вычисленные в предположении, что $\vec{x}'$ нормально распределен.

Теорема 3. Пусть $x(t)$ и $y(t)$ стационарны и стационарно связаны, и $x(t)$ имеет неограниченный спектр и может иметь любое распределение с фиксированной функцией корреляции. Тогда

$$C \geqslant \frac{1}{4\pi} \int_{-\infty}^{\infty} \log \frac{f_{xx}(\lambda) f_{y'y'}(\lambda)}{f_{xx}(\lambda) f_{y'y'}(\lambda) - |f_{x'y'}(\lambda)|^2} d\lambda, \tag{4.14}$$

где $f_{xx}(\lambda)$ — спектральная плотность процесса $x(t)$,
$f_{y'y'}(\lambda)$ — спектральная плотность процесса $y'(t)$,
$f_{x'y'}(\lambda)$ — взаимная спектральная плотность процессов $x'(t)$ и $y'(t)$, полученная в предположении, что $x'(t)$ — гауссовский процесс.

§ 5. ТЕОРЕМА КОДИРОВАНИЯ ДЛЯ НЕПРЕРЫВНЫХ КАНАЛОВ

Смысл и значение понятия пропускной способности непрерывных каналов получают завершенность лишь после изложения теоремы о кодировании, аналогичной соответствующим теоремам для дискретных каналов. Эта теорема гласит:

Если источник имеет ε-энтропию $H_{\varepsilon}(x)$, меньшую или равную пропускной способности C

канала связи с шумами, то существует код, при котором продукция источника может быть передана по данному каналу и воспроизведена на его выходе с точностью, сколь угодно близкой к ε_0. Это невозможно, если $H_{\varepsilon_0}(x) > C$.

Необходимость условия $H_{\varepsilon_0}(x) \leqslant C$ для возможности передачи сигналов без ухудшения точности следует уже из самого определения пропускной способности. Что касается доказательства достаточности этого условия, то Шэннон предложил [7] строить доказательство по аналогии с доказательствами «дискретных» теорем. Идея построения такой аналогии сводится к тому, чтобы пространство сигналов разбить на большое число малых ячеек и рассматривать этот случай как дискретный, применив далее все рассуждения предыдущей главы. Указав далее на смысл ε_0-энтропии как такого количества информации, по которому можно восстановить непрерывный сигнал с ошибкой не большей ε_0, можно завершить доказательство теоремы.

Однако при таком подходе подразумевается кодирование и передача информации в дискретном виде и восстановление непрерывного сигнала уже в приемной системе. Поэтому представляет интерес отыскание доказательств, не предполагающих в процессе передачи преобразования непрерывных величин в дискретные, и обратно. Некоторые общие вопросы построения таких доказательств обсуждены А. Н. Колмогоровым [20]. Я. Г. Синай [91] оценил наименьшую дисперсию ошибки, появляющейся при передаче по непрерывному каналу в предположении, что кодирование и декодирование осуществляется с помощью линейных фильтров. Это, однако, не решает общей проблемы нахождения оптимальных непрерывных кодов, остающейся пока открытой.

ГЛАВА XII

СЛОЖНЫЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ

§ 1. ПОСЛЕДОВАТЕЛЬНОЕ И ПАРАЛЛЕЛЬНОЕ СОЕДИНЕНИЯ КАНАЛОВ СВЯЗИ

Проблемы, которые возникают при рассмотрении сложных информационных систем, весьма разнообразны. К числу наиболее общих из них относится следующая. Пусть задано произвольное соединение определенного числа каналов связи с известной пропускной способностью C_k каждого канала; какова пропускная способность C системы в целом?

Рассмотрение сложных систем естественно начать с наиболее простых из них, — с последовательного и параллельного соединений каналов связи.

Для определения свойств системы в целом недостаточно лишь указать, что каналы включены, например, последовательно. Многое будет зависеть от того, как именно осуществлено сопряжение соседних каналов: осуществляется ли оптимальное декодирование и кодирование на стыках каналов, или просто выходной алфавит одного канала преобразуется побуквенно во входной алфавит другого и т. п.

Рассмотрим случай, когда в параллельном или последовательном соединении каналов реализуются условия безошибочной передачи (шумы отсутствуют) по каждому из каналов или допустимы задержки, обеспечивающие возможность оптимального кодирования в каждом канале при наличии шумов. Результаты получаются немедленно из общих положений и являются очевидными:

1. Пропускная способность C последовательного соединения каналов равна минимальному числу из множества $\{C_k\}$ пропускных способностей соединяемых каналов:

$$C = \min_k C_k. \tag{1.1}$$

2. Пропускная способность параллельного соединения каналов равна сумме пропускных способностей соединяемых каналов:

$$C = \sum_{\kappa} C_{\kappa}. \tag{1.2}$$

Соотношения (1.1) и (1.2) могут служить основой для определения характеристик более сложных систем. Ясно, например, что если сложная система может быть разбита на блоки, включенные параллельно или последовательно, то для определения ее пропускной способности можно воспользоваться формулами (1.1) и (1.2), подразумевая под C_{κ} пропускные способности блоков, которые сами могут состоять из нескольких каналов связи. Правда, не всякое соединение каналов может быть представлено до конца в виде чисто параллельно-последовательной схемы, в особенности если связь по некоторым каналам системы однонаправленная. Для оценки пропускной способности таких систем требуется более общий подход, который будет рассмотрен в следующем параграфе.

Обратимся теперь к случаю, когда соединения между последовательными каналами системы таковы, что ошибки, возникшие в одном из каналов, не исправляются на стыке со следующим каналом, который в свою очередь имеет не нулевое рассеяние информации, и так далее. К таким системам относятся, например, радиорелейные линии передач, дальние телефонные линии с рядом промежуточных пунктов усиления и т. п. Общих результатов для таких систем пока не имеется, однако, для одного важного случая можно получить [102] некоторые полезные оценки.

Пусть имеется n последовательно соединенных каналов связи с шумами. Так как сигнал x_{κ} на выходе κ-го канала зависит только от сигнала $x_{\kappa-1}$ и только через $x_{\kappa-1}$ зависит от сигналов остальных (предыдущих) каналов, то последовательность $\{x_1, x_2, \ldots, x_{\kappa}, \ldots, x_n\}$ является простой марковской цепью. По выходному сигналу x_n судят об отправленном сигнале x_1. Что можно сказать о количестве информации $I(x_1, x_n)$, о том, как эта величина связана с содержанием информации о промежуточных звеньях цепи?

Количество информации, содержащееся в x_n относительно x_1, определяется соотношением

$$I(x_1, x_n) = \sum_{x_1, x_n} p(x_1, x_n) \log \frac{p(x_n|x_1)}{p(x_n)}. \tag{1.3}$$

Выражение под знаком логарифма можно подвергнуть тождественным преобразованиям так, чтобы включить в явном виде вероятности, связанные с κ-м элементом марковской цепи:

$$\frac{p(x_n|x_1)}{p(x_n)} = \frac{p(x_n,\ x_\kappa)}{p(x_n)\cdot p(x_\kappa)} \cdot \frac{p(x_n|x_1)}{p(x_n|x_\kappa)}. \tag{1.4}$$

Второй множитель легко привести к виду

$$\frac{p(x_n|x_1)}{p(x_n|x_\kappa)} = \frac{p(x_n|x_1)\,p(x_\kappa|x_1)}{p(x_n,\ x_\kappa|x_1)}. \tag{1.5}$$

Так как

$$\sum_{x_\kappa} p(x_m,\ x_\kappa,\ x_l) = p(x_m,\ x_l), \tag{1.6}$$

то (1.3) с учетом (1.4) и (1.5) можно записать как

$$I(x_1,\ x_n) = \sum_{x_\kappa,\ x_n} p(x_\kappa,\ x_n) \log \frac{p(x_\kappa,\ x_n)}{p(x_\kappa)\,p(x_n)} -$$

$$- \sum_{x_1} p(x_1) \sum_{x_\kappa x_n} p(x_\kappa,\ x_n|x_1) \log \frac{p(x_\kappa,\ x_n|x_1)}{p(x_\kappa|x_1)\,p(x_n|x_1)}. \tag{1.7}$$

Второй член в (1.7) естественно толковать как математическое ожидание количества информации в x_n относительно x_κ при известном x_1. Окончательно имеем:

$$I(x_1,\ x_n) = I(x_\kappa,\ x_n) - MI(x_\kappa,\ x_n|x_1), \quad (\kappa = 1,\ 2,\ \ldots,\ n), \tag{1.8}$$

Это соотношение обобщает результат Р. Л. Добрушина [67], полученный им для $n = 3$.

При больших n вычисления по формуле (1.7) становятся громоздкими, поэтому представляет интерес дать просто вычисляемые оценки для $I(x_1,\ x_n)$. Из общих положений теории информации следуют очевидные неравенства:

$$I(x_1,\ x_n) < \min_{\kappa} I(x_\kappa,\ x_{\kappa-1}) \leqslant H(x_\kappa). \tag{1.9}$$

Привлекая уравнение Колмогорова-Чемпена, можно получить еще одну оценку. Для дискретных марковских процессов уравнение Колмогорова-Чемпена записывается в виде

$$p(x_\kappa|x_{\kappa-2}) = \sum_{x_{\kappa-1}} p(x_\kappa|x_{\kappa-1})\,p(x_{\kappa-1}|x_{\kappa-2}). \tag{1.10}$$

Применив это равенство $n-2$ раза, получим:

$$p(x_n|x_1) = \sum_{x_{n-1}} p(x_n|x_{n-1}) \cdot \sum_{x_{n-2}} p(x_{n-1}|x_{n-2}) \cdot \sum_{x_{n-3}} \ldots$$

$$\ldots \sum_{x_3} p(x_1|x_3) \sum_{x_2} p(x_3|x_2) \cdot p(x_2|x_1) =$$

$$= \sum_{x_{n-1}, \ldots, x_2} \prod_{\kappa=1}^{n-1} p(x_{\kappa+1}|x_\kappa). \qquad (1.11)$$

Подставив (1.11) в (1.3) и применив к результату общее неравенство Иенсена для выпуклой функции $f(x)$:

$$\sum_{\kappa} p_\kappa f(x_\kappa) < f\left(\sum_{\kappa} p_\kappa x_\kappa\right) \qquad (1.12)$$

(где $\sum_{\kappa} p_\kappa = 1$), получим после некоторых преобразований искомую оценку:

$$I(x_1, x_n) < H(x_n) - MH(x_2, x_3, \ldots, x_n|x_1) + \log(n-2). \qquad (1.13)$$

§ 2. ПРОИЗВОЛЬНОЕ СОЕДИНЕНИЕ КАНАЛОВ СВЯЗИ В СЛОЖНУЮ СЕТЬ

Пусть имеется некоторая достаточно сложная сеть, состоящая из каналов связи. Для общности предположим, что часть этих каналов обеспечивает связь лишь в одном направлении, другие каналы позволяют передавать информацию в любом направлении; пропускные способности каждого из каналов известны (для простоты будем считать, что двухсторонние каналы имеют одинаковые пропускные способности в обоих направлениях; в случае необходимости это ограничение легко снимается). В этой сети произвольным образом указываются два пункта (например, пункты A и B на рис. 29. Направление связи указано стрелками, двусторонние каналы стрелок не имеют; пропускные способности обозначены цифрами). Требуется а) определить пропускную способность системы связи между пунктами A и B, образованной всеми каналами сети, б) указать, как распределить поток информации по каналам, чтобы полностью реализовать пропускную способность системы*). Обратимся сначала к первой задаче.

Структура системы либо вообще может не позволить представить ее в виде параллельно-последовательной системы, либо допускает это, но остается весьма громоздкой. Поэтому желательно развить более общий подход, чем рассмотренный в начале предыдущего параграфа. Такой подход можно позаимствовать из теории графов: легко видеть, что рис. 29 явля-

*) Отметим, что эта задача легко может быть сформулирована так, чтобы охватить проблему организации грузоперевозок по сети дорог, распределения электроэнергии в сложной энергосистеме и т. п.

ется типичным изображением графа. Основную теорему, которая разрешает поставленную задачу, доказали Дж. Б. Данциг и Д. Р. Фулкерстон [92]; ее детальное рассмотрение в применении к информационным сетям привели П. Элиас, А. Фейнстейн и К. Э. Шэннон [109].

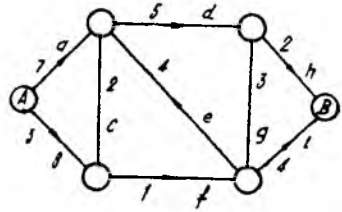

Рис. 29.

Для дальнейшего необходимо ввести понятие разреза сети. Разрезом называется совокупность ветвей графа, удаление которой из сети приведет к распадению сети на две или более несвязанных части, причем заданные пункты A и B окажутся в разных частях сети*). Ясно, что любой путь от A до B включает в себя по крайней мере один элемент разреза. Примерами разрезов сети на рис. 29 являются совокупности (d, e, f), (b, c, e, g, h), (a, b, c, f). Определим далее простой разрез как такой разрез, из которого нельзя изъять какую-либо ветвь без того, чтобы оставшиеся ветви не перестали составлять разрез. Например, (d, e, f) и (b, c, e, g, h) являются простыми разрезами; разрез (a, b, c, f) не является простым. Удаление простого разреза из сети приводит к ее распадению точно на две части с рассматриваемыми пунктами A и B в разных частях. Каждому простому разрезу припишем величину, называемую ценой разреза, которую определим как сумму пропускных способностей каналов, входящих в разрез. При вычислении цены разреза необходимо учитывать направленность связи и суммировать только пропускные способности, соответствующие рассматриваемому направлению. Так, цена разреза (d, e, f) равна $5+0+1=6$ для связи от A к B и равна $0+4+0=4$ для связи от B к A; цена разреза $(a, в)$ равна 10 в одном направлении и 0 в другом; разрез $(в, c, e, g, h)$ имеет цену 10 для связи от A к B.

Сформулируем теперь основную теорему для сложных сетей

Пропускная способность системы связи между двумя заданными узлами сложной

*) Таким образом, разрез определяется относительно двух заданных узлов графа; однако, для краткости это обычно не оговаривается.

сети равна минимальной цене простого разреза:

$$C_{AB} = \min_{\{\sigma\}} \sum_{\sigma} C_{ik}. \qquad (2.1)$$

Здесь C_{ik} — пропускная способность канала связи между пунктами i и k в направлении от A к B; σ — множество каналов, входящих в простой разрез; $\{\sigma\}$ — множество простых разрезов.

Хотя справедливость теоремы представляется почти очевидной, воспроизведем простое ее доказательство [109]. Рассмотрим некоторое произвольное распределение потоков информации, идущей от A к B. Естественные ограничения, которым удовлетворяет это распределение, состоят в том, что а) скорости передачи информации по элементарным каналам не превосходят соответствующих пропускных способностей ($R_{ik} \leqslant C_{ik}$), б) потоки информации по каналам стационарны, а емкости запоминающих устройств в узлах сети конечны, так что выполняется «закон Кирхгофа»: алгебраическая сумма потоков информации, относящихся к одному узлу, равна нулю.

Пусть минимальная цена простого разреза s есть C_s. При некотором произвольном распределении потоков информации алгебраическая сумма скоростей передачи информации по каналам, входящим в разрез s, не превосходит, очевидно, цены этого разреза:

$$R = \sum_{s} R_{ik} \leqslant C_s. \qquad (2.2)$$

Так как, согласно „закону Кирхгофа", алгебраическая сумма потоков информации для каждого (и для всех вместе) узла в правой (от разреза) части сети равна нулю, то общий поток информации от A к B равен $R \leqslant C_{AB}$. А так как $\max R = C_{AB}$, то цена разреза C_s и есть пропускная способность системы в целом, что и требовалось доказать. Укажем, что теоремы о параллельном и последовательном соединении каналов связи (см. § 1) легко получаются из доказанной теоремы как частные случаи.

Обратимся теперь ко второй основной проблеме эксплуатации сложной сети—проблеме распределения потоков информации с тем, чтобы действительно реализовать потенциальную способность C_{AB} сети. Необходимо прежде всего доказать, что в принципе возможно подобрать такое распределение потоков информации, которое, удовлетворяя условиям а) и б), сформулированным выше, обеспечивало бы передачу информации со скоростью C_{AB}.

Введем понятие редуцированной сети, соответствующей любой заданной сети с минимальной ценой разреза C_{AB}.

Редуцированная сеть обладает следующими свойствами:

1. Граф редуцированной сети тот же, что и граф исходной сети.

2. Пропускная способность C'_{ik} каждой ветви графа редуцированной сети меньше или равна пропускной способности C_{ik} соответствующей ветви исходной сети (в частности, C'_{ik} может быть равной нулю):

$$C_{ik} \geqslant C'_{ik} \geqslant 0. \tag{2.3}$$

3. Каждая ветвь редуцированной сети входит по крайней мере в один из минимальных разрезов*).

4. Минимальная цена простого разреза редуцированной сети равна минимальной цене C_{AB} простого разреза исходной сети.

Один из способов построения редуцированной сети состоит в следующем [109]. Если данная ветвь графа не входит в какой-либо минимальный разрез, то будем уменьшать ее пропускную способность до тех пор, пока либо цена рассматриваемого разреза не сравняется с минимальной, либо пропускная способность ветви не достигнет нуля. Затем та же операция последовательно проделывается с каждой ветвью, не входящей хотя бы в один минимальный разрез, пока таких ветвей вообще не останется. Построенная таким образом сеть, очевидно, удовлетворяет всем перечисленным выше условиям.

Ясно, что одной и той же исходной сети каналов может соответствовать некоторое множество редуцированных сетей, так как порядок, в котором редуцируются отдельные ветви, является произвольным. С другой стороны, очевидно, что если потоки информации по редуцированной сети таковы, что реализуется пропускная способность C_{AB}, то данное распределение потоков информации будет реализовывать C_{AB} и в исходной нередуцированной сети. Таким образом, достаточно доказать, что распределение потоков, при котором скорость передачи информации равна C_{AB}, может быть найдено для редуцированной сети.

Доказательство существования указанного распределения потоков информации основывается на том, что редуцирование сети можно всегда провести так, что редуцированная сеть будет параллельно-последовательной схемой, для которой справедливость доказываемого утверждения очевидна. Ясно, что в связи с неоднозначностью процесса редуцирования может существовать несколько распределений потоков информа-

*) Для краткости иногда будем называть минимальным разрезом разрез, цена которого минимальна; минимальных разрезов с одинаковой ценой может быть несколько.

ции, обеспечивающих передачу информации с одинаковой скоростью, близкой к $C_{\text{ав}}$.

Основываясь на приведенных результатах, можно разрешать и более сложные проблемы, связанные с информационными сетями. Пусть, например, имеется сеть с несколькими входами и несколькими выходами с заданными пропускными способностями (рис. 30). Требуется указать, при каких условиях возможна эксплуатация входных и выходных каналов на полную мощность. Построив вспомогательную сеть (см. рис. 31) и привлекая рассмотренные выше свойства сложных сетей, сразу получаем, что такими условиями являются:

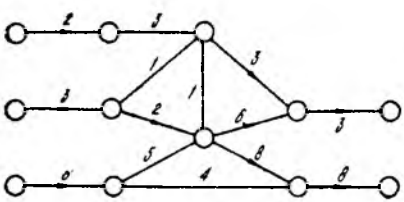

Рис. 30.

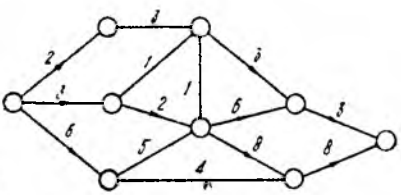

Рис. 31.

1) равенство суммы пропускных способностей входных каналов сумме пропускных способностей выходных каналов (пусть эта сумма равна C);

2) минимальная цена простого разреза вспомогательной сети должна быть не меньше C.

§ 3. СИСТЕМЫ С ОБРАТНОЙ СВЯЗЬЮ

Большой интерес представляют различные информационные системы, в структуре которых имеются петли обратных связей. Важность изучения таких систем следует уже из того, что класс систем с обратными связями чрезвычайно широк: он охватывает все системы автоматического регулирования и автоматического управления, а также целый ряд систем связи и других информационных систем. Влияние обратных связей проявляется в различных системах по-разному. В одних системах введение обратных связей не изменяет их потенциальных возможностей, свойства других систем количественно изменяются при введении обратных связей, а для такого важного класса систем, как системы автоматического управления, наличие или отсутствие обратных связей является не второстепенным признаком, а определяет все основные качества системы. В связи с таким разнообразием в проявлении действия обратных связей на свойства системы представляется целесообразным рассматривать отдельно несколько классов информационных систем с обратными связями.

Рассмотрим сначала системы связи с обратной связью. Модель такой системы, изображенная на рис. 32, отражает многие реальные системы связи. Например, при телефонном разговоре принимающий телефонограмму обычно переспрашивает или просит повторить плохо расслышанные слова; если требуется особая надежность при приеме, получатель

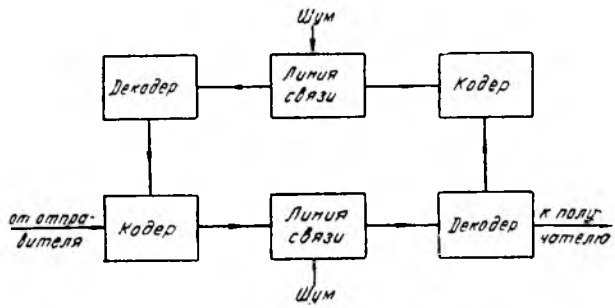

Рис. 32.

может повторить весь текст, чтобы отправитель подтвердил правильность принятого текста. Другим примером использования обратной связи может служить импульсная автоматическая система связи при наличии шума с квантованием на три области [104]. Если принятый сигнал превысил верхний порог, приемник фиксирует наличие импульса; при недостижении нижнего порога фиксируется отсутствие импульса; если же принятый сигнал оказался между порогами квантования — считается, что сигнал «неразборчив», и по каналу обратной связи запрашивается повторение передачи символа.

Во всех подобных случаях обратная связь используется для улучшения качества передачи информации в прямом направлении. И это качество действительно улучшается, но это связано, очевидно, с введением избыточности и соответствующим уменьшением скорости передачи информации. В связи с этим возникает принципиальный вопрос: как наличие обратной связи изменяет потенциальные характеристики системы связи, в первую очередь — пропускную способность системы?

Для каналов без памяти легко доказать следующую теорему:

Пропускную способность системы связи нельзя увеличить введением обратной связи, какова бы ни была пропускная способность канала обратной связи.

Это утверждение является простым следствием теоремы

о сложных сетях, доказанной в предыдущем параграфе*). Действительно, систему связи с обратной связью можно представить графом рис. 33. Каналы прямой и обратной связи образуют простой разрез, цена которого в прямом направлении равна пропускной способности C канала связи и не меняется, как бы не изменялась пропускная способность $C_{обр}$ канала обратной связи. Это и доказывает сформулированную теорему.

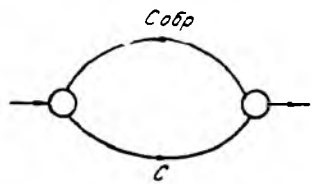

Рис. 33.

Несмотря на невозможность увеличения пропускной способности канала с помощью обратной связи, использование обратных связей представляет большой интерес в том отношении, что позволяет осуществить достаточно эффективное кодирование более простыми методами, в частности, с несколько (иногда — значительно) меньшей необходимой емкостью памяти, чем при кодировании для канала без обратной связи. Однако многие вопросы, связанные с такой возможностью, еще мало исследованы.

Положение существенным образом меняется, если шумы в канале связи не являются независимыми от символа к символу («канал с памятью»). При этом на основе сведений, поступающих по каналу обратной связи, можно предсказать (статистически) поведение шума на некоторое время вперед, а значит — и повысить пропускную способность системы в целом. Ясно, что для реализации такой возможности необходимо выполнение ряда дополнительных условий. Прежде всего, нужно, чтобы за время передачи достаточно длинной группы символов характер шума не сильно изменился; далее, канал обратной связи должен обеспечивать достаточно быстрое получение информации о шуме, чтобы эта информация не устарела и не стала бесполезной. Полное исследование возникающих здесь возможностей до сих пор не проведено. Один интересный частный случай рассмотрен Р. Л. Добрушиным [93]. Им получены количественные оценки пропускной способности в предположении, что память прямого канала и пропускная способность обратного канала столь велики, что можно считать сведения, поступающие по каналу обратной связи,

*) Прямое доказательство, не основывающееся на указанной теореме, дано Р. Л. Добрушиным [93].

безошибочными и не теряющими своей ценности сколь угодно долго.

Перейдем теперь к рассмотрению принципиально другого класса систем с обратными связями — систем автоматического управления. Теория информации не может здесь претендовать на достаточно исчерпывающее описание, так как многие вопросы разработки систем автоматического управления не являются чисто информационными. Однако некоторые аспекты управления требуют оценки информационных характеристик элементов контура управления. Важно, например, указать, какова необходимая пропускная способность канала обратной связи для поддержания процесса в стационарном режиме, какова необходимая емкость памяти управляющей системы и т. п.

Для иллюстрации того, как система автоматического управления может рассматриваться методами теории информации, обратимся к достаточно простому случаю [94] дискретного во времени управления некоторым процессом по схеме, приведенной на рис. 34. Для наших целей процесс управления необходимо рассматривать как процесс «циркуляции» информации по петле обратной связи.

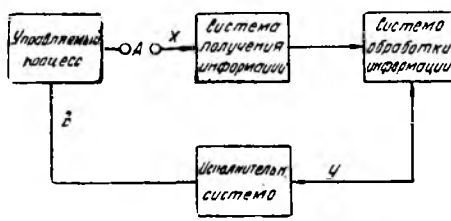

Рис. 34.

Разомкнем контур управления, например, в точке A и подадим на правую клемму некоторый известный сигнал x, а затем пронаблюдаем через интервал дискретности сигнал x' на левой клемме разрыва A. В силу наличия шумов в измерителях и системе обработки информации, ошибок исполнительной системы и влияния внешних возмущающих факторов, действующих на сам управляемый процесс, сигнал x' не совпадает с x, однако, содержит некоторую информацию о нем. В среднем эта информация определяется величиной

$$I(x,\ x') = H(x') - MH(x'|x). \qquad (3.1)$$

Поскольку управление определяется не по средним характеристикам, а непосредственно по сигналу x, введем в рассмотрение условную энтропию $H(x'|x)$, тождественно преобразуя (3.1) к виду

$$I(x,\ x') = H(x') - H(x'|x) - S, \tag{3.2}$$

где $S = H(x'|x) - MH(x'|x)$. Величину S естественно назвать приращением энтропии приведенных шумов в системе управления, имея в виду, что $H(x'|x)$ (а следовательно, и $MH(x'|x)$) отличны от нуля лишь в силу влияния случайных воздействий в различных частях системы управления.

Если теперь замкнуть контур управления, ликвидировав разрыв A, то информационное соотношение (3.2) остается в силе, если рассматривать x и x' как сигналы о состоянии управляемого процесса в моменты времени, отстоящие друг от друга на интервал дискретности управления. В соответствии с этим на n-м шаге управления обозначим x через x_{n-1}, x' через x_n; энтропии $H(x')$ и $H(x'|x)$ будут соответственно выражать энтропии управляемого процесса в $(n-1)$-й и n-й моменты времени. Тогда для замкнутого контура управления имеем:

$$I_{n,\,n-1} = H_{n-1} \quad H_n + S_{n.\,n-1}. \tag{3.3}$$

Это основное уравнение выражает общие свойства системы автоматического управления и позволяет сформулировать [94] несколько утверждений типа теорем.

1. Изменение энтропии управляемого процесса за интервал дискретности управления равно разности между приращением энтропии приведенных шумов и количеством информации, прошедшим по контуру управления за тот же интервал времени. Это утверждение следует из записи (3.3) в виде

$$\Delta H = H_n - H_{n-1} = S_{n.\,n-1} - I_{n.\,n-1}. \tag{3.4}$$

2. Если управление обеспечивает стационарный характер сигнала x, то $H_n = H_{n-1}$ и $I_{n.\,n-1} = S_{n.\,n-1}$. Это означает, что для поддержания стационарного режима управления необходимо обеспечить передачу по контуру управления количества информации, численно равного приращению энтропии приведенных шумов.

3. Если процесс неуправляем, т. е. $I_{n.\,n-1} = 0$, то $H_n - H_{n-1} = S_{n.\,n-1}$. Приращение энтропии приведенных шумов равно возрастанию энтропии неуправляемого процесса. Это еще раз поясняет смысл введения величины S.

Теоремы такого рода могут облегчить решение ряда практически важных задач при проектировании систем автомати-

ческого управления. Однако необходимо указать, что для общего случая непрерывного управления требуется провести обобщение, которое вовсе не очевидно.

Рассмотренные системы с обратной связью (т. е. системы связи и одноконтурные системы автоматического управления) не исчерпывают всех возможных вариантов; построение достаточно общей информационной теории таких систем остается пока нерешенной задачей.

§ 4. КАНАЛЫ СВЯЗИ СО СЛУЧАЙНЫМИ ПАРАМЕТРАМИ

К числу сложных информационных систем могут быть отнесены многолучевые (многопутевые) системы связи. Они характеризуются двумя основными особенностями: 1) выходной сигнал системы является определенной комбинацией некоторого множества сигналов, пришедших от источника по различным путям, 2) свойства среды на пути каждого из лучей меняются во времени неконтролируемым, случайным образом. Примерами таких каналов являются системы связи на тропосферном или ионосферном рассеянии, системы коротковолновой связи с несколькими лучами и т. п.

Изучение таких систем может вестись различными методами и на разных уровнях абстракции. С позиций теории информации впервые, по-видимому, рассматривал такие системы А. Фейнстейн [103] в 1953—1954 гг. В последующие годы основные работы по определению пропускной способности различных разновидностей многолучевых систем выполнены И. А. Овсеевичем и М. С. Пинскером [95]-[99], В. И. Сифоровым [100]-[101] и Б. С. Цыбаковым [105]-[108].

Приведем здесь результаты чисто феноменологического подхода к рассмотрению многолучевых систем. Идея такого подхода состоит в том, что всякую многолучевую систему можно заменить некоторым эквивалентным однолучевым каналом связи, параметры которого случайно изменяются, причем так, что выходной сигнал однолучевого канала оказывается тождественным реальному выходному сигналу многолучевой системы. Например, флуктуации амплитуды полезного сигнала на выходе приемника, вызываемые в действительности изменением фазовых соотношений между лучами, можно рассматривать как результат эквивалентного случайного изменения коэффициента затухания среды в однолучевом канале.

Примем теперь следующие конкретные предположения.

1) Приемная система чувствительна только к амплитуде принимаемого сигнала.

2) Амплитуда a полезной компоненты принимаемого сигнала флуктуирует в соответствии с плотностью вероятностей вида

$$p(a) = \frac{2a}{\sigma^2} \exp\left(-\frac{a^2 + a_0^2}{\sigma^2}\right) I_0\left(\frac{2aa_0}{\sigma^2}\right). \qquad (4.1)$$

Известно, что в ряде практически важных случаев это распределение хорошо соответствует экспериментальным данным. Физический смысл параметров распределения (4.1): a_0 — амплитуда прямой волны, σ^2 — средняя мощность рассеянной компоненты полезного сигнала. Мощность всей полезной компоненты на выходе приемника равна a^2P, где P — средняя мощность полезного сигнала.

3) Ширина спектра полезного сигнала ограничена полосой F герц.

4) Выходной сигнал всей системы содержит кроме полезной компоненты шумовую составляющую; будем считать шум аддитивным и нормальным со средней мощностью N.

5) Условия связи таковы, что в каждый момент значение величины a известно и кодирование обеспечивает при этом максимальную скорость передачи информации. Это возможно, например, при достаточно медленных замираниях и наличии обратной связи.

В силу условий (4) и (5) и теоремы Шэннона о пропускной способности гауссовых каналов, при заданном a пропускная способность канала равна

$$C(a) = F \ln\left(1 + \frac{a^2P}{N}\right), \qquad (4.2)$$

а средняя пропускная способность определится как

$$C = \int_0^\infty C(a)\, p(a)\, da =$$

$$= F \int_0^\infty \frac{2a}{\sigma^2} \exp\left(-\frac{a^2 + a_0^2}{\sigma^2}\right) I_0\left(\frac{2a_0a}{\sigma^2}\right) \ln\left(1 + \frac{a^2P}{N}\right) da. \qquad (4.3)$$

Так как этот интеграл не вычисляется в конечном виде, представим функцию Бесселя в виде ряда

$$I_0\left(\frac{2a_0a}{\sigma^2}\right) = \sum_{n=0}^{\infty} \left(\frac{a_0}{\sigma^2}\right)^{2n} \cdot \frac{a^{2n}}{(n!)^2}. \qquad (4.4)$$

Если теперь ввести функцию

$$Q(\mu) = \int_0^\infty \exp(-u\mu) \ln\left(1 + \frac{\sigma^2P}{N}\, u\right) du =$$

$$= -\frac{1}{\mu}\exp\left(\frac{\mu P}{N}\right)\cdot \mathrm{Ei}\left(-\frac{\mu P}{\sigma^2 N}\right), \qquad (4.5)$$

где $\mathrm{Ei}(x)$ — интегральная показательная функция,

$$\mathrm{Ei}(x) = \int_{-\infty}^{x} \frac{e^t}{t}\,dt,$$

то легко убедиться, что *)

$$C = F\exp\left(-\frac{a_0^2}{\sigma^2}\right)\sum_{n=0}^{\infty}\frac{(-1)^n}{(n!)^2}\left(\frac{a_0^2}{\sigma^2}\right)^n \frac{\partial^n}{\partial\mu^n}Q(\mu) =$$

$$= F\exp\left(-\frac{a_0^2}{\sigma^2}\right)\sum_{n=0}^{\infty}\frac{(-1)^{n+1}}{(n!)^2}\left(\frac{a_0^2}{\sigma^2}\right)^n \times$$

$$\times\left[\frac{\partial^n}{\partial\mu^n}\frac{1}{\mu}\exp\left(\frac{\mu P}{\sigma^2 N}\right)\mathrm{Ei}\left(-\frac{\mu P}{\sigma^2 N}\right)\right]_{\mu=1} \qquad (4.6)$$

Интересно рассмотреть один частный случай. Пусть прямой луч отсутствует, $a_0 = 0$ (случай „чистого рассеяния"). Тогда все слагаемые в (4.6), кроме первого, равны нулю, и

$$C = -F\exp\left(\frac{P}{\sigma^2 N}\right)\mathrm{Ei}\left(-\frac{P}{\sigma^2 N}\right). \qquad (4.7)$$

Определим коэффициент уменьшения пропускной способности при рассеянии как

$$\eta\left(\frac{P'}{N}\right) = \frac{C}{C_1}, \qquad (4.8)$$

где $P' = \sigma^2 P$, $C_1 = F\ln\left(1+\frac{P'}{N}\right)$; смысл величины η становится ясным, если учесть, что C_1 — пропускная способность обычного гауссова канала при мощности полезного сигнала, равной мощности рассеянного сигнала. На рис. 35 представлена зависимость η от P'/N. Из этого графика следует, что уменьшение пропускной способности за счет рассеяния не превышает 17 % — факт, впервые отмеченный В. И. Сифоровым в 1957 г.

*) Формула (4.6) получена Б. С. Цыбаковым [107].

Из условий (1)-(5), в рамках которых получена формула (4.6), наибольшее ограничение накладывается условием (5): представляется весьма трудным реализовать достаточно быстродействующую систему кодирования и декодирования, следящую за изменением уровня полезного сигнала. Поэтому представляет интерес рассмотрение случая, когда такая подстройка невозможна или не производится.

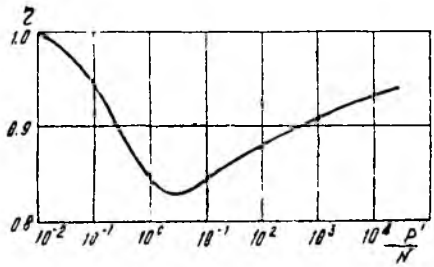

Рис. 35.

Для нахождения пропускной способности канала со случайными параметрами необходимо найти максимальное количество информации, которое может содержаться в выходном сигнале y относительно входного сигнала x, если $y = ax + z$, где z — нормальный шум, a — случайный коэффициент с известными статистическими свойствами. В общем виде решить эту задачу не удается из-за значительных вычислительных трудностей. Однако при некоторых упрощающих предположениях можно получить интересные соотношения.

Основное предположение, которое мы примем, будет состоять в том, что флуктуации коэффициента передачи среды a достаточно малы, так что $\frac{P}{N} \cdot \sigma^2 \ll 1$. Тогда можно провести все вычисления, даже не конкретизируя вида распределения $p(a)$, потребовав лишь, чтобы третий начальный момент распределения $p(a)$ был порядка σ^3 (что практически всегда имеет место).

Задача определения пропускной способности C сводится к отысканию максимального значения функционала

$$I(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_\xi(x)\, p_{\eta/\xi}(y \mid x) \ln \frac{p_{\eta/\xi}(y \mid x)}{\int_{-\infty}^{\infty} p_\xi(x)\, p_{\eta/\xi}(y \mid x)\, dx}\, dx\, dy \qquad (4.9)$$

путем вариации $p_\xi(x)$. Так как дисперсия флуктуаций σ^2 мала, то мы рассматриваем случай, близкий к отсут-

ствию флуктуаций. При $a = \text{const}$ имеем обычный гауссов канал, оптимальная плотность вероятностей $p_\xi(x)$ для которого нормальна $(0, P)$. Естественно, что при малых σ^2 оптимальная плотность будет мало отличаться от нормальной, поэтому

$$p_\xi(x) = N_x(0,P) + \sigma^2 f(x). \tag{4.10}$$

Поправочная функция $f(x)$ обладает в силу условий

$$\int_{-\infty}^{\infty} p_\xi(x)\,dx = 1,$$

$$\int_{-\infty}^{\infty} x^2 p_\xi(x)\,dx = P, \tag{4.11}$$

свойствами

$$\int_{-\infty}^{\infty} f(x)\,dx = 0$$

$$\int_{-\infty}^{\infty} x^2 f(x)\,dx = 0. \tag{4.12}$$

Получим теперь приближенное выражение для $p_{\eta/\xi}(y/x)$. Очевидно $p_{\eta/\xi}(y/x)$ представляется сверткой распределений $p(a)$ и $p_\zeta(z)$, или, через характеристические функции,

$$p_{\eta/\xi}(y/x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} g_a(tx)\, g_z(t)\, e^{-ity}\,dt. \tag{4.13}$$

Разложим $g_a(t)$ в ряд

$$g_a(t) = e^{it\bar{a}}\left[1 + \sigma^2\frac{(it)^2}{2!} + g_a'''(t_1)\,\frac{(it)^3}{3!} + \ldots\right] \tag{4.14}$$

и ограничимся первыми двумя членами (воспользовавшись введенным выше ограничением на третий момент). Тогда, вычислив (4.13) с точностью до членов порядка σ^3, учитывая, что $p_\zeta(z) = N_z(0,N)$, имеем

$$p_{\eta/\xi}(y/x) = p_\zeta(y-\bar{a}x) + \frac{\partial^2 p_\zeta(y-\bar{a}x)}{\partial y^2}\cdot\frac{x^2\sigma^2}{2} =$$

$$= N_y(\bar{a}x, N) + \frac{\partial^2 N_y(\bar{a}x, N)}{\partial y^2}\cdot\frac{x^2\sigma^2}{2}. \tag{4.15}$$

Теперь формула (4.9) может быть с точностью до членов порядка σ^2 записана в виде

$$C = 2FI_{\max} = F2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} N_x(0, P)\, p_{\eta|\xi}(y|x) \ln \times$$

$$\times \frac{p_\zeta(y - ax)}{\int_{-\infty}^{\infty} N_x(0, P)\, p_\zeta(y - \bar{a}x)dx}\, dx\, dy +$$

$$+ \left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x) p_\zeta(y - \bar{a}x) \ln \frac{p_\zeta(y - \bar{a}x)}{\int_{-\infty}^{\infty} N_x(0, P)\, p_\zeta(y - \bar{a}x)\, dx}\, dx dy - \right.$$

$$\left. - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)\, p_\zeta(y - \bar{a}x)\, dx dy \right] \cdot \frac{\sigma^2}{2} \cdot 2F. \qquad (4.16)$$

После несложных преобразований и вычисления двух последних интегралов можно показать, что оба они равны нулю в силу свойств (4.12) функции $f(x)$.

Подставив далее в первый интеграл (4.16) $p_{\eta|\xi}(y|x)$ в виде (4.15) и $p_\zeta(z) = N_z(0, N)$, получим:

$$C = F\left[\ln\left(1 + \frac{\bar{a}^2 P}{N}\right) - \frac{\bar{a}^2}{1 + \frac{\bar{a}^2 p}{N}} \cdot \frac{P^2}{N^2} \cdot \sigma^2\right]. \qquad (4.17)$$

Воспользовавшись малостью $\sigma^2 \dfrac{P^2}{N^2}$, имеем:

$$\ln\left(1 + \frac{\bar{a}^2 P}{N}\right) - \frac{\bar{a}^2}{1 + \frac{\bar{a}^2 P}{N}} \cdot \frac{P^2}{N^2} \cdot \sigma^2 =$$

$$= \ln\left(1 + \frac{\bar{a}^2 P}{N}\right) + \ln \exp\left(- \frac{\bar{a}^2}{1 + \frac{\bar{a}^2 P}{N}} \cdot \frac{P^2}{N^2} \cdot \sigma^2\right) =$$

$$= \ln\left(1 + \frac{\bar{a}^2 P}{N}\right) + \ln\left(1 - \frac{\bar{a}^2 \cdot \sigma^2 \cdot \frac{P^2}{N^2}}{1 + \frac{\bar{a}^2 P}{N}}\right) =$$

$$= \ln \left[\left(1 + \frac{\overline{a^2} P}{N} \right) \left(1 - \frac{\overline{a^2}\sigma^2 \frac{P^2}{N^2}}{1 + \frac{\overline{a^2} P}{N}} \right) \right] =$$

$$= \ln \left[1 + \frac{\overline{a^2} P}{N} \left(1 - \sigma^2 \frac{P}{N} \right) \right] =$$

$$= \ln \left[1 + \frac{\overline{a^2} P}{N \left(1 + \sigma^2 \frac{P}{N} \right)} \right] = \ln \left(1 + \frac{\overline{a^2} P}{N + \sigma^2 P} \right).$$

Окончательно:

$$C = F \ln \left(1 + \frac{\overline{a^2} P}{N + \sigma^2 P} \right). \qquad (4.18)$$

Из этой формулы следует, что флуктуации полезной компоненты снижают пропускную способность канала. При малых флуктуациях C приблизительно равна пропускной способности гауссова канала, мощность шумов в котором равна $N + \sigma^2 P$, а мощность полезного сигнала—средней мощности полезной компоненты.

Формула (4.18) в более частных предположениях была впервые получена Фейстейном; вывод, приведенный здесь, дан Цыбаковым. В наиболее общих предположениях формула (4.18) получена Овсеевичем и Пинскером, которые показали, что выражение (4.18) является нижней гранью для пропускной способности рассматриваемых каналов.

ЛИТЕРАТУРА

К главе I

1. Нейман М. С. Об общей теории сигналов и общей теории автоматических процессов. Радиотехника, 1955, 10, № 5, 13—16.

2. Тарасенко Ф. П. Некоторые общие вопросы теории структуры сигналов. «Вычислительная техника. Автоматика. Теория информации». Труды СФТИ, 1961, 40, 3—7.

К главе II

3. Габор (Gabor D.) A Summary of Communication Theory. Proc. First London Sympos. on Inform. Theory, 1952.

4. Дуб Дж. Вероятностные процессы. ИЛ, 1958.

5. Железнов Н. А. Некоторые вопросы спектрально-корреляционной теории нестационарных сигналов. Радиотехника и электроника, 1959, № 3, 359—373.

6. Лэннинг Дж., Бэттин Р. Случайные процессы в задачах автоматического регулирования, ИЛ, 1958.

7. Шэннон, Уивер (Shannon C. E., Weaver W.) The Mathematical Theory of Communication. University Illinois Press, 1949, 3—89. (неполный русский перевод в сборнике «Теория передачи электрических сигналов при наличии помех» под редакцией Н. А. Железнова, ИЛ, 1953).

8. Яглом А. М. Введение в теорию стационарных случайных функций, УМН, 1952, т. 7, вып. 5.

К главе III

См. 7, 5.

9. Александров М. И. Распределение фазы и огибающей смеси синусоидального сигнала с шумом. Вестник НИИ МРТП, 1955, 3 (54), 36—42.

10. Берж К. Теория графов и ее применения. ИЛ, 1962.

11. Биллингс (Billings A. K.) Sampling of Signals without D. C.-component. Electronic and Radio-Engeneer, 1959, № 2,70.

12. Бунимович В. И. Флуктуационный процесс как колебание со случайными амплитудой и фазой, ЖТФ, 1949, 19, 1231—1259

13. Бунимович В. И. Флуктуационные процессы в радиоприемных устройствах, Сов. радио, 1951.

14. Гаврилов М. А., Шастова Г. А. Основные вопросы теории

построения сигналов. Сессия АН СССР по научным проблемам автоматизации производства. 1956, т. 4, 90—119.

15. Габор (Gabor D.) Communication Theory and Physics. Phil. Mag., 1950, 41, 1161.

16. Железнов Н. А. О принципиальных вопросах теории сигналов, Радиотехника, 1957, № 11, 3—13.

17. Железнов Н. А. Принцип дискретизации непрерывных сигналов с неограниченным спектром. Радиотехника и электроника, 1958, № 1, 3—18.

18. Задэ, Хаггинс (Zadeh L., Huggins M.) Signal Flow Graphs PIRE, 1957, № 10, 1413—1414.

19. Колмогоров А. Н., Экстраполяция и интерполяция случайных процессов. Бюллетень МГУ, 1941, т. 2, вып. 6.

20. Колмогоров А. Н. Теория передачи информации, Сессия АН СССР, 15—20 октября 1956 г. Пленарные заседания. Изд. АН СССР, 1957, 66—99.

21. Котельников В. А. Теория потенциальной помехоустойчивости, ГЭИ, 1956.

22. А. Г. Майер, Е. А. Леонтович. Об одном неравенстве, связанном с интегралом Фурье. ДАН СССР 1934, т. IV, № 7, 353—360.

23. Мидлтон (Middleton D.) Some Results in the Theory of Noises. Quart. Appl. Math., 1948, 5, № 4, 445—498.

24. Мурский В. Л. Об эквивалентных преобразованиях контактных схем. Проблемы кибернетики, № 5, 61—76.

25. Мэсон (Mason S.) Signal Flow Graphs PIRE, 1953, 41, № 9, 1144—1156.

26. Мэсон (Mason S.) Feedback Theory and Signal Flow Graphs. PIRE, 1956, 44, № 7, 920—926.

27. Пугачев В. С. Теория случайных функций, ГИТТЛ, 1957.

28. Райс (Rice S. O.) Mathematical Analysis of Noise. BSTJ, 1944, 23, № 3, 282—332; 1945, 24, № 1, 46—156 (русский перевод в сборнике «Теория передачи электрических сигналов при наличии помех» — под редакцией Н. А. Железнова, ИЛ, 1953).

29. Рыжик И. М., Градштейн И. С. Таблицы интегралов, сумм, рядов и произведений, ГИТТЛ, 1951.

30. Тракселл Д. Синтез систем автоматического регулирования, Машгиз, 1959.

31. Турбович И. Т. Некоторое обобщение теоремы Котельникова, Радиотехника, 1956, № 4, 5—14.

32. Турбович И. Т. К вопросу о применимости теоремы Котельникова к функциям времени с неограниченным спектром, Радиотехника, 1958, № 8, 11—12.

33. Турбович И. Т. Аналитическое представление функции времени с неограниченным спектром, Радиотехника, 1959, № 3, 22—27.

34. Фурдуев В. В. О некоторых основных понятиях теории сигналов, Радиотехника, 1957, № 4.

35. Хаггинс (Huggins W.) Signal Flow Graph and Random Signals. PIRE, 1957, № 1, 74—86.

36. Харкевич А. А. Спектры и анализ, ГИТТЛ, 1957.

37. Цыбаков Б. С., Яковлев В. П. О точности восстановления непрерывной функции, представленной конечным рядом Котельникова, Радиотехника и электроника, 1959, № 3, 542.

38. Шэннон (Shannon C. E.) Communication in the Presence of Noise. PIRE, 1949, 37, 10—21 (русский перевод в сборнике «Теория информации и ее приложения» под редакцией А. А. Харкевича, ГИФМЛ, 1959, 82—112).

К главе IV

См. 7, 20, 27.

39. Блэкман (Blachman) Bounds for Entropy. Journ. Appl. Phys., 1953, 24, № 10, 1340.

40. Добрушин Р. Л. Об измерении энтропии стационарных случайных последовательностей. Теория вероятностей и ее применения, 1958, т. 3, вып. 4.

41. Джейнс (Jaynes) Information Theory and Statistical Mechanics. Phys. Rev., 1957, № 4, 520—530.

42. Крамер Г. Математические методы статистики, ИЛ, 1949.

43. Макмиллан (McMillan B.) The Basic Theorems of Information Theory, Ann. Math. Stat., 1953, 24, № 2, 196—219.

44. Прайс (Price R.) On Entropy Equivalence in the Time — and Frequency Domains. PIRE, 1955, 43, № 4, 484—485.

45. Тарасенко Ф. П. Зависимость энтропии случайной функции от порядка распределения и статистической связи аргументов распределения, Изв. вузов, Физика, 1958, № 1.

46. Тарасенко Ф. П. Об энтропийных характеристиках случайных процессов с непрерывным временем. Труды СФТИ, 1961, вып. 40, 24—28.

47. Хинчин А. Я. Понятие энтропии в теории вероятностей, УМН, 1953, 3 (55), 3—20.

48. Хинчин А. Я. Об основных теоремах теории информации, УМН, 1956, 1 (67), 17—75.

К главе V

49. Анисимов С., Вислобоков А. Некоторые философские вопросы кибернетики. Коммунист, 1960, № 2, 108—118.

50. Берг А. И. Некоторые проблемы кибернетики. Вопросы философии, 1960, № 5.

51. Бриллюэн Л. Наука и теория информации, ГИФМЛ, 1960.

52. Влэдуц Г. Э., Налимов В. В., Стяжкин Н. И. Научная и техническая информация как одна из задач кибернетики, УФН, 1959, XIX, вып. 1, 13—56.

53. Винер Н. Кибернетика. Сов. радио, 1958.

54. Винер Н. Кибернетика и общество, ИЛ, 1959.

55. Иванов С. Г. Некоторые философские вопросы кибернетики, Ленинград, 1960.

56. Ленин В. И. Материализм и эмпирнокритицизм. Сочинения, т. 14, стр. 81.

57. Мак-Кэй (McKay) The Place of 'Meaning' in the Theory of Information. Information Theory 3-rd London Symposium, 1955; London, 1956, 215—225.

58. Новик И. Б. О некоторых методологических проблемах кибернетики. Сб. «Кибернетику на службу коммунизму», 1, 1961, 34—54.

59. Ровенский З., Уемов А., Уемова Е. Машина и мысль (философский очерк о кибернетике), Госполитиздат, 1960.

60. Соболев С. Л., Китов А. И., Ляпунов А. А. Основные черты кибернетики, Вопросы философии, 1955, № 4, 136—148.

61. Ерохин В. ε-энтропия дискретного случайного объекта. Т. Вер. и ее прим., 1958, т. III, вып. 1, 103—107.

62. Харкевич А. А. О ценности информации. Проблемы кибернетики, № 4, 53—57, ГИФМЛ, 1960.

63. Яглом А. М., Яглом И. М. Вероятность и информация, ГИФМЛ, 1960.

64. Вудворд, Дэвис (Woodward Ph. M., Davies I. L.) Information Theory and Inverse Probability in Telecommunikation PIRE, 1952, № 58,

37—44, (русский перевод в сборнике «Теория передачи электрических сигналов при наличии помех под редакцией Н. А. Железнова, ИЛ, 1953).
65. Гельфанд И. М., Яглом А. М. О вычислении количества информации о случайной функции, содержащейся в другой такой функции. УМН, 1957, XII, 1 (73), 3—52.
66. Гольдман С. Теория информации, ИЛ, 1958.
67. Добрушин Р. Л. Общая формулировка основной теоремы Шэннона в теории информации, УМН, 1959, XIV, 6 (60), 3—104.
68. Фэно (Fano R. M.) The Statistical Theory of Information, Nuovo Cimento, 1959, 13, Suppl. № 2, 353—372.
69. Харди Г. Г., Литтлвуд Дж. Е., Полиа Г. Неравенства, ИЛ, 1948.
70. Хартлн (Hartley R. V. L.). Transmission of Information, BSTJ, 1928, 7, № 3, 535—563 (русский перевод в сборнике «Теория информации и ее приложения» под редакцией А. А. Харкевича).

К главе VI

См. 51, 68.

К главе VII

См. 7, 20, 61.

К главе VIII

См. 7, 21, 48.

71. Бородин Л. Ф., Зотова Е. Н. Параметры систем передачи дискретных сообщений. НДВШ, Радиотехника и электроника, 1958, № 1, 27—36.
72. Закревский А. Д. Метод синтеза функционально-устойчивых автоматов. ДАН СССР, 1959, 129, 4, 729—731.
73. Закревский А. Д. Функциональная устойчивость релейных схем. Вычислительная техника. Автоматика. Теория информации». Труды СФТИ, 1961, вып. 40. 112—126.
74. Зюко А. Г. К определению общетехнических характеристик систем связи. Сборник трудов НТОРиЭ им. А. С. Попова, 1958, вып. II, 5—11.
75. Мур Э. Ф., Шэннон К. Э. Надежная схема из ненадежных реле. Кибернетический сборник № 1, ИЛ, 1960, 109—148.
76. Нейман Дж. Вероятностная логика и синтез надежных организмов из ненадежных компонент. Сборник статей «Автоматы», ИЛ, 1956, 68—139.
77. Сифоров В. И. К теории каналов радиосвязи с многолучевым распространением. Сборник трудов НТОРиЭ им. А. С. Попова, 1958, вып. II, 56—86.
78. Сифоров В. И. Параметры систем бинарного кодирования. Электросвязь, 1957, № 1.
79. Харкевич А. А. Очерки общей теории связи. ГИТТЛ, 1955.
80. Харрингтон (Harrington J. V.) An analysis of the Detection of Repeated Signals in Noise by Binary Integration. IRE Trans., 1955, IT-I, № 1, 1.

К главе IX

См. 7, 63.

81. Фейнстейн А. Основы теории информации. ИЛ, 1960.

К главе X

См. 7, 20, 43, 48, 67, 81.

82. Борнард Г. А. Простые доказательства простых частных слу-

чаев теоремы кодирования. Сборник статей «Теория передачи сообщений» под редакцией Сифорова В. И., ИЛ, 1957, 32—39, 40—42.

83. Заремба С. К. Заключение к основной теореме для дискретного канала с шумами. Сборник «Теория передачи сообщений» под редакцией В. И. Сифорова, ИЛ, 1957, 28—31.

84. Коды с обнаружением и исправлением ошибок. Сборник статей под редакцией А. М. Петровского, ИЛ, 1956.

85. Шэннон (Shannon C. E.) Certain Results in Coding Theory for Noisy Channels. Information. Control, 1957, sept.

86. Элиас (Elias P.) Coding for Noisy Channels. IRE Conv. Rec., 1955, pt. 4.

К главе XI

См. 7, 20, 38, 65.

87. Гельфаид И. М., Колмогоров А. Н., Яглом А. М. Количество информации и энтропия для непрерывных распределений. Труды Третьего Всесоюзного математического съезда. Изд. АН СССР, 1958, т. III, 300—320.

88. Пиискер М. С. Количество информации о гауссовском случайном процессе, содержащейся во втором процессе, стационарно с ним связанном. ДАН СССР, 1954, 99, № 2, 213—216.

89. Пинскер М. С. Количество информации об одном стационарном случайном процессе, содержащееся в другом стационарном случайном процессе. Труды Третьего Всесоюзного математического съезда, Изд. АН СССР, 1956, т. I, 125.

90 Пинскер М. С. Вычисление скорости создания сообщений стационарным случайным процессом и пропускной способности стационарного каиала. ДАН СССР, 1956, III, № 4, 753—756.

91. Синай Я. Г. Наименьшая ошибка и наилучший способ передачи стационарных сообщений при линейном кодировании в случае гауссовских каналов связи. Проблемы передачи информации, 1959, вып. 2, 40—48.

К главе XII

См. 67, 77.

92. Данциг Дж. Б., Фулкерстон Д. Р. Теорема о максимальном потоке и минимальном разрезе в сетях. Сборник статей «Линейные неравенства и смежные вопросы» под редакцией Куна и Таккера. Перевод с английского под редакцией Канторовича и Новожилова, ИЛ, 1959, 318—324.

93. Добрушин Р. Л. Передача информации по каналу с обратной связью. Теория вероятностей и ее применения, 1958, вып. 4, 395—412.

94. Красовский А. А., Поспелов Г. С. Осиовы автоматики и технической кибернетики. Изд. ВВИА им. Жуковского, 1961.

95. Овсеевич И. А., Пиискер М. С. Оценка пропускной способиости канала связи, параметры которого являются случайными функционалами времени. Радиотехника, 1957, № 10, 40—46.

96. Овсеевич И. А., Пинскер М. С. Оценка пропускной способности некоторых релейных каналов связи. Радиотехника, 1958, № 4, 15—25.

97. Овсеевич И. А., Пинскер М. С. О пропускной способности многопутевой системы информации. ИАН ОТН, Эиергетика и автоматика, 1959, № 1, 133—135.

98. Овсеевич И. А., Пинскер М. С. Оптимальное линейное предыскажение и корректирование сигнала при передаче его по многопутевой системе. ИАН ОТН, Энергетика и автоматика, 1959, № 2, 49—59.

99. Овсеевич И. А., Пинскер М. С. Скорость передачи информации и пропускная способность многопутевой системы и прием по методу линейно-операторного преобразования. Радиотехника, 1959, № 3, 9—21.

100. Сифоров В. И. О собственной пропускной способности каналов

связи со случайными изменениями параметров. НДВШ, Радиотехника и электроника, 1958, № 1, 7—11.

101. Сифоров В. И. К теории каналов радиосвязи с многолучевым распространением. Сборник трудов НТОРиЭ им. А. С. Попова, вып. II, 1958, 56—86.

102. Тарасенко Ф. П. Передача информации по марковской цепи. «Вычислительная техника. Автоматика. Теория информации». Труды СФТИ, 1961, вып. 40, 15—17.

103. Фейнстен (Feinstein A.) Some Jnformation Theory Aspects of Propagation Through Time Varying Media, IRE. Conv Rec. 1954, pt. I, 85—97.

104. Харрис, Хаупстайн, Шварц (Harris B., Hauptstein A., Schwartz L. S.) Optimum Decision Feedback Systems, IRE Conv. Rec., 1957, pt. 2, 3—10.

105. Цыбаков Б. С. О пропускной способности двухлучевых каналов связи. Радиотехника и электроника, 1959, № 7, 1116—1123.

106. Цыбаков Б. С. О пропускной способности каналов с большим числом лучей. Радиотехника и электроника, 1959, № 9, 1427—1433.

107. Цыбаков Б. С. Пропускная способность некоторых многолучевых каналов связи. Радиотехника и электроника, 1959, № 10, 1602—1608.

108. Цыбаков Б. С. О пропускной способности однолучевого канала со случайными изменениями поглощения НДВШ, Радиотехника и электроника, 1959, № 2, 44—50.

109. Элиас, Фейнстейн, Шэинон (P. Elias, A. Feinstein, C. E. Shannon). A Note on the Maximum Flow Through a Network. IRE Trans. 1956, IT—2, № 4 117—119.

СОДЕРЖАНИЕ

Феликс Петрович Тарасенко

ВВЕДЕНИЕ В КУРС ТЕОРИИ ИНФОРМАЦИИ

Редактор издательства М. И. Волкова

Корректоры М. И. Сваровская, О. Л. Болдырева

К301725. Сдано в набор 11/XII-62 г. Подписано к печати 20/VIII-63 г. Бумага 60×92¹/₁₆. Объем 15 п. л.; 7,5 бум. л. Заказ 6199-62 г. Тираж 5000 Цена в переплете 1 р. 20 коп.

Издательство Томского университета, пр. Ленина, 34.
Типография № 1 Полиграфиздата, г. Томск, Советская, 47.

Цена 1 руб. 20 коп.