ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ
УНИВЕРСИТЕТ имени академика С.П. КОРОЛЕВА»

В. А. ФУРСОВ

ЛЕКЦИИ ПО ТЕОРИИ ИНФОРМАЦИИ

Под редакцией Н.А. Кузнецова

Допущено учебно-методическим советом по прикладной математике и информатике УМО по классическому университетскому образованию в качестве учебного пособия для студентов высших учебных заведений, обучающихся по специальности и направлению «Прикладная математика и информатика» и по направлению «Информационные технологии»

С А М А Р А
Издательство СГАУ
2006

УДК 519.72
ББК 32.811
Ф 95

Инновационная образовательная программа "Развитие центра компетенции и подготовка специалистов мирового уровня в области аэрокосмических и геоинформационных технологий"

Рецензенты: д-р физ.-мат. наук, В. М. Чернов,
д-р техн. наук,. О. В. Горячкин

Фурсов В. А.

Ф 95 **Лекции по теории информации:** учеб. пособие / *В. А. Фурсов;* под ред. Н.А. Кузнецова – Самара: Изд-во Самар. гос. аэрокосм. ун-та, 2006. – 148 с. : ил.

ISBN Б-7883-0458-X

В учебном пособии рассматриваются модели сигналов, основы теории информации и кодирования, а также некоторые вопросы приема и обработки информации. Книга составлена как сборник лекций, каждая из которых посвящена одной теме. Дается конспективное изложение основных вопросов. Лекции занимают промежуточное положение между справочниками и солидными изданиями и адресованы студентам, обучающимся по учебным планам бакалавров и специалистов.

Печатается по решению Редакционно-издательского совета Самарского государственного аэрокосмического университета.

УДК 519.72
ББК 32.811

ISBN Б-7883-0458-X

ПРЕДИСЛОВИЕ

Идея подготовки настоящего пособия возникла в связи с переходом к подготовке прикладных математиков по двухступенчатой схеме. В учебных планах подготовки бакалавров по направлению 510200 предусматривается лекционный курс теории информации и кодирования объемом около 35 часов. В рамках указанного сравнительного небольшого объема необходимо было сохранить достаточно полную и глубокую подготовку, которая традиционно обеспечивалась учебным планом подготовки по специальности 010200.

В 1977 году в Куйбышевском авиационном институте (ныне Самарский государственный аэрокосмический университет) вышло в свет учебное пособие [10] (автор В.А. Сойфер). В нем рассматриваются вопросы теории информации и кодирования, которые составляют основу курса. Наряду с этим в учебные программы входят также разделы, посвященные рассмотрению моделей сигналов, а также вопросам их обнаружения и восстановления параметров. Это нашло отражение в изданиях других авторов [3], [7]. Вместе с тем в указанных книгах либо недостаточно внимания уделено фундаментальным теоремам теории информации [7], либо имеет место перегруженность техническими вопросами реализации методов [3], что не является задачей подготовки специалистов и бакалавров по прикладной математике.

В связи с этим, потребовалось пересмотреть структуризацию материала с целью придания курсу большей компактности. При отборе материала авторы стремились дать основные теоретические сведения, на которых базируется ряд последующих специальных дисциплин. В частности, включены вопросы помехоустойчивого кодирования с использованием линейных последовательных машин, задачи обнаружения и оценивания. Вместе с тем от многих излагаемых, например, в [3] вопросов, связанных со схемными решениями, пришлось отказаться.

Книга составлена как сборник лекций, каждая из которых посвящена одной теме, что по замыслу авторов должно облегчить самостоятельную работу над курсом. В учебном пособии дается краткое конспективное изложение основных вопросов. Вместе с тем, авторы стремились к тому, чтобы в пособии нашли отражение ключевые вопросы математического описания сигналов, теории информации и кодирования. По замыслу лекции должны занять промежуточное положение между справочниками и солидными изданиями.

Авторы выражают признательность заведующему кафедрой технической кибернетики СГАУ, члену-корреспонденту РАН В.А. Сойферу, внимательно прочитавшему рукопись и высказавшему ряд полезных советов по содержанию учебного пособия, а также А.В. Гаврилову, и Н.Е. Козину, выполнившим набор текста рукописи книги.

Учебное пособие подготовлено при финансовой поддержке Министерства образования и науки РФ, Администрации Самарской области и Американского фонда гражданских исследований и развития (CRDF).

ВВЕДЕНИЕ

Понятие информации. Предмет и задачи курса

Термин «Информация» относится к числу наиболее часто употребляемых. Он широко используется в лингвистике, психологии, биологии и других науках. Однако в разных областях знаний в него вкладывают разный смысл. Разнообразие информационных процессов и широкий интерес к ним в разных областях знаний породили много толкований определений понятия «информация», а также определений количества информации.

Условно все подходы к определению количества информации [6] можно разделить на пять видов:

1) энтропийный;

2) алгоритмический;

3) комбинаторный;

4) семантический;

5) прагматический.

Первые три вида дают количественное определение сложности описываемого объекта или явления. Четвертый – описывает содержательность и новизну передаваемого сообщения для получателя (пользователя) сообщения. Наконец, пятый вид обращает внимание на полезность полученного сообщения для пользователя.

Термин «информация» происходит от латинского слова «informatio», что означает «разъяснения» и, по сути, предполагает наличие некоторого диалога между отправителями и получателями информации. Следовательно, информационное взаимодействие можно представить пятикомпонентной (пятимерной векторной) величиной, состоящей из компонент:

1) физической;

2) сигнальной;

3) лингвистической;

4) семантической;

5) прагматической.

Заметим, что приведенное разбиение информационного взаимодействия на пять компонентов носит условный характер и возможно частичное пересечение в этом разбиении. Так, отдельные составляющие передаваемого сообщения можно отнести к физической или сигнальной, сигнальной или лингвистической компонентам.

Например, рассмотрим процесс передачи информации на примере устной речи. Процесс этот многокомпонентный (векторный). Первая компонента – физическая, т.е. для успешного осуществления процесса передачи информации необходимо наличие источника акустического сигнала (голосовых связок человека), среды для распространения акустических колебаний и приемника колебаний (уха). Вторая компонента – сигнальная: амплитудно и частотно модулированные акустические колебания. Третья компонента – синтаксическая: необходимо, чтобы собеседники знали хотя бы один общий язык. Четвертая компонента – семантическая, т.е. в передаваемом сообщении должно присутствовать содержательное описание объекта или явления, неизвестное получателю информации. Наконец, пятая компонента – прагматическая: необходимо наличие желания (мотивации) передавать и принимать сообщение.

На сложный, многокомпонентный характер информации указывал еще А.Н. Колмогоров [5]: «Подчеркну и качественно новое и неожиданное, что содержится... в теории информации. По первоначальному замыслу «информация» не есть скалярная величина. Различные виды информации могут быть чрезвычайно разнообразны... было совершенно неясно, можно ли качественно различные информации... считать эквивалентными».

Один из центральных вопросов, по которому существуют разные точки зрения, состоит в следующем: информация – это свойство объекта или резуль-

тат взаимодействия. Мы будем придерживаться точки зрения А.Н. Колмогорова: информация существует независимо от того, воспринимается она или нет, но проявляется только при взаимодействии. Информация – это характеристика внутренней организованности материальной системы по множеству состояний, которые она может принимать.

Приведем пример. По срезу дерева опытный специалист может дать заключение относительно его возраста, эволюции климатических условий, в которых развивалось дерево, и др., однако получить эту информацию он сможет лишь в результате анализа конкретного среза дерева. Другими словами, информация объективно существует независимо от нашего сознания, но выявляется при взаимодействии с конкретным объектом.

Факт объективного существования информации независимо от нашего сознания для некоторых исследователей послужил поводом для пропаганды весьма неординарной точки зрения, что информация является третьей (наряду с материей и энергией) субстанцией материального мира. Эта точка зрения наиболее уязвима, поскольку для информации пока не сформулированы фундаментальные законы сохранения и перехода в эквивалентных количествах в материю и/или энергию. Например, при сжигании дерева информация о нем, если она не была установлена и сохранена ранее, безвозвратно теряется. Тем не менее следует подчеркнуть, что информация всегда проявляется в материально-энергетической форме в виде сигналов, хотя это не материя и не энергия, которые переходят друг в друга. Информация может исчезать и появляться.

В настоящем пособии термин «информация» понимается в узком смысле, принятом при описании так называемых информационных систем [3,4,7], [10]. К ним относятся телекоммуникационные и вычислительные сети, автоматизированные системы управления и контроля и т.п. В данном случае понятие «количество информации» определяется как частота употребления знаков. Количество информации в указанном смысле не отражает ни семантики, ни прагматической ценности информации.

Информационные системы – это класс технических систем, предназначенных для хранения, передачи и преобразования информации. Соответственно информация – это сведения, являющиеся объектом хранения, передачи и преобразования, а теория информации – раздел кибернетики, занимающийся математическим описанием методов передачи, хранения, извлечения (обработки) и классификации информации. Заметим, что сама информация, как правило, используется для осуществления каких-либо управляющих воздействий.

Таким образом, предметом нашего рассмотрения является теория информации в классическом смысле – решение теоретических вопросов, касающихся повышения эффективности и функционирования информационных систем, в частности, систем связи. Она включает в себя:

1) анализ сигналов, как средства передачи информации;
2) анализ информационных характеристик источников сообщения и каналов связи;
3) теорию кодирования;
4) методы приема и обработки информации.

Каждый из указанных разделов может быть (и, как правило, является) предметом самостоятельного глубокого изучения в соответствующих дисциплинах различных специальностей информационного направления. В настоящем курсе мы стремились акцентировать внимание на наиболее общих фундаментальных законах, имеющих существенное значение для восприятия указанных разделов как единого целого. На наш взгляд, таким общим фундаментом являются теоремы К. Шеннона о кодировании и информационная теория оценивания, большой вклад в развитие которой внес Я.З. Цыпкин.

Лекция 1

Модели детерминированных сигналов

1.1 Понятие модели сигнала

Для перенесения информации в пространстве и времени она представляется в форме сообщений. Сообщение, вне зависимости от его содержания, всегда отображается в виде сигнала. Построение сигнала по определенным правилам, обеспечивающим соответствие между сообщением и сигналом, называют кодированием.

Кодирование в широком смысле – преобразование сообщения в сигнал. Кодирование в узком смысле – представление исходных знаков, называемых символами, в другом алфавите с меньшим числом знаков. Оно осуществляется с целью повышения надежности и преобразования сигналов к виду, удобному для передачи по каналам связи.

Сигналы могут быть непрерывными и дискретными как по времени, так и по множеству значений, т.е. возможен один из четырех типов сигнала:

1) непрерывный (по множеству значений и времени);
2) непрерывный по множеству значений, дискретный по времени;
3) дискретный по множеству значений, непрерывный по времени;
4) дискретный (по множеству значений и времени).

Иногда в теории связи рассматривают также сигналы непрерывные по времени и значениям, но дискретные по параметру.

Носителем сигнала всегда является объект или процесс, однако математическая модель сигнала абстрагируется от его физической природы и описывает лишь существенные с точки зрения изучаемого явления черты. Модель сигнала может даже противоречить физическим свойствам реальных объектов. Например, математическая модель сигнала в виде суммы бесконечного числа гармонических функций не может быть реализована на практике, однако эта абстракция позволяет выявить важные закономерности.

В реальных информационных системах осуществляется передача только той информации, которая не известна получателю. Поэтому можно предсказать лишь вероятность каждого сообщения, а аналитической моделью сигнала, строго говоря, может быть только случайный процесс. Тем не менее основой для изучения случайных сигналов является анализ детерминированных сигналов, рассматриваемых как элементы множества (ансамбля) реализаций. В настоящем разделе изучаются модели детерминированных сигналов.

1.2 Обобщенное спектральное представление детерминированных сигналов

Для анализа прохождения сложного сигнала $u(t)$ через линейную систему его обычно представляют в виде

$$u(t)=\sum_{k=1}^{n} c_k \varphi_k(t), \quad t \in [t_1,t_2], \tag{1.1}$$

где $\varphi_k(t)$ – так называемые базисные функции, а c_k – безразмерные коэффициенты. Если базисные функции заданы, $u(t)$ полностью определяется коэффициентами c_k, которые называют дискретным спектром сигнала. За пределами интервала $[t_1,t_2]$ сигнал (1.1) считается условно продолжающимся. При рассмотрении ряда задач такое допущение может оказаться неприемлемым.

Для представления сигналов конечной длительности используют интеграл:

$$u(t)=\int_{-\infty}^{\infty} S(\alpha)\cdot\varphi(\alpha,t)d\alpha, \tag{1.2}$$

где $S(\alpha)$ – спектральная плотность, описывающая непрерывный спектр, а $\varphi(\alpha,t)$ – базисная функция, зависящая от параметра α.

Совокупность методов, в которых используется представление сигнала в виде (1.1) и/или (1.2) называют обобщенной спектральной теорией сигналов. При этом рассматриваются частные случаи, различающиеся видом используемых базисных функций. Основное требование, обычно предъявляемое к базисным функциям, – простота вычисления коэффициентов c_k. Этому требованию

отвечают так называемые ортогональные на отрезке $[t_1,t_2]$ базисные функции, удовлетворяющие условию

$$\int_{t_1}^{t_2}\varphi_k(t)\cdot\varphi_j(t)dt=\begin{cases}0, & \text{при} \quad j\neq k,\\ \mu, & \text{при} \quad j=k.\end{cases} \tag{1.3}$$

Если умножить все $\varphi_j(t)$, $j=\overline{1,n}$ на $1/\sqrt{\mu}$, то при $j=k$

$$\int_{t_1}^{t_2}\varphi_k(t)\cdot\varphi_j(t)dt=1. \tag{1.4}$$

Такую систему функций называют ортонормированной.

Предположим, что базисные функции удовлетворяют условию (1.4). Умножим обе части (1.1) на $\varphi_j(t)$ и проинтегрируем на интервале $[t_1,t_2]$:

$$\int_{t_1}^{t_2}u(t)\cdot\varphi_j(t)\cdot dt=\int_{t_1}^{t_2}\sum_{k=1}^{n}c_k\varphi_k(t)\cdot\varphi_j(t)\cdot dt=\sum_{k=1}^{n}c_k\int_{t_1}^{t_2}\varphi_k(t)\cdot\varphi_j(t)\cdot dt.$$

В силу (1.3) все интегралы в правой части последнего равенства при $j\neq k$ равны нулю. Поэтому, с учетом (1.4), имеем

$$c_k=\int_{t_1}^{t_2}u(t)\cdot\varphi_k(t)\cdot dt. \tag{1.5}$$

Из последнего равенства видно, что коэффициенты c_k, $k=\overline{1,n}$ могут вычисляться независимо друг от друга, а сложность их вычисления определяется лишь видом аналитического выражения базисной функции. Указанное, связанное с условиями (1.3), (1.4), свойство является причиной широкого использования ортогональных функций при изучении свойств сигналов. В частности, применяются следующие системы ортогональных функций: система тригонометрических функций, система функций Хаара, полиномы Лежандра, полиномы Лаггерра, полиномы Чебышева, полиномы Эрмита и др.

1.3 Временная форма представления сигналов

Произвольную функцию (непрерывный сигнал) $u(t)$ можно представить в виде совокупности примыкающих друг к другу импульсов бесконечно малой длительности с амплитудой, равной значению сигнала в текущий момент времени:

$$u(t)=\int_{-\infty}^{\infty} u(\tau)\cdot\delta(\tau-t)\cdot d\tau, \tag{1.6}$$

где $\delta(\tau-t)$ – дельта-функция:

$$\delta(\tau-t)=\begin{cases}\infty, & \text{при} \quad t=\tau,\\ 0 & \text{при} \quad t\neq\tau.\end{cases} \qquad \int_{-\infty}^{\infty}\delta(\tau-t)\cdot d\tau=1.$$

Нетрудно заметить, что представление (1.6) является частным случаем обобщенного спектрального представления (1.2) с базисной функцией $\delta(\tau-t)$.

С помощью дельта-функции можно построить дискретную так называемую решетчатую функцию:

$$u_g(t)=\sum_{k=-\infty}^{\infty} u(t)\cdot\delta(t-k\Delta t). \tag{1.7}$$

Функция $u_g(t)$ равна $u(k\Delta t)$ в точках $t=k\Delta t$, где Δt - период следования импульсов, и нулю в остальных точках. Пределы суммирования в (1.7) так же, как в (1.1), могут быть установлены конечными, исходя из условий физической реализуемости.

1.4 Частотное представление периодических сигналов

Рассмотрим представление детерминированных сигналов с применением в качестве базисных функций $\varphi(t)=e^{pt}$, при $p=\pm j\omega$. Такое представление называется преобразованием Фурье. В силу формулы Эйлера $\cos\omega t=\left(e^{j\omega t}+e^{-j\omega t}\right)/2$ преобразование Фурье дает возможность представить сложный сигнал в виде суммы гармоник [12].

Предположим, что функция $u(t)$, описывающая детерминированную реализацию сигнала на интервале $[t_1, t_2]$, удовлетворяет условиям Дирихле (непрерывна или имеет конечное число точек разрыва первого рода, а также конечное число экстремумов) и повторяется с периодом $T = t_2 - t_1$ при $t \in (-\infty, +\infty)$. Используя указанную выше базисную функцию $\varphi(t) = e^{\pm j\omega t}$, функцию $u(t)$ можно представить в виде

$$u(t) = \frac{1}{2} \sum_{k=-\infty}^{\infty} A(jk\omega_1) \cdot e^{jk\omega_1 t}, \tag{1.8}$$

где

$$A(jk\omega_1) = \frac{2}{T} \int_{t_1}^{t_2} u(t) \cdot e^{-jk\omega_1 t} dt, \tag{1.9}$$

а период $T = t_2 - t_1 = 2\pi/\omega_1$.

Коэффициенты $A(jk\omega_1)$ в данном спектральном представлении называют комплексным спектром периодического сигнала $u(t)$, а значение $A(jk\omega_1)$ для конкретного k – комплексной амплитудой. Комплексный спектр дискретный, но путем замены $k\omega_1 = \omega$ для него можно построить огибающую:

$$A(j\omega) = \frac{2}{T} \int_{t_1}^{t_2} u(t) \cdot e^{-j\omega t} dt. \tag{1.10}$$

Как всякое комплексное число, комплексный спектр можно представить:

а) в показательной форме:

$$A(jk\omega_1) = A(k\omega_1) \cdot e^{-j\varphi(k\omega_1)}, \tag{1.11}$$

где $A(k\omega_1)$ – спектр амплитуд, а $\varphi(k\omega_1)$ – спектр фаз (также дискретный);

б) в алгебраической форме:

$$A(jk\omega_1) = A_k - jB_k, \tag{1.12}$$

где

$$A_k = \frac{2}{T} \int_{t_1}^{t_2} u(t) \cdot \cos(k\omega_1 t) dt, \qquad B_k = \frac{2}{T} \int_{t_1}^{t_2} u(t) \cdot \sin(k\omega_1 t) dt.$$

Представление (1.12) получается из (1.9) путем замены по формуле Эйлера: $e^{-jk\omega_1 t} = \cos(k\omega_1 t) - j\sin(k\omega_1 t)$. Ясно, что $A(k\omega_1) = \sqrt{A_k^2 + B_k^2}$, а $\varphi(k\omega_1) = arctg(B_k / A_k)$. Из равенства, определяющего в (1.12) вещественную часть A_k при $k = 0$, получаем равенство для постоянной составляющей сигнала:

$$\frac{A_0}{2} = \frac{1}{T}\int_{t_1}^{t_2} u(t)dt. \tag{1.13}$$

Объединяя в (1.8) комплексно-сопряженные составляющие, можно получить ряд Фурье в тригонометрической форме:

$$\begin{aligned} u(t) &= \frac{A_0}{2} + \frac{1}{2}\sum_{k=1}^{\infty}\left[A(jk\omega_1)\cdot e^{jk\omega_1 t} + A(-jk\omega_1)\cdot e^{-jk\omega_1 t}\right] = \\ &= \frac{A_0}{2} + \frac{1}{2}\sum_{k=1}^{\infty}\left[A(k\omega_1)\cdot e^{j[k\omega_1 t - \varphi(k\omega_1)]} + A(k\omega_1)\cdot e^{-j[k\omega_1 t - \varphi(k\omega_1)]}\right] = \\ &= \frac{A_0}{2} + \sum_{k=1}^{\infty} A(k\omega_1)\cos(k\omega_1 t - \varphi(k\omega_1)). \end{aligned} \tag{1.14}$$

Спектры амплитуд – $A(k\omega_1)$ и фаз – $\varphi(k\omega_1)$ могут быть представлены спектральными диаграммами в виде совокупности линий, каждая из которых соответствует определенной частоте (одному из слагаемых). Поэтому эти спектры называют линейчатыми. Сигналы, линейчатые спектры которых включают гармоники некратных частот, называются почти периодическими.

1.5 Распределение энергии в спектре периодического сигнала

В соответствии с (1.14) энергию, выделяемую периодическим сигналом за время, равное периоду T, можно представить в виде

$$\begin{aligned} &\int_0^T |u(t)|^2 dt = \int_0^T \left|\frac{A_0}{2} + \frac{1}{2}\sum_{k=1}^{\infty}\left[A(jk\omega_1)\cdot e^{jk\omega_1 t} + A(-jk\omega_1)\cdot e^{-jk\omega_1 t}\right]\right|^2 dt = \\ &= \frac{A_0^2}{4}\int_0^T dt + \frac{A_0}{2}\left\{\sum_{k=1}^{\infty} A(jk\omega_1)\int_0^T e^{jk\omega_1 t}dt + \sum_{k=1}^{\infty} A(-jk\omega_1)\int_0^T e^{-jk\omega_1 t}dt\right\} \\ &+ \frac{1}{2}\sum_{k=1}^{\infty}\sum_{l=1}^{\infty} A(jk\omega_1)A(-jl\omega_1)\int_0^T e^{j(k-l)\omega_1 t}dt. \end{aligned}$$

Можно показать, что

$$\int_0^T e^{jk\omega_1 t}dt = \int_0^T e^{-jk\omega_1 t}dt = 0, \qquad \text{а} \int_0^T e^{j(k-l)\omega_1 t}dt = \begin{cases} 0 & \text{при} \quad k \neq l, \\ T & \text{при} \quad k = l. \end{cases}$$

С учетом этого окончательно получаем

$$\int_0^T \left|u(t)\right|^2 dt = \frac{T}{2}\left[\frac{A_0^2}{2} + \sum_{k=1}^{\infty}\left|A(jk\omega_1)\right|^2\right]. \tag{1.15}$$

Из (1.15) следует, что средняя за период энергия сложного периодического сигнала равна сумме средних энергий, выделяемых каждой гармоникой.

1.6 Частотное представление непериодических сигналов

Предположим, что соответствующая реальному непериодическому сигналу функция $u(t)$ удовлетворяет условиям Дирихле и абсолютно интегрируема: $\int_{-\infty}^{\infty}\left|u(t)\right|\cdot dt < \infty$. Тогда спектральное представление непериодического сигнала $u(t)$ можно строить путем увеличения периода периодического сигнала до бесконечности. Для этого поступим следующим образом.

Подставим выражение (1.9) для комплексной амплитуды $A(jk\omega_1)$ периодического сигнала в (1.8). С учетом того, что $T = 2\pi/\omega_1$, имеем

$$u(t) = \frac{1}{2}\sum_{k=-\infty}^{\infty}\left[\frac{\omega_1}{\pi}\int_{t_1}^{t_2} u(t)\cdot e^{-jk\omega_1 t}dt\right]\cdot e^{jk\omega_1 t}.$$

Далее осуществим предельный переход при $T \to \infty$. При этом сумма переходит в интеграл, $\omega_1 = \Delta\omega \to d\omega$, $k\omega_1 \to \omega$. В результате получаем:

$$u(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} u(t)\cdot e^{-j\omega t}dt\right]\cdot e^{j\omega t}d\omega.$$

Введя в последнем равенстве для интеграла в квадратных скобках обозначение $S(j\omega)$, запишем пару преобразований Фурье:

$$u(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(j\omega)\cdot e^{j\omega t}d\omega, \tag{1.16}$$

$$S(j\omega)=\int_{-\infty}^{\infty}u(t)\cdot e^{-j\omega t}dt\,. \qquad (1.17)$$

Комплексную функцию $S(j\omega)$ называют комплексной спектральной плотностью или спектральной характеристикой. Так же, как в случае периодического сигнала, для непериодического сигнала имеют место следующие представления спектральной характеристики:

а) показательная форма:

$$S(j\omega)=S(\omega)\cdot e^{-j\varphi(\omega)}, \qquad (1.$$

$$S(j\omega)=S(\omega)\cdot e^{-j\varphi(\omega)}, \text{ где } S(\omega)=|S(j\omega)| \text{ – спектральная плотность амплитуд, а } \varphi(\omega) \text{ – спектр фаз;} \qquad (1.18)$$

где $S(\omega)=|S(j\omega)|$ – спектральная плотность амплитуд, а $\varphi(\omega)$ – спектр фаз;

б) алгебраическая форма (получается из (1.17) путем замены $e^{-j\omega t}=\cos(\omega t)-j\sin(\omega t)$):

$$S(j\omega)=A(\omega)-jB(\omega), \qquad (1.19)$$

где

$$A(\omega)=\int_{-\infty}^{+\infty}u(t)\cdot\cos(\omega t)dt,\ B(\omega)=\int_{-\infty}^{+\infty}u(t)\cdot\sin(\omega t)dt\,. \qquad (1.20)$$

При этом

$$S(\omega)=|S(j\omega)|=\sqrt{|A(\omega)|^2+|B(\omega)|^2},\ \varphi(\omega)=arctg\left[B(\omega)/A(\omega)\right]. \qquad (1.21)$$

Подставляя $S(j\omega)$ из (1.18) в (1.16), имеем

$$u(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}S(\omega)\cdot e^{j\left[\omega t-\varphi(\omega)\right]}d\omega=$$

$$\frac{1}{2\pi}\left[\int_{-\infty}^{\infty}S(\omega)\cdot\cos\left[\omega t-\varphi(\omega)\right]d\omega+j\int_{-\infty}^{\infty}S(\omega)\cdot\sin\left[\omega t-\varphi(\omega)\right]d\omega\right].$$

Второй интеграл от нечетной функции равен нулю, а первый (в силу четности подынтегральной функции) можно записать только для положительных частот. Таким образом, получаем тригонометрическую форму ряда Фурье:

$$u(t)=\frac{1}{\pi}\int_{0}^{\infty}S(\omega)\cdot\cos\left[\omega t-\varphi(\omega)\right]d\omega\,, \qquad (1.22)$$

которая дает возможность ясного физического толкования.

В заключение рассмотрим еще одно интересное свойство. Для функции $u(t)$, заданной на интервале $[t_1,t_2]$, в соответствии с (1.17), можно записать

$$S(j\omega)=\int_{t_1}^{t_2}u(t)\cdot e^{-j\omega t}dt\,. \tag{1.23}$$

Сравнивая правые части (1.10) и (1.23), нетрудно заметить, что имеет место равенство $A(j\omega)=\frac{2}{T}\cdot S(j\omega)$, т.е. по $S(j\omega)$ одиночного импульса можно построить линейчатый спектр их периодической последовательности.

1.7 Распределение энергии в спектре непериодического сигнала

Выражение для величины, характеризующей энергию, выделяемую сигналом, с учетом (1.16), можно записать в виде:

$$\int_{-\infty}^{\infty}\left[u(t)\right]^2dt=\int_{-\infty}^{\infty}u(t)\left[\frac{1}{2\pi}\int_{-\infty}^{\infty}S(j\omega)\cdot e^{j\omega t}d\omega\right]\cdot dt\,.$$

Перепишем последнее равенство, изменив порядок интегрирования:

$$\int_{-\infty}^{\infty}\left[u(t)\right]^2dt=\frac{1}{2\pi}\int_{-\infty}^{\infty}S(j\omega)\left[\int_{-\infty}^{\infty}u(t)\cdot e^{j\omega t}dt\right]\cdot d\omega\,. \tag{1.24}$$

Сравнивая правые части (1.17) и (1.24), нетрудно заметить, что выражение в квадратных скобках в (1.24) не что иное, как $S(-j\omega)$, следовательно

$$\int_{-\infty}^{\infty}\left[u(t)\right]^2dt=\frac{1}{2\pi}\int_{-\infty}^{\infty}S(j\omega)S(-j\omega)\cdot d\omega\,.$$

Теперь с учетом свойства $\left|S(j\omega)\right|^2=S(j\omega)S(-j\omega)$ можно окончательно записать так называемое равенство Парсеваля:

$$\int_{-\infty}^{\infty}\left|u(t)\right|^2dt=\frac{1}{2\pi}\int_{-\infty}^{\infty}\left|S(j\omega)\right|^2d\omega=\frac{1}{\pi}\int_{0}^{\infty}\left|S(j\omega)\right|^2d\omega\,. \tag{1.25}$$

В соответствии с этим равенством энергию, выделяемую непериодическим сигналом за время его существования, можно определить, интегрируя квадрат модуля спектральной характеристики в интервале частот.

1.8 Соотношение между длительностью сигналов и шириной их спектров

Предположим, что сигнал $u(t)$ определенной продолжительности имеет спектральную характеристику $S(j\omega)$. Найдем соответствующую характеристику $S_\lambda(j\omega)$ для сигнала $u(\lambda t)$, длительность которого изменена в λ раз:

$$S_\lambda(j\omega)=\int_{-\infty}^{\infty}u(\lambda t)\cdot e^{-j\omega t}dt=\frac{1}{\lambda}\int_{-\infty}^{\infty}u(\tau)\cdot e^{-j\frac{\omega\tau}{\lambda}}d\tau=\frac{1}{\lambda}S\left(j\frac{\omega}{\lambda}\right), \qquad (1.26)$$

где $\tau=\lambda t$.

Из (1.26) видно, что спектр укороченного (удлиненного) в λ раз сигнала в λ раз шире (уже), при этом коэффициент $1/\lambda$ изменяет только амплитуды гармоник и на ширину спектра не влияет. Указанное свойство связано с тем, что переменные t и ω входят в показатель степени экспоненциальной функции прямого и обратного преобразования Фурье в виде произведения. Из этого следует, что длительность сигнала и ширина его спектра не могут быть одновременно ограничены конечными интервалами. В частности, имеет место соотношение:

$$\Delta t\cdot\Delta f=Const,$$

где Δt – длительность импульса, Δf – ширина спектра.

Лекция 2

Модели случайных сигналов

2.1 Случайный процесс как модель сигнала

Более адекватной моделью сигнала при изучении вопросов передачи и преобразования информации является случайный процесс, для которого рассматривавшиеся выше детерминированные функции рассматриваются как отдельные реализации.

Случайным процессом называют случайную функцию времени $U(t)$, значения которой в каждый момент времени являются случайной величиной. Случайные процессы, могут быть непрерывными и дискретными как по времени, так и по множеству состояний, т.е. по аналогии с классификацией детерминированных сигналов возможен один из четырех типов случайного процесса:

1) непрерывный случайный процесс (множество состояний – континуум, а изменения состояний возможны в любой момент времени);

2) непрерывная случайная последовательность (изменения состояний допускаются лишь в конечном или счетном числе моментов времени);

3) дискретный случайный процесс (изменения состояний могут происходить в произвольные моменты времени, но множество состояний конечно);

4) дискретная случайная последовательность (состояния из конечного множества могут изменяться в конечном или счетном числе моментов времени).

Для описания свойств случайного процесса может использоваться N-мерная плотность вероятности $p_N(U_1,U_2,...,U_N;t_1,t_2,...,t_N)$ системы N случайных величин $U_1=U(t_1),U_2=U(t_2),...,U_N=U(t_N)$, взятых в моменты времени t_1, t_2,...,t_N. В частности, одномерная плотность вероятности $p_1(U;t)$ характеризует распределение случайной величины в произвольный момент времени t, а двумерная плотность $p_2(U_1,U_2;t_1,t_2)$ дает вероятность совместной реализа-

ции значений случайных величин в произвольные моменты времени t_1, t_2. Имеет место соотношение

$$p_1(U_1;t_1) = \int_{-\infty}^{\infty} p_2(U_1,U_2;t_1,t_2)\cdot dU_2 . \qquad (2.1)$$

Оперирование с плотностью вероятности, в особенности высокого порядка, чрезвычайно трудоемко. Поэтому для характеристики случайного процесса обычно используют моментные функции первого и второго порядка: математическое ожидание, дисперсию и корреляционную функцию.

Математическим ожиданием случайного процесса $U(t)$ называют неслучайную функцию времени $m_u(t)$, значение которой в каждый момент времени равно математическому ожиданию случайной величины в соответствующем сечении случайного процесса:

$$m_u(t) = M\{U(t)\} = \int_{-\infty}^{\infty} U\cdot p_1(U;t)\cdot dU , \qquad (2.2)$$

где $p_1(U;t)$ – одномерная плотность вероятности.

Дисперсией случайного процесса $U(t)$ называют неслучайную функцию времени $D_u(t)$, значение которой в каждый момент времени равно дисперсии случайной величины в соответствующем сечении случайного процесса:

$$D_u(t) = M\left\{\left[\overset{\circ}{U}(t)\right]^2\right\} = \int_{-\infty}^{\infty}\left[U(t)-m_u(t)\right]^2\cdot p_1(U;t)\cdot dU , \qquad (2.3)$$

где $\overset{\circ}{U}(t) = U(t) - m_u(t)$ – центрированная случайная величина в сечении t.

Корреляционной (автокорреляционной) функцией случайного процесса $U(t)$ называют неслучайную функцию $R_u(t_1,t_2)$ двух аргументов, которая для каждой пары произвольно выбранных значений t_1, t_2 равна корреляционному моменту соответствующих сечений случайного процесса:

$$R_u(t_1,t_2) = M\left\{\overset{\circ}{U}(t_1)\overset{\circ}{U}(t_2)\right\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\overset{\circ}{U}(t_1)\overset{\circ}{U}(t_2)\cdot p_2(U_1,U_2;t_1,t_2)\cdot dU_1 dU_2 , \qquad (2.4)$$

где $\overset{\circ}{U}(t_1)=\left[U(t_1)-m_u(t_1)\right]$, $\overset{\circ}{U}(t_2)=\left[U(t_2)-m_u(t_2)\right]$.

Часто во многих отношениях удобнее использовать нормированную автокорреляционную функцию:

$$\rho_u(t_1,t_2)=R_u(t_1,t_2)\Big/\left(\sigma_u(t_1)\cdot\sigma_u(t_2)\right), \tag{2.5}$$

где $\sigma_u(\cdot)=\sqrt{D_u(\cdot)}$. При произвольном $t=t_1=t_2$ автокорреляционная функция (2.4) вырождается в дисперсию (2.3): $R_u(t_1,t_2)=D_u(t)$, а соответствующая нормированная автокорреляционная функция (2.5) равна единице.

Для характеристики связи между двумя случайными процессами, например $U(t)$ и $V(t)$, рассматривают также функцию взаимной корреляции:

$$R_{uv}(t_1,t_2)=M\left[\overset{\circ}{U}(t_1)\overset{\circ}{V}(t_2)\right]. \tag{2.6}$$

С точки зрения изменчивости указанных характеристик во времени различают стационарные и нестационарные случайные процессы. Процесс $U(t)$ называют стационарным в узком смысле, если описывающие его плотности вероятности не зависят от начала отсчета времени.

Случайный процесс называют стационарным в широком смысле, если

$$m_u(t)=m_u=Const, \tag{2.7}$$

$$D_u(t)=D_u=Const, \tag{2.8}$$

$$R_u(t,t+\tau)=R_u(\tau), \tag{2.9}$$

т.е. математическое ожидание (2.2) и дисперсия (2.3) постоянны, а корреляционная функция не зависит от начала отсчета времени и является функцией одного аргумента $\tau=t_2-t_1$. Легко заметить, что условие постоянства дисперсии (2.8) как частный случай вытекает из требования к корреляционной функции (2.9) при $\tau=0$: $D_u(t)=R_u(t,t)=R_u(0)=Const$.

Обычно предполагается, что стационарный процесс является эргодичным, т.е. среднее по ансамблю реализаций равно среднему по времени на одной длинной реализации:

$$m_u = \lim_{T\to\infty}\frac{1}{T}\int_0^T u(t)\cdot dt = u_0\,, \tag{2.10}$$

$$D_u = \lim_{T\to\infty}\frac{1}{T}\int_0^T \left[u(t)-u_0\right]^2\cdot dt\,, \tag{2.11}$$

$$R_u(\tau) = \lim_{T\to\infty}\frac{1}{T}\int_0^T \left[u(t)-u_0\right]\cdot\left[u(t+\tau)-u_0\right]\cdot dt\,, \tag{2.12}$$

где $u(t)$ – некоторая реализация случайного процесса $U(t)$.

2.2 Спектральное представление случайных сигналов

Подобно детерминированным сигналам, случайный процесс может быть представлен в виде суммы спектральных составляющих. Для этого используется так называемое каноническое разложение случайных процессов $U(t)$ в виде

$$U(t) = m_u(t) + \sum_k C_k\cdot\varphi_k(t), \tag{2.13}$$

где $m_u(t)$ – математическое ожидание случайного процесса (2.2), $\varphi_k(t)$ – неслучайные базисные (координатные) функции, а C_k – некоррелированные случайные величины с математическими ожиданиями, равными нулю, и дисперсией D_k, т.е.

$$M\left[C_kC_l\right] = \begin{cases} D_k & \text{при} \quad k = l, \\ 0 & \text{при} \quad k \neq l. \end{cases} \tag{2.14}$$

Слагаемые $C_k\cdot\varphi_k(t)$ называют элементарными случайными процессами. Случайность такого процесса проявляется через случайную величину C_k, которую называют коэффициентом канонического разложения.

Найдем корреляционную функцию случайного процесса $U(t)$, представленного каноническим разложением (2.13):

$$R_u(t_1,t_2) = M\left[\mathring{U}(t_1)\cdot\mathring{U}(t_2)\right] = M\left[\sum_k C_k\varphi_k(t_1)\sum_l C_l\varphi_l(t_2)\right] =$$

$$= \sum_{k,l} M\left[C_kC_l\right]\varphi_k(t_1)\varphi_l(t_2).$$

Поскольку по предположению C_k, C_l – некоррелированны, с учетом условий (2.14) выражение для корреляционной функции принимает вид

$$R_u(t_1,t_2)=\sum_k \varphi_k(t_1)\cdot\varphi_k(t_2)\cdot D_k. \qquad (2.15)$$

Представление корреляционной функции в виде суммы (2.15) называют каноническим разложением корреляционной функции случайного процесса $U(t)$.

Доказано [8], что всякому каноническому разложению случайного процесса (2.13) соответствует каноническое разложение корреляционной функции (2.15). Справедливо и обратное утверждение: всякому разложению корреляционной функции вида (2.15) соответствует каноническое разложение центрированного случайного процесса.

Полагая в выражении (2.15) $t_1=t_2=t$, получим формулу для дисперсии случайного процесса:

$$D_u(t)=R_u(t,t)=\sum_k D_k\cdot\left[\varphi_k(t)\right]^2. \qquad (2.16)$$

Таким образом, при выбранном наборе координатных функций центрированный случайный процесс характеризуется совокупностью дисперсий коэффициентов разложения, которую можно рассматривать как обобщенный спектр случайного процесса.

Для построения представлений (2.13), (2.15) и/или (2.16) необходимо найти координатные функции $\varphi_k(t)$ некоррелированных случайных величин C_k, что во многих случаях представляет значительные трудности.

Если $\varphi_k(t)$ – ортогональные координатные функции, а $\int_{-T/2}^{T/2} m_u^2(t)\cdot dt<\infty$, неслучайную функцию $m_u(t)$ на интервале T также можно разложить по аналогии с (1.1):

$$m_u(t)=\sum_k m_{uk}\varphi_k(t), \qquad (2.17)$$

где

$$m_{uk} = \int_{-T/2}^{T/2} m_u(t) \cdot \varphi_k(t) dt .$$

Подставляя $m_u(t)$ из (2.17) в (2.13) для случайного процесса $U(t)$ с отличным от нуля средним значением, каноническое разложение получаем в виде

$$U(t) = \sum_k (m_{uk} + C_k) \cdot \varphi_k(t) . \qquad (2.18)$$

Соотношение (2.18) может рассматриваться как обобщенное спектральное представление типа (1.1) для случайного сигнала.

2.3 Частотное представление стационарных случайных сигналов, дискретные спектры

Предположим, что случайный процесс задан на конечном интервале времени $[-T, T]$. Тогда соответствующая корреляционная функция $R_u(\tau)$ должна рассматриваться на интервале $4T$, т.к. при $-T < t_1, t_2 < T$ должны выполняться неравенства $-2T < \tau < 2T$.

Считая $R_u(\tau)$ условно продолжающейся с периодом $4T$, можно записать

$$R_u(\tau) = \frac{1}{2} \sum_{k=-\infty}^{\infty} D_k \cdot e^{jk\omega_1 \tau} , \qquad (2.19)$$

$$D_k = \frac{1}{2T} \int_{-2T}^{2T} R_u(\tau) \cdot e^{-jk\omega_1 \tau} d\tau \quad (k = 0, \pm 1, \pm 2, \ldots), \qquad (2.20)$$

где

$$\omega_1 = 2\pi/(4T) = \pi/(2T) .$$

С учетом того, что $R_u(\tau)$ – четная функция, (2.20) можно представить в виде

$$D_k = \frac{1}{T} \int_0^{2T} R_u(\tau) \cdot e^{-jk\omega_1 \tau} d\tau .$$

Положив в (2.19) $\tau = t_1 - t_2$, можно записать

$$R_u(t_1 - t_2) = \frac{1}{2} \sum_{k=-\infty}^{\infty} D_k \cdot e^{jk\omega_1 t_1} e^{-jk\omega_1 t_2} . \qquad (2.21)$$

Сравнивая последнее выражение с (2.15), нетрудно заметить, что (2.21) – суть каноническое разложение корреляционной функции. Как указывалось ранее, ему соответствует каноническое разложение центрированного случайного процесса:

$$\mathring{U}(t)=\frac{1}{2}\sum_{k=-\infty}^{\infty}C_k\cdot e^{jk\omega_1 t}, \tag{2.22}$$

где

$$C_k:\ M\left\{\left[C_k\right]^2\right\}=D_k.$$

В общем случае в правую часть (2.22) необходимо добавить математическое ожидание стационарного случайного процесса – m_u.

При объединении экспоненциальных составляющих с одинаковыми по абсолютной величине индексами разных знаков стационарный случайный процесс на ограниченном интервале времени представляется суммой гармоник:

$$U(t)=m_u+\sum_{k=1}^{\infty}\left(a_k\cos k\omega_1 t+b_k\sin k\omega_1 t\right), \tag{2.23}$$

где

$$\omega_1=\pi/(2T),\ m_u=M\left\{U(t)\right\},$$

$$M\left\{a_k\right\}=M\left\{b_k\right\}=0,\ M\left\{\left[a_k\right]^2\right\}=M\left\{\left[b_k\right]^2\right\}=D_k.$$

Из представления спектрального разложения в тригонометрической форме (2.23) видно, что получающиеся спектры являются линейчатыми, т.е. каждой гармонике на спектральной диаграмме будет соответствовать вертикальный отрезок, длина которого пропорциональна дисперсии амплитуды D_k.

2.4 Частотное представление стационарных случайных сигналов, непрерывные спектры

Для описания стационарного случайного процесса при любом $-\infty<t<\infty$ построим интегральное каноническое разложение. Для этого несколько изменим формулу (2.19):

$$R_u(\tau)=\frac{1}{2}\sum_{k=-\infty}^{\infty}\frac{D_k}{\Delta\omega}e^{jk\omega_1\tau}\Delta\omega\,, \tag{2.24}$$

где $\Delta\omega=\omega_{k+1}-\omega_k=\pi/(2T)$ – интервал частот между соседними гармониками.

Обозначим

$$S_u(k\omega_1)=\frac{D_k}{\Delta\omega}=\frac{2T}{\pi}D_k\,. \tag{2.25}$$

Функцию $S_u(k\omega_1)$ называют средней плотностью дисперсии стационарного процесса. Это дисперсия, приходящаяся на единицу длины частотного интервала между соседними гармониками.

С учетом обозначения (2.25) формула (2.24) примет вид

$$R_u(\tau)=\frac{1}{2}\sum_{k=-\infty}^{\infty}S_u(k\omega_1)\cdot e^{jk\omega_1\tau}\Delta\omega\,. \tag{2.26}$$

Подставляя в (2.25) выражение для D_k из (2.20), можно также записать

$$S_u(k\omega_1)=\frac{1}{\pi}\int_{-2T}^{2T}R_u(\tau)\cdot e^{-jk\omega_1\tau}d\tau\,. \tag{2.27}$$

Далее осуществим в (2.26), (2.27) предельный переход при $T\to\infty$. При этом сумма переходит в интеграл, $S_u(k\omega_1)\to S_u(\omega)$, $k\omega_1\to\omega$, $\Delta\omega\to d\omega$. В результате получаем:

$$R_u(\tau)=\frac{1}{2}\int_{-\infty}^{\infty}S_u(\omega)\cdot e^{j\omega\tau}d\omega\,, \tag{2.28}$$

$$S_u(\omega)=\frac{1}{\pi}\int_{-\infty}^{\infty}R_u(\tau)\cdot e^{-j\omega\tau}d\tau\,. \tag{2.29}$$

Величина $S_u(\omega)\cdot d\omega$, фигурирующая в (2.28), по смыслу введенного обозначения (2.25) представляет собой дисперсию, приходящуюся на спектральные составляющие в интервале частот $[\omega,\omega+d\omega]$. Функцию $S_u(\omega)$, характеризующую распределение дисперсии случайного процесса по частотам, называют *спектральной плотностью* стационарного случайного процесса.

По аналогии с (2.21) выражение для интегрального канонического разложения корреляционной функции $R_u(\tau)$ можно записать, положив в (2.28) $\tau = t_1 - t_2$:

$$R_u(\tau) = \frac{1}{2}\int_{-\infty}^{\infty} S_u(\omega)\cdot e^{j\omega t_1} e^{-j\omega t_2} d\omega . \qquad (2.30)$$

Подобно разложению корреляционной функции по той же схеме можно построить разложение случайного процесса. Для этого формулу (2.22) представим в виде

$$\overset{\circ}{U}(t) = \frac{1}{2}\sum_{k=-\infty}^{\infty} \frac{C_k}{\Delta\omega}\cdot e^{jk\omega_1 t}\Delta\omega .$$

Далее введем обозначение $G_u(\omega_k) = C_k/\Delta\omega$ и подобно тому, как мы это сделали в (2.26), (2.27), осуществим предельный переход при $T \to \infty$. В результате получим каноническое разложение стационарной случайной функции:

$$\overset{\circ}{U}(t) = \frac{1}{2}\int_{-\infty}^{\infty} G_u(\omega)\cdot e^{j\omega t} d\omega . \qquad (2.31)$$

В силу отмечавшегося выше соответствия между разложением (2.21) корреляционной функции и разложением (2.22) случайного процесса очевидно, что $G_u(\omega)d\omega$ в (2.31) является случайной функцией с дисперсией $S_u(\omega)d\omega$, приходящейся на спектральные составляющие в интервале частот $(\omega, \omega + d\omega)$.

2.5 Спектральная плотность мощности

Перейдем к одностороннему спектру для положительных частот. С использованием формулы Эйлера представим (2.29) в виде двух слагаемых:

$$S_u(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} R_u(\tau)\cdot\cos\omega\tau d\tau - \frac{j}{\pi}\int_{-\infty}^{\infty} R_u(\tau)\cdot\sin\omega\tau d\tau .$$

Поскольку $R_u(\tau)$ – четная функция, второе слагаемое равно нулю, а первый интеграл можно записать для положительных частот:

$$S_u(\omega) = \frac{2}{\pi}\int_{0}^{\infty} R_u(\tau)\cdot\cos\omega\tau d\tau . \qquad (2.32)$$

Отсюда, в частности, следует, что $S_u(\omega)$ также действительная и четная функция. Следовательно, в (2.28) также можно ограничиться положительными частотами:

$$R_u(\tau) = \int_0^\infty S_u(\omega) \cdot \cos\omega\tau \cdot d\omega .$$

Положив в последнем равенстве $\tau = 0$, получаем

$$R_u(0) = D_u = \int_0^\infty S_u(\omega) \cdot d\omega . \qquad (2.33)$$

Поскольку дисперсия характеризует мощность сигнала:

$$D_u = M\left\{\left[\overset{\circ}{U}\right]^2\right\} = P_u ,$$

спектральную плотность $S_u(\omega)$ часто называют спектральной плотностью мощности.

Лекция 3

Преобразование непрерывных сигналов в дискретные

3.1 Формулировка задачи дискретизации

Дискретизация сигнала – это преобразование функции непрерывного аргумента в функцию дискретного времени. Она заключается в замене непрерывного сигнала $u(t)$ совокупностью координат:

$$[c_1, c_2, ..., c_N] = A[u(t)], \tag{3.1}$$

где $A[\cdot]$ – некоторый оператор.

С точки зрения простоты реализации целесообразно использовать линейные операторы. В частности, для определения координат сигнала удобно использовать соотношение

$$c_i = Au(t) = \int_T \varphi_i(t) \cdot u(t) \cdot dt, \quad i = \overline{1, N}, \tag{3.2}$$

где $\varphi_i(t), \quad i = \overline{1, N}$ – заданные базисные (в частности, могут использоваться ортогональные) функции.

При последующем использовании дискретного сигнала для целей управления обычно осуществляют его восстановление с использованием некоторого заданного оператора:

$$u^*(t) = B[c_1, c_2, ..., c_N], \tag{3.3}$$

Если дискретизация осуществлялась оператором вида (3.2), для восстановления непрерывного сигнала в соответствии с (1.1) может использоваться оператор

$$u^*(t) = \sum_{i=1}^{N} c_i \varphi_i(t). \tag{3.4}$$

Дискретизация по соотношению (3.2), вследствие применения операции интегрирования, обладает высокой помехоустойчивостью. Однако при этом имеет место задержка сигнала на время интегрирования T. Поэтому чаще дискретизация сводится к замене сигнала совокупностью его мгновенных значений (выборок) $u(t_i), \quad i = 1, 2, \ldots$, которая описывается соотношением (1.6).

Это достигается использованием в (3.2) дельта-функции: $\varphi_i(t) = \delta(t - t_i)$. В результате получается решетчатая функция (1.7), а координаты c_i сигнала определяются как

$$c_i = u(t_i). \tag{3.5}$$

Если шаг дискретизации $\Delta t_i = t_i - t_{i-1} = Const$ – дискретизация называется равномерной.

При восстановлении непрерывного сигнала по выборкам для обеспечения простоты реализации устройств широко применяются неортогональные базисные функции, в частности, используются степенные алгебраические полиномы вида

$$u^*(t) = \sum_{i=0}^{N} a_i \cdot t^i \quad \text{или} \quad u^*(t) = \sum_{i=0}^{N} a_i \cdot (t - t_0)^i,$$

где a_i – действительные коэффициенты.

Представление непрерывного сигнала совокупностью равноотстоящих отсчетов – наиболее распространенный вид дискретизации. Обычно она осуществляется с целью дальнейшего преобразования сигнала в цифровую форму. В результате цифрового кодирования дискретного сигнала происходит его квантование – замена в соответствующие моменты времени мгновенных значений сигнала ближайшими разрешенными. При этом сигнал оказывается дискретным как по времени, так и по множеству значений.

Важное достоинство цифровой формы представления сигнала состоит в том, что много уровней квантования можно представить небольшим количеством разрядов. Кроме того, при представлении в цифровой форме могут быть реализованы сложные алгоритмы обработки на ЭВМ, включая построение кодов, обнаруживающих и исправляющих ошибки.

3.2 Критерии качества восстановления непрерывного сигнала

Для оценки качества восстановления сигнала используются следующие критерии.

Равномерное приближение (критерий наибольшего отклонения):

$$\max_{t\in T}\left|u(t)-u^{*}(t)\right|\le\varepsilon_{доп}. \tag{3.6}$$

Равномерное приближение для ансамбля реализаций:

$$\sup_{u_i(t)\in U}\left|u_i(t)-u_i^{*}(t)\right|\le\varepsilon_{доп}. \tag{3.7}$$

Критерий среднеквадратического отклонения (СКО):

$$\sigma=\sqrt{\frac{1}{T}\int_{T}\left|u(t)-u^{*}(t)\right|^{2}dt}\le\sigma_{доп}. \tag{3.8}$$

СКО для ансамбля N реализаций – σ_{Σ} вычисляется усреднением по ансамблю с учетом вероятностей реализаций p_i, $i=\overline{1,N}$:

$$\sigma_{\Sigma}=\sum_{i=1}^{N}p_i\sigma_i\le\sigma_{\Sigma,доп}. \tag{3.9}$$

Интегральный критерий:

$$\varepsilon=\frac{1}{T}\int_{T}\left|u(t)-u^{*}(t)\right|dt\le\varepsilon_{доп}. \tag{3.10}$$

Величину интегрального критерия ε_{Σ} для N реализаций вычисляют путем усреднения по ансамблю:

$$\varepsilon_{\Sigma}=\sum_{i=1}^{N}p_i\varepsilon_i. \tag{3.11}$$

Применяют также вероятностный критерий, определяемый как допустимый уровень вероятности $\mathrm{P}_{доп}$ того, что ошибка не превысит допустимого значения $\varepsilon_{доп}$:

$$\mathrm{P}\left\{\left|u(t)-u^{*}(t)\right|\le\varepsilon_{доп}\right\}\le\mathrm{P}_{доп}. \tag{3.12}$$

Использование одного из указанных критериев (3.6)-(3.12) в каждом конкретном случае зависит от требований к системе и доступной априорной информации.

3.3 Теорема Котельникова

Как отмечалось выше, наиболее широко используется равномерная дискретизация. При этом для выбора величины шага дискретизации используется модель сигнала в виде эргодического случайного процесса, каждая реализация которого представляет собой функцию с ограниченным спектром. Теоретической основой этого подхода является следующая теорема Котельникова.

Любая функция $u(t)$, допускающая преобразование Фурье и имеющая непрерывный спектр, ограниченный полосой частот от 0 до $f_c = \omega_c / 2\pi$, полностью определяется дискретным рядом своих мгновенных значений, отсчитанных через интервалы времени $\Delta t = 1/(2 \cdot f_c) = \pi / \omega_c$.

Доказательство. Поскольку по предположению функция $u(t)$ имеет ограниченный спектр, т.е. $S(j\omega) = 0$ при $|\omega| > \omega_c$, в соответствии с (1.16) можно записать равенство

$$u(t) = \frac{1}{2\pi} \int_{-\omega_c}^{+\omega_c} S(j\omega) \cdot e^{j\omega t} d\omega . \qquad (3.13)$$

Функцию $S(j\omega)$ на *конечном интервале* $[-\omega_c, \omega_c]$ можно разложить в ряд Фурье. Пару преобразований Фурье запишем, полагая $S(j\omega)$ условно продолжающейся с периодом $2\omega_c$ и формально заменив в (1.8), (1.9) t на ω, а ω_1 на $\Delta t = \pi / \omega_c$:

$$S(j\omega) = \frac{1}{2} \sum_{-\infty}^{+\infty} A_k \cdot e^{jk\Delta t\omega} , \qquad (3.14)$$

$$A_k = \frac{1}{\omega_c} \int_{-\omega_c}^{\omega_c} S(j\omega) \cdot e^{-jk\Delta t\omega} d\omega . \qquad (3.15)$$

Сравним соотношения (3.15) и (3.13), предварительно переписав равенство (3.13) для дискретных моментов времени $t_k = k\Delta t$:

$$u(k\Delta t) = \frac{1}{2\pi} \int_{-\omega_c}^{\omega_c} S(j\omega) \cdot e^{j\omega k\Delta t} d\omega . \qquad (3.16)$$

Нетрудно заметить, что

$$A_k = \frac{2\pi}{\omega_c} \cdot u(-k\Delta t). \tag{3.17}$$

Подставляя значение A_k из (3.17) в (3.14), можно записать:

$$S(j\omega) = \frac{\pi}{\omega_c} \sum_{-\infty}^{+\infty} u(-k\Delta t) \cdot e^{jk\Delta t\omega}.$$

В последнем равенстве знак минус перед k можно поменять на обратный, т.к. суммирование ведется как по положительным, так и по отрицательным числам:

$$S(j\omega) = \frac{\pi}{\omega_c} \sum_{-\infty}^{+\infty} u(k\Delta t) \cdot e^{-jk\Delta t\omega}. \tag{3.18}$$

Теперь подставим $S(j\omega)$ из (3.18) в (3.13):

$$u(t) = \frac{1}{2\omega_c} \int_{-\omega_c}^{+\omega_c} \left(\sum_{-\infty}^{+\infty} u(k\Delta t) \cdot e^{-jk\Delta t\omega} \right) \cdot e^{j\omega t} d\omega = \frac{1}{2\omega_c} \sum_{-\infty}^{+\infty} u(k\Delta t) \int_{-\omega_c}^{+\omega_c} e^{j\omega(t-k\Delta t)} d\omega.$$

После выполнения интегрирования в правой части последнего равенства получаем

$$u(t) = \sum_{-\infty}^{+\infty} u(k\Delta t) \frac{\sin \omega_c (t - k\Delta t)}{\omega_c (t - k\Delta t)} = \sum_{-\infty}^{+\infty} u(k\Delta t) \operatorname{sinc} \omega_c (t - k\Delta t). \tag{3.19}$$

Итак, мы выразили функцию $u(t)$ через ее дискретные значения, взятые в моменты времени $t_k = k\Delta t$. Предположим $t = n\Delta t$, где n – некоторое целое число. Поскольку $\Delta t = \pi/\omega_c$, для любых целых k и n

$$\omega_c (n\Delta t - k\Delta t) = (n-k)\omega_c \Delta t = (n-k)\pi.$$

Следовательно,

$$\frac{\sin \omega_c (t - k\Delta t)}{\omega_c (t - k\Delta t)} = \begin{cases} 1, & \text{если} \quad t = k\Delta t, \\ 0, & \text{если} \quad t = n\Delta t, \quad n \neq k. \end{cases}$$

Это означает, что значения функции $u(t)$ в моменты времени $t_k = k\Delta t$ представляют собой не что иное, как ее отсчеты. Таким образом, функция с ограниченным спектром может быть представлена рядом (3.19), коэффициенты

которого представляют собой отсчеты значений функции, взятые через интервалы времени

$$\Delta t = \frac{\pi}{\omega_c} = \frac{1}{2 \cdot f_c} \ . \qquad (3.20)$$

На основании этого можно представить следующую схему передачи-приема. На передающей стороне мгновенные значения сигнала $u(t)$ передаются через интервалы Δt, определяемые по соотношению (3.20). На приемной стороне последовательность импульсов пропускают через идеальный фильтр нижних частот с частотой среза f_c. Тогда при длительной передаче теоретически сигнал на выходе фильтра будет точно воспроизводить переданный непрерывный сигнал $u(t)$.

В действительности реальный сигнал всегда имеет конечную длительность, следовательно, его спектр неограничен. Ошибка возникает не только за счет принудительного ограничения спектра, но и за счет конечного числа отсчетов в интервале времени T, которых в соответствии с теоремой будет $N = 2 f_c T$.

Модель сигнала с ограниченным спектром имеет также принципиальное теоретическое неудобство. Она не может отражать основное свойство сигнала – способность нести информацию. Дело в том, что поведение функции с ограниченным спектром можно точно предсказать на всей оси времени, если она точно известна на сколь угодно малом отрезке времени.

Тем не менее теорема Котельникова имеет важное прикладное значение. На практике ширину спектра f_c определяют как интервал частот, вне которого спектральная плотность меньше некоторой заданной величины. При таком допущении функция на интервале T с некоторой степенью точности (зависящей от точности представления спектральной плотности) определяется посредством $N = 2 f_c T$ отсчетов, т.е. общий смысл теоремы Котельникова сохраняется.

3.4 Квантование сигналов

Физически реализуемый непрерывный сигнал $u(t)$ всегда ограничен некоторым диапазоном $[u_{\min}, u_{\max}]$. Вдобавок часто устройство может воспроизводить лишь конечное множество фиксированных значений сигнала из этого диапазона. В частности, непрерывная шкала мгновенных значений $u_n = u_{\max} - u_{\min}$ может быть разбита на n одинаковых интервалов, а разрешенные значения сигнала равноотстоят друг от друга, тогда говорят о равномерном квантовании. Если постоянство интервала (шага квантования) не соблюдается, то квантование неравномерное.

Из множества мгновенных значений, принадлежащих i-му интервалу (шагу квантования), только одно значение u_i' является разрешенным (i-й уровень квантования), а любое другое округляется до u_i'. Предположим, равномерное квантование с шагом $\Delta = (u_{\max} - u_{\min})/n$ осуществляется так, что уровни квантования u_i' размещаются в середине каждого шага. Ясно, что при этом ошибка квантования минимальна и не превышает $0{,}5\Delta$. Определим для этого случая среднеквадратическое отклонение (СКО) ошибки квантования.

В общем случае СКО ошибки квантования σ_i для i-го шага определяется соотношением

$$\sigma_i = \sqrt{\int_{u_{i-1}}^{u_i} \left(u(t) - u_i'\right)^2 p(u)\,du}\,, \qquad (3.21)$$

где $p(u)$ – функция плотности вероятности мгновенных значений сигнала U. Если шаги квантования малы по сравнению с диапазоном изменения сигнала, плотность $p(u)$ в пределах каждого шага можно считать постоянной и равной, например, $p(u_i')$. Тогда, вводя новую переменную $y = u(t) - u_i'$, для указанного способа квантования в соответствии с (3.21) имеем

$$\sigma_i = \sqrt{p\left(u_i^{'}\right)\int\limits_{-\frac{\Delta_i}{2}}^{\frac{\Delta_i}{2}} y_i^2 dy_i} = \sqrt{p\left(u_i^{'}\right)\frac{\Delta_i^3}{12}}\,. \qquad (3.22)$$

С учетом того, что $p\left(u_i^{'}\right) > 0$ и $\Delta_i > 0$ для всех $i = \overline{1,n}$, в соответствии с (3.22) можно записать дисперсию ошибки квантования на i-м шаге:

$$\sigma_i^2 = \left[p\left(u_i^{'}\right)\Delta_i\right]\frac{\Delta_i^2}{12}\,. \qquad (3.23)$$

Оказывается, она равна величине $\Delta_i^2/12$, умноженной на вероятность $p\left(u_i^{'}\right)\Delta_i$ попадания мгновенного значения сигнала в данный интервал. Дисперсия полной ошибки определяется как математическое ожидание дисперсий $\Delta_i^2/12$ на отдельных шагах:

$$\sigma^2 = \sum_{i=1}^{n}\left[p\left(u_i^{'}\right)\Delta_i\right]\frac{\Delta_i^2}{12}\,.$$

Если интервалы одинаковы, т.е. $\Delta_i = \Delta$ для всех $i = \overline{1,n}$, с учетом условия нормировки $\sum_{i=1}^{n}\left[p\left(u_i^{'}\right)\Delta\right] = 1$, получаем

$$\sigma^2 = \frac{\Delta^2}{12}\sum_{i=1}^{n}\left[p\left(u_i^{'}\right)\Delta\right] = \frac{\Delta^2}{12}\,.$$

Если на квантуемый сигнал воздействует помеха, он может попасть в интервал, соответствующий другому уровню квантования. Интуитивно ясно (и это можно строго показать), что в случае, когда помеха ξ имеет равномерное распределение $p(\xi) = 1/a$, где $a/2$ – амплитуда помехи, симметричной относительно мгновенного значения сигнала, вероятность неправильного квантования сигнала резко возрастает при $a > \Delta$. Воздействие нормально распределенной помехи с параметрами $\left(0,\sigma^2\right)$ эквивалентно воздействию равномерно распределенной помехи при $a = 3\sigma$.

Лекция 4

Меры неопределенности дискретных множеств

4.1 Вероятностное описание дискретных ансамблей

Пусть $Z=\{z_1,z_2,z_3,\ldots,z_N\}$ – множество, состоящее из N элементов. Говорят, что на множестве Z задано распределение вероятностей $p(z)$, если каждому z_i поставлено в соответствие число $p(z_i)$, такое, что для всех $i=\overline{1,N}$ $p(z_i)\geq 0$, а $\sum p(z_i)=1$. Множество Z вместе с заданным на нём распределением вероятностей называется дискретным вероятностным ансамблем или просто дискретным ансамблем и обозначается $\{Z,p(z)\}$.

Пусть $Z=\{z_1,z_2,...,z_N\}$ и $V=\{v_1,v_2,...,v_K\}$ – два конечных множества. Произведением множеств $\{ZV\}$ называется множество, элементы которого представляют собой все возможные упорядоченные пары произведений z_iv_j, $i=\overline{1,N}$, $j=\overline{1,K}$. Если каждой паре z_i,v_j поставлена в соответствие вероятность $p(z_i,v_j)$, то имеем произведение ансамблей $\{ZV,p(zv)\}$. Для элементов объединенного ансамбля имеют место обычные свойства вероятностей:

$$\sum_{j=1}^{K}p(z_i,v_j)=p(z_i),\qquad \sum_{i=1}^{N}p(z_i,v_j)=p(v_j). \tag{4.1}$$

Из указанных свойств, в частности, следует, что если задано произведение ансамблей, то всегда могут быть найдены исходные ансамбли $\{Z,p(z)\}$ и $\{V,p(v)\}$. Обратное возможно лишь в случае, когда элементы исходных ансамблей независимы, при этом $p(z_i,v_j)=p(z_i)p(v_j)$. В общем случае для зависимых ансамблей $p(z_i,v_j)=p(z_i)p(v_j/z_i)=p(v_j)p(z_i/v_j)$, т.е. для определения вероятности элемента объединенного ансамбля необходимо задание условной вероятности появления элемента одного из ансамблей, при условии, что реализовался элемент другого ансамбля:

$$p(z_i / v_j) = \frac{p(z_i, v_j)}{p(v_j)}, \qquad p(v_j / z_i) = \frac{p(z_i, v_j)}{p(z_i)}. \tag{4.2}$$

4.2 Энтропия, как мера неопределенности выбора

Пусть задан дискретный ансамбль с N возможными состояниями:

$$Z = \begin{bmatrix} z_1, z_2, \ldots, z_i, \ldots, z_N \\ p_1, p_2, \ldots, p_i, \ldots, p_N \end{bmatrix}, \; p_i = p(z_i) \geq 0, \; \sum p_i = 1. \tag{4.3}$$

Интуитивно ясно, чем больше величина N, тем больше неопределенность выбора конкретного элемента ансамбля. Это наталкивает на мысль принять число N в качестве меры неопределенности выбора. Однако при $N = 1$ неопределенность выбора равна 0, хотя мера отлична от нуля. По-видимому, это неудобство послужило одной из причин введения следующей меры неопределенности:

$$H(Z) = \log_a N. \tag{4.4}$$

Мера предложена Р. Хартли в 1928 г. Свойства меры Хартли:

1) она является монотонной функцией числа элементов;
2) при $N = 1$ $H(Z) = 0$, т.е. мера равна нулю, когда неопределенность отсутствует;
3) мера аддитивна, т.е. объединение, например, двух множеств Z и V с числом элементов N и M, можно рассматривать как одно множество, включающее $N \times M$ различных комбинаций $z_i v_j$, $i = \overline{1, N}$, $j = \overline{1, M}$, при этом

$$H(ZV) = \log_a (NM) = \log_a N + \log_a M.$$

К сожалению, мера Р. Хартли не учитывает того факта, что вероятности p_i, $i = \overline{1, N}$ в (4.3) могут быть различны. Поэтому она используется лишь в случае равновероятных элементов множества. При неравновероятных элементах неопределенность меньше. Например, неопределенность выбора в случае двух элементов с априорными вероятностями 0,9 и 0,1 меньше, чем в случае равновероятных элементов (0,5; 0,5). Поэтому естественным является требование, чтобы мера неопределенности была непрерывной функцией вероятностей p_i,

$i=\overline{1,N}$ элементов. Удовлетворяющая этому требованию мера предложена К. Шенноном и называется энтропией:

$$H(Z)=-\sum_{i=1}^{N}p(z_i)\log_a p(z_i). \qquad (4.5)$$

Основание a логарифма, вообще говоря, не имеет значения. Если логарифм десятичный ($\lg$), энтропия и количество информации определяются в десятичных единицах *дитах*, если логарифм натуральный ($\ln$), единицей измерения является *нит*. Наиболее широко используется двоичная единица информации – *bit* (сокращение от английского *binary digit*), соответствующая логарифму по основанию два ($\log_2$), которая и будет использоваться далее.

Для независимо реализуемых элементов множества в качестве меры может использоваться априорная частная неопределенность:

$$H(z_i)=-\log_2 p(z_i). \qquad (4.6)$$

Нетрудно заметить, что мера К. Шеннона (4.5), характеризующая неопределённость источника в целом, получается усреднением частных неопределенностей (4.6) по всем элементам множества.

Покажем связь меры К. Шеннона с мерой Р. Хартли. Если все элементы множества равновероятны, т.е. $p_i=1/N$ для всех $i=\overline{1,N}$, то

$$H(Z)=-\sum_{i=1}^{N}\frac{1}{N}\log_2\frac{1}{N}=\log_2 N. \qquad (4.7)$$

Таким образом, мера Р. Хартли – частный случай меры К. Шеннона для равновероятных элементов. Можно также показать, что мера К. Шеннона является обобщением меры Хартли на случай неравновероятных элементов.

4.3 Свойства энтропии

1. Энтропия – величина вещественная и неотрицательная. Свойство легко проверяется по формуле (4.5) с учетом того, что $0\le p(z_i)\le 1$ для всех $i=\overline{1,N}$.

2. Энтропия величина ограниченная. При $0<p_i\le 1$ это свойство непосредственно следует из формулы (4.5). При $p=0$ имеем:

$$\lim_{p\to 0}\left(-p\log_2 p\right)=\lim_{p\to 0}\frac{\log_2\frac{1}{p}}{\frac{1}{p}}=\lim_{\alpha\to\infty}\frac{\log_2\alpha}{\alpha}=\lim_{\alpha\to\infty}\frac{\log_2 e}{\alpha\cdot 1}=0$$

(здесь произведена замена $1/p=\alpha$ и далее раскрыта неопределенность по правилу Лопиталя). Таким образом, при любых значениях $0\le p_i\le 1$, $i=\overline{1,N}$ $H(Z)<\infty$.

3. По ходу доказательства свойства 2 нетрудно заметить, что $H(Z)=0$, если вероятность одного из элементов множества равна 1.

4. Энтропия максимальна, когда все элементы множества равновероятны и

$$H_{\max}(Z)=\max_{\forall p_i}H(Z)=\log_2 N. \qquad (4.8)$$

Будем искать максимум (4.5) при условии $\sum p_i=1$. Функция Лагранжа для соответствующей задачи на безусловный экстремум

$$F(p,\lambda)=-\sum_{p=1}^{N}p_i\log_2 p_i+\lambda\left(\sum_{i=1}^{N}p_i-1\right)\to extr.$$

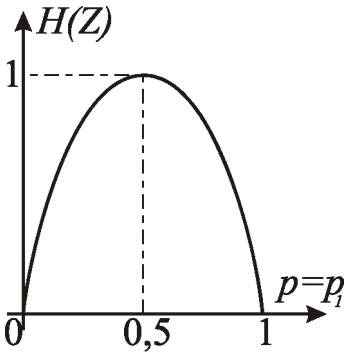

Рис. 4.1. Изменение энтропии в случае двух элементов

Необходимые условия экстремума:

$$\frac{\partial F(p,\lambda)}{\partial p_i}=-\log_2 p_i-\log_2 e+\lambda=0,$$

$$\frac{\partial F(p,\lambda)}{\partial \lambda}=\sum_{i=1}^{N}p_i-1=0,$$

откуда следует $p_i=2^{\lambda-\log_2 e}=Const=1/N$. Проверкой легко убедиться, что указанное значение доставляет максимум.

5. В частном случае множества с двумя элементами зависимость энтропии от вероятности одного из элементов имеет вид, показанный на рис. 4.1. В этом можно убедиться, применяя соотношения и выводы, полученные при рассмотрении свойств 2 и 3 к соотношению (4.5), которое в данном случае принимает вид

$$H(Z)=-p\log_2 p-(1-p)\log_2(1-p). \qquad (4.9)$$

6. Энтропия объединения статистически независимых множеств равна сумме энтропий исходных множеств. При установлении этого свойства используется свойство вероятностей независимых элементов:

$$p(z_i, v_j) = p(z_i) \cdot p(v_j).$$

Поскольку при этом $\log_2 p(z_i, v_j) = \log_2 p(z_i) + \log_2 p(v_j)$ имеем

$$\begin{aligned} H(ZV) &= -\sum_{i=1}^{N}\sum_{j=1}^{K} p(z_i, v_j)\log_2 p(z_i, v_j) = \\ &= -\sum_{i=1}^{N}\sum_{j=1}^{K} p(z_i)\,p(v_j)\log_2\left[p(z_i)\,p(v_j)\right] = \\ &= -\sum_{i=1}^{N} p(z_i)\log_2 p(z_i)\underbrace{\sum_{j=1}^{K} p(v_j)}_{1} - \sum_{j=1}^{K} p(v_j)\log_2 p(v_j)\underbrace{\sum_{i=1}^{N} p(z_i)}_{1} = \\ &= H(Z) + H(V). \end{aligned} \tag{4.10}$$

Аналогично могут быть получены формулы для объединения любого числа независимых источников.

В заключение подчеркнем, что энтропия характеризует только среднюю неопределенность выбора одного элемента из множества, полностью игнорируя их содержательную сторону.

4.4 Условная энтропия и её свойства

Часто имеют место связи между элементами разных множеств или между элементами одного множества. Пусть объединенный ансамбль $\{ZV\}$ задан матрицей вероятностей всех его возможных элементов $z_i v_j$, $i = \overline{1,N}$, $j = \overline{1,K}$:

$$\begin{bmatrix} p(z_1, v_1) & p(z_2, v_1) & \dots & p(z_N, v_1) \\ p(z_1, v_2) & p(z_2, v_2) & \dots & p(z_N, v_2) \\ \dots & \dots & \dots & \dots \\ p(z_1, v_K) & p(z_2, v_K) & \dots & p(z_N, v_K) \end{bmatrix}. \tag{4.11}$$

Суммируя вероятности по строкам и столбцам (4.11) в соответствии с (4.1), можно определить также ансамбли $\{Z, p(z)\}$ и $\{V, p(v)\}$:

$$\{Z, p(z)\} = \begin{bmatrix} z_1 & z_2 & \dots & z_N \\ p(z_1) & p(z_2) & \dots & p(z_N) \end{bmatrix},$$

$$\{V, p(v)\} = \begin{bmatrix} v_1 & v_2 & \dots & v_K \\ p(v_1) & p(v_2) & \dots & p(v_K) \end{bmatrix}.$$

Поскольку в случае зависимых элементов

$$p(z_i, v_j) = p(z_i)\, p(v_j / z_i) = p(v_i)\, p(z_i / v_j), \tag{4.12}$$

с использованием первого из указанных в (4.12) равенств можно записать

$$\begin{aligned} H(ZV) &= -\sum_{ij} p(z_i, v_j) \log_2 p(z_i, v_j) = \\ &= -\sum_i p(z_i) \log_2 p(z_i) \sum_j p(v_j / z_i) - \\ &\quad -\sum_i p(z_i) \sum_j p(v_j / z_i) \log_2 p(v_j / z_i). \end{aligned} \tag{4.13}$$

По условию нормировки $\sum_j p(v_j / z_i) = 1$ для любого $i = \overline{1, N}$, поэтому первое слагаемое в правой части является энтропией $H(Z)$ ансамбля $\{Z, p(z)\}$. Вторая сумма (по j) во втором слагаемом характеризует частную неопределенность, приходящуюся на одно состояние ансамбля V при условии, что реализовалось состояние z_i ансамбля Z. Ее называют *частной условной энтропией* и обозначают $H_{z_i}(V)$:

$$H_{z_i}(V) = -\sum_{j=1}^{K} p(v_j / z_i) \log_2 p(v_j / z_i). \tag{4.14}$$

Величина $H_Z(V)$, получаемая усреднением частной условной энтропии по всем элементам z_i:

$$H_Z(V) = \sum_{i=1}^{N} p(z_i) H_{z_i}(V), \tag{4.15}$$

называется *полной условной энтропией* или просто условной энтропией. Таким образом, (4.13) с учетом (4.14), (4.15) можно записать в виде

$$H(ZV) = H(Z) + H_Z(V). \tag{4.16}$$

Используя второе равенство в (4.12), по аналогии можно записать:

$$H(ZV) = H(V) + H_V(Z). \qquad (4.17)$$

Можно также показать, что в случае объединения любого числа множеств $\{ZVW...\}$ с зависимыми элементами имеет место равенство

$$H(ZVW...) = H(Z) + H_Z(V) + H_{ZV}(W) + ... \quad .$$

Подчеркнем, что условная энтропия всегда меньше или равна безусловной:

$$H_V(Z) \le H(Z), \;\; H_Z(V) \le H(V). \qquad (4.18)$$

Справедливость неравенств (4.18) интуитивно понятна: неопределенность выбора элемента из некоторого множества может только уменьшиться, если известен элемент другого множества, с элементами которого существует взаимосвязь. Из (4.16)–(4.18), в частности, следует

$$H(ZV) \le H(Z) + H(V). \qquad (4.19)$$

Полезно дать геометрическую интерпретацию соотношений (4.16)–(4.19). На рис. 4.2 наглядно показаны различия, которые имеют место при вычислении энтропии объединенного множества в случае независимых (а) и зависимых (б) элементов.

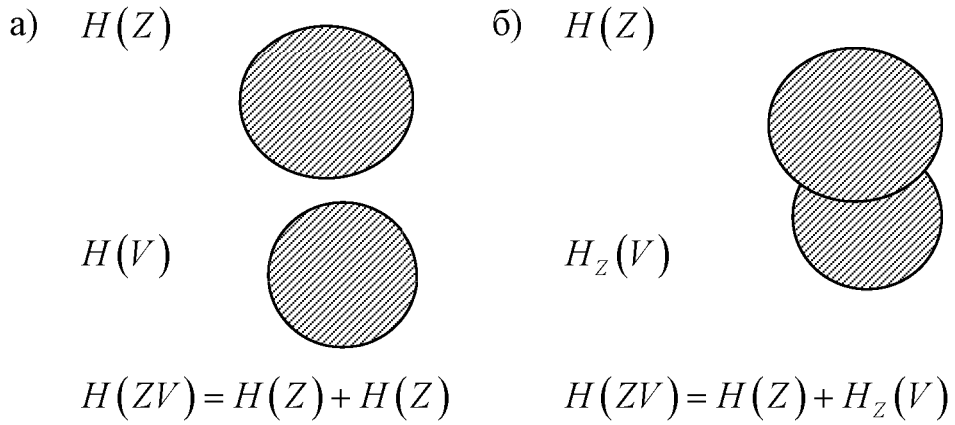

Рис. 4.2. Иллюстрация формирования энтропии объединенных ансамблей

Часто имеет место другой тип связи, а именно: статистическая зависимость между элементами последовательности. Если имеет место связь только между двумя соседними элементами последовательности, она характеризуется условной вероятностью $p(z_i / z_j)$. Последовательность элементов, обладающую указанным свойством, называют *односвязной цепью Маркова.* Связь каждого эле-

мента с двумя предшествующими характеризуется условной вероятностью $p\left(z_i / z_j z_k\right)$, а соответствующая последовательность называется *двусвязной цепью Маркова.*

Для односвязной цепи Маркова в предположении, что известен (принят) элемент z_j из алфавита объема N, частная условная энтропия

$$H\left(Z / z_j\right) = -\sum_{i=1}^{N} p\left(z_i / z_j\right) \log_2 p\left(z_i / z_j\right).$$

При этом полная (средняя) условная энтропия определяется как

$$H(Z) = -\sum_{j=1}^{N} p\left(z_j\right) \sum_{i=1}^{N} p\left(z_i / z_j\right) \log_2 p\left(z_i / z_j\right). \qquad (4.20)$$

Аналогично для двусвязной цепи Маркова

$$H\left(Z / z_j z_k\right) = -\sum_{i=1}^{N} p(z_i / z_j z_k) \log_2 p(z_i / z_j z_k),$$

$$H(Z) = -\sum_{j,k} p(z_j, z_k) \sum_{i} p(z_i / z_j z_k) \log_2 p(z_i / z_j z_k). \qquad (4.21)$$

Можно построить выражения для энтропии и при более протяженной связи между элементами последовательности.

Лекция 5

Меры неопределенности непрерывных случайных величин

5.1 Понятие дифференциальной энтропии

Перейдем к рассмотрению источников информации, выходные сигналы которых являются непрерывной случайной величиной. Множество возможных состояний такого источника составляет континуум, а вероятность любого конкретного значения равна 0, что делает невозможным применение, например, меры (4.5). Построим меры неопределенности таких источников, опираясь на введенные ранее меры для дискретных ансамблей.

Мы можем приближенно оценить неопределенность выбора какого-либо значения непрерывной случайной величины по формуле (4.5), если ограничим диапазон ее допустимых значений и разобьем этот диапазон, например, на равные интервалы, вероятность попадания в каждый из которых отлична от нуля и определяется как

$$\mathrm{P}\{z_i \le Z < z_i + \Delta z\} \cong p\left(z_i^*\right)\Delta z .$$

Здесь $p\left(z_i^*\right)$ – ордината плотности распределения $p(z)$ непрерывной случайной величины при значении z_i^*, принадлежащем интервалу $[z_i, z_i + \Delta z]$.

Заменяя в (4.5) $p(z_i)$ его приближенным значением $p\left(z_i^*\right)\cdot\Delta z$, имеем

$$\begin{aligned} H(Z) &= -\sum_{i=1}^{N} p\left(z_i^*\right)\Delta z \log_2\left(p\left(z_i^*\right)\Delta z\right) = \\ &= -\sum_{i=1}^{N} p\left(z_i^*\right)\log_2 p\left(z_i^*\right)\Delta z - \log_2\Delta z \sum_{i=1}^{N} p\left(z_i^*\right)\Delta z. \end{aligned} \tag{5.1}$$

Далее осуществим предельный переход при $\Delta z \to 0$. При этом сумма переходит в интеграл, $\Delta z \to dz$, а $\sum_{i=1}^{N} p(z_i^*)\Delta z \to 1$. С учетом того, что в общем случае диапазон изменения непрерывной случайной величины $(-\infty;+\infty)$, получаем:

$$H(Z) = -\int_{-\infty}^{+\infty} p(z)\log_2 p(z)dz - \lim_{\Delta z\to 0}\log_2\Delta z . \tag{5.2}$$

Из формулы (5.2) следует, что энтропия непрерывной случайной величины равна бесконечности независимо от вида плотности вероятности. Этот факт, вообще говоря, не является удивительным, так как вероятность конкретного значения непрерывного сигнала равна 0, а множество состояний бесконечно. Ясно, что использовать такую меру на практике не представляется возможным.

Для получения конечной характеристики информационных свойств используется только первое слагаемое, называемое *дифференциальной энтропией*:

$$h(Z) = -\int_{-\infty}^{+\infty} p(z)\log_2 p(z)dz. \qquad (5.3)$$

Термин дифференциальная энтропия связан с тем, что для ее определения в формуле (5.3) используется дифференциальный закон распределения $p(z)$. Возникает естественный вопрос: не является ли это соглашение искусственным и не имеющим смысла.

Оказывается, что дифференциальная энтропия имеет смысл средней неопределённости выбора случайной величины с произвольным законом распределения за вычетом неопределённости случайной величины, равномерно распределённой в единичном интервале.

Действительно, энтропия (5.2) равномерно распределённой на интервале δ случайной величины Z_r определяется как

$$H(Z_r) = -\int_{-\infty}^{\infty} \frac{1}{\delta}\log_2 \frac{1}{\delta} dz - \lim_{\Delta z \to 0} \log_2 \Delta z_r.$$

При $\delta = 1$

$$H(Z_r) = -\lim_{\Delta z \to 0} \log_2 \Delta z_r \qquad (5.4)$$

Сравнивая (5.2) и (5.4), нетрудно заметить, что при $\Delta z = \Delta z_r$

$$H(Z) - H(Z_r) = h(z). \qquad (5.5)$$

5.2 Понятие дифференциальной условной энтропии

Рассмотрим теперь ситуацию, когда (далее две) непрерывные случайные величины статистически связаны. Как и ранее, разобьем диапазоны допустимых значений случайных величин на равные интервалы так, что

$$\mathrm{P}\{z_i \le Z < z_i + \Delta z, \quad v_j \le V < v_j + \Delta v\} \cong p(z_i^*, v_j^*) \cdot \Delta z \Delta v, \tag{5.6}$$

где $p(z_i^*, v_j^*)$ – ордината двумерной плотности распределения в точке (z_i^*, v_j^*), принадлежащей прямоугольнику со сторонами Δz, Δv: $(z_i \le z_i^* < z_i + \Delta z$, $v_j \le v_j^* < v_j + \Delta v)$. Подставляя приближенные значения вероятностей (5.6) в формулу энтропии (4.5), получаем

$$H(Z,V) = -\sum_i \sum_j p(z_i^*, v_j^*) \log_2 p(z_i^*, v_j^*) \Delta z \Delta v -$$
$$-\log_2 \Delta z \sum_i \sum_j p(z_i^*, v_j^*) \Delta z \Delta v - \log_2 \Delta v \sum_i \sum_j p(z_i^*, v_j^*) \Delta z \Delta v.$$

С учетом того, что $p(z_i^*, v_j^*) = p(z_i^*) p(v_j^* / z_i^*)$, первое слагаемое в правой части последнего равенства можно представить в виде суммы

$$-\sum_i p(z_i^*) \log_2 p(z_i^*) \Delta z \sum_j p(v_j^* / z_i^*) \Delta v - \sum_i \sum_j p(z_i^*, v_j^*) \log_2 p(v_j^* / z_i^*) \Delta v \Delta z .$$

Далее осуществляя предельный переход при $\Delta z \to 0$, $\Delta v \to 0$, с учетом того, что по условию нормировки

$$\lim_{\substack{\Delta z \to 0 \\ \Delta v \to 0}} \sum_i \sum_j p(z_i^*, v_j^*) \Delta z \Delta v = 1,$$

$$\lim_{\Delta v \to 0} \sum_i \sum_j p(v_j^* / z_i^*) \Delta v = 1,$$

$$\lim_{\Delta z \to 0} \sum_i \sum_j p(z_i^*) \Delta z = 1,$$

получаем

$$H(Z,V) = -\int_{-\infty}^{\infty} p(z) \log_2 p(z) dz - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(z,v) \log_2 p(v/z) dz dv - \tag{5.7}$$
$$- \lim_{\Delta z \to 0} \log_2 \Delta z \quad - \lim_{\Delta v \to 0} \log_2 \Delta v.$$

Первое и третье слагаемое – суть энтропия $H(Z)$ непрерывного источника (5.2), выходным сигналом которого является случайная величина Z, а величина

$$H_Z(V) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(z,v)\log_2 p(v/z)dzdv - \lim_{\Delta v\to 0}\log_2 \Delta v \qquad (5.8)$$

является *условной энтропией* непрерывной случайной величины. Она, как и следовало ожидать, в силу второго слагаемого в правой части равна бесконечности. Поэтому, как и в случае одного независимого источника, принимают во внимание только первое слагаемое:

$$h_Z(V) = -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p(z,v)\log_2 \frac{p(z,v)}{p(z)}dzdv\,. \qquad (5.9)$$

Величину (5.9) называют *условной дифференциальной энтропией.*

Условная дифференциальная энтропия характеризует среднюю неопределенность выбора непрерывной случайной величины с произвольным законом распределения при условии, что известны результаты реализации другой, статистически связанной с ней непрерывной случайной величины, за вычетом средней неопределенности выбора случайной величины, имеющей равномерное распределение на единичном интервале.

Сопоставляя (5.2), (5.3), (5.7), (5.8), (5.9), дифференциальную энтропию двух непрерывных статистически связанных источников можно представить в виде

$$h(ZV) = h(Z) + h_Z(V) = h(V) + h_V(Z)\,. \qquad (5.10)$$

Второе равенство в (5.10) получается по той же схеме, что и первое, при $p(z_i^*, v_j^*) = p(v_j^*)p(z_i^*/v_j^*)$. Заметим также, что в соответствии с (5.7), (5.8) для непрерывных источников можно выписать равенства, аналогичные (4.16) и (4.18) для дискретных сообщений: $H(ZV) = H(Z) + H_Z(V) = H(V) + H_V(Z)$, однако они имеют лишь теоретическое значение, поскольку оперировать на практике с бесконечными неопределенностями не представляется возможным.

5.3 Свойства дифференциальной энтропии

Дифференциальная энтропия, в отличие от энтропии дискретного источника, является относительной мерой неопределенности, т.к. её значения зависят от масштаба непрерывной величины. Действительно, предположим, что непрерывная случайная величина Z изменилась в k раз. Поскольку всегда должно выполняться условие нормировки:

$$\int_{-\infty}^{+\infty} p(kz)d(kz) = k\int_{-\infty}^{+\infty} p(kz)dz = 1,$$

имеет место следующее соотношение для плотностей исходной и масштабированной величин

$$p(kz) = \frac{p(z)}{k}. \tag{5.11}$$

С учетом (5.11) в соответствии с (5.3) имеем

$$\begin{aligned} h(kZ) &= -\int_{-\infty}^{+\infty} p(kz)\cdot\log_2 p(kz)\cdot d(kz) = \\ &= -\int_{-\infty}^{+\infty} p(z)\left[\log_2 p(z) - \log_2 k\right]dz = \\ &= -\int_{-\infty}^{+\infty} p(z)\log_2 p(z)dz + \log_2 k\int_{-\infty}^{+\infty} p(z)dz = h(Z) + \log_2 k. \end{aligned} \tag{5.12}$$

Из (5.12) следует, что из-за выбора различных k дифференциальная энтропия может принимать положительные, отрицательные и нулевые значения.

Дифференциальная энтропия не зависит от параметра сдвига $\Theta = Const$, т.е. $h(Z+\Theta) = h(Z)$. Действительно, используя замену $V = Z + \Theta$, при которой пределы интегрирования не изменяются, а $dz = dv$ имеем:

$$\begin{aligned} h(Z+\Theta) &= -\int_{-\infty}^{+\infty} p(z+\Theta)\log_2 p(z+\Theta)dz = \\ &= -\int_{-\infty}^{+\infty} p(v)\log_2 p(v)dv = h(V). \end{aligned} \tag{5.13}$$

5.4 Распределения, обладающие максимальной дифференциальной энтропией

Сформулируем следующую задачу. Определить плотность $p(z)$, обеспечивающую максимальное значение функционала

$$h(Z) = -\int_{\alpha}^{\beta} p(z)\log_2 p(z)dz, \tag{5.14}$$

при ограничении

$$\int_{\alpha}^{\beta} p(z)dz = 1. \tag{5.15}$$

Функция Лагранжа в указанной (изопериметрической) задаче имеет вид

$$F(p,\mu) = p(z)\log_2 p(z) + \mu \cdot p(z), \tag{5.16}$$

где μ, в данном случае постоянный, неопределенный множитель Лагранжа. Необходимые условия экстремума (5.16) даются соотношением

$$\frac{\partial F(p,\mu)}{\partial p} = \log_2 p(z) + \log_2 e + \mu = 0. \tag{5.17}$$

Искомая плотность $p(z) = 1/(\beta - \alpha)$, $\alpha \le z \le \beta$ получается в результате совместного решения (5.15), (5.17). Это означает, что если единственным ограничением для случайной величины является область возможных значений: $Z \in [\alpha, \beta]$, то максимальной дифференциальной энтропией обладает равномерное распределение вероятностей в этой области.

Снимем теперь ограничение на область возможных значений, но добавим ограничение на величину дисперсии:

$$h(Z) = -\int_{-\infty}^{\infty} p(z)\log_2 p(z)dz \to max, \tag{5.18}$$

при

$$\int_{-\infty}^{\infty} p(z)dz = 1, \tag{5.19}$$

$$\int_{-\infty}^{\infty} z^2 p(z)dz = \sigma^2. \tag{5.20}$$

Функция Лагранжа в данном случае принимает вид

$$F(p,\mu_1,\mu_2) = p(z)\log_2 p(z) + \mu_1 \cdot p(z) + \mu_2 z^2 p(z),$$

а соответствующее уравнение Эйлера

$$\frac{\partial F(p,\mu)}{\partial p} = \log_2 p(z) + \log_2 e + \mu_1 + \mu_2 z^2 = 0. \qquad (5.21)$$

Непосредственной подстановкой можно убедиться, что гауссовская плотность

$$p(z) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{z^2}{2\sigma^2}\right\}$$

удовлетворяет необходимому условию (5.21) экстремума (в данном случае максимума) функционала (5.18) и заданным изопериметрическим ограничениям (5.19), (5.20). Заметим, что при выводе для простоты математическое ожидание мы приняли равным нулю, поскольку дифференциальная энтропия все равно не зависит от параметра сдвига.

Лекция 6

Количество информации как мера снятой неопределенности

6.1 Количество информации при передаче отдельного элемента дискретного сообщения

Предположим, что задан некоторый дискретный источник информации, характеризующийся дискретным вероятностным ансамблем:

$$Z = \begin{bmatrix} z_1 & z_2 & \cdots & z_N \\ p(z_1) & p(z_2) & \cdots & p(z_N) \end{bmatrix},$$

где z_i, $i = \overline{1,N}$ – его возможные состояния. Каждому состоянию источника можно поставить в соответствие отдельный первичный сигнал. Некоторую заданную совокупность первичных сигналов, поступающих с выхода источника информации на вход канала связи, принято называть сообщением, а z_i – элементом сообщения.

Если состояния источника реализуются независимо друг от друга, то частная априорная неопределённость появления на входе канала элемента сообщения z_i в соответствии с (4.6) определяется как

$$H(z_i) = -\log_2 p(z_i). \tag{6.1}$$

Предположим, что статистическая связь между помехой и элементами сообщения отсутствует и известны условные вероятности того, что вместо z_i принимается v_j:

$$p(z_i / v_j), \quad i = \overline{1,N}, \quad j = \overline{1,K}.$$

Таким образом, если на выходе канала получен элемент v_j, то становится известной апостериорная вероятность $p(z_i / v_j)$. Следовательно, можно определить апостериорную частную неопределённость:

$$H_{v_j}(z_i) = -\log_2 p(z_i / v_j). \tag{6.2}$$

Частное количество информации, полученное в результате того, что стал известен элемент v_j, определим как разность априорной и апостериорной неопределенностей:

$$I\left(z_i, v_j\right) = H\left(z_i\right) - H_{v_j}\left(z_i\right) =$$

$$= -\log_2 p\left(z_i\right) + \log_2 p\left(z_i / v_j\right) = \log_2 \frac{p\left(z_i / v_j\right)}{p\left(z_i\right)}. \qquad (6.3)$$

Таким образом, частное количество информации равно величине неопределённости, которая снята в результате получения элемента сообщения v_j.

6.2 Свойства частного количества информации

1. Частное количество информации уменьшается с ростом априорной вероятности $p\left(z_i\right)$, увеличивается с ростом апостериорной вероятности $p\left(z_i / v_j\right)$ и в зависимости от соотношения между ними может быть положительным, отрицательным и нулевым (свойство непосредственно следует из (6.3)).

2. Если $p\left(z_i / v_j\right) = p\left(z_i\right)$, то в соответствии с (6.3) $I\left(z_i, v_j\right) = 0$.

3. При отсутствии помехи частное количество информации равно частной априорной неопределенности элемента z_i: $I(z_i, v_j) = H(z_i) = -\log_2 p(z_i)$, поскольку при этом $H_{v_j}\left(z_i\right) = 0$.

4. Частное количество информации о z_i, содержащееся в v_j, равно частному количеству информации о v_j, содержащемуся в z_i. Действительно:

$$I(z_i, v_j) = \log_2 \frac{p\left(z_i / v_j\right)}{p(z_i)} = \log_2 \frac{p(v_j) p\left(z_i / v_j\right)}{p(v_j) p(z_i)} =$$

$$\log_2 \frac{p(z_i) p\left(v_j / z_i\right)}{p(z_i) p(v_j)} = \log_2 \frac{p\left(v_j / z_i\right)}{p(v_j)} = I(v_j, z_i).$$

6.3 Среднее количество информации в любом элементе дискретного сообщения

Априорная неопределённость в среднем на один элемент сообщения характеризуется энтропией (4.5):

$$H(Z) = -\sum_{i=1}^{N} p(z_i) \cdot \log_2 p(z_i), \tag{6.4}$$

а апостериорная неопределенность – условной энтропией (4.15):

$$H_V(Z) = -\sum_{j=1}^{K} p(v_j) \sum_{i=1}^{N} p\left(z_i / v_j\right) \log_2 p\left(z_i / v_j\right). \tag{6.5}$$

В соответствии с (6.4), (6.5) по аналогии с частным количеством информации количество информации в среднем на один элемент сообщения определим как

$$I(Z,V) = H(Z) - H_V(Z) =$$

$$= -\sum_i p(z_i) \log_2 p(z_i) + \sum_j p(v_j) \sum_i p\left(z_i / v_j\right) \log_2 p\left(z_i / v_j\right).$$

В последнем равенстве ничего не изменится, если первое слагаемое в правой части умножить на $\sum_{j=1}^{K} p\left(v_j / z_i\right) = 1$. Тогда, с учетом того, что

$$\sum_i p(z_i) \sum_j p\left(v_j / z_i\right) = \sum_j p(v_j) \sum_i p\left(z_i / v_j\right) = \sum_{ij} p\left(z_i, v_j\right),$$

и используя свойства логарифма, формулу для количества информации в среднем на один элемент сообщения можно записать в виде

$$I(Z,V) = \sum_{ij} p(z_i, v_j) \log_2 \frac{p\left(z_i / v_j\right)}{p(z_i)} = \sum_{ij} p(z_i, v_j) \log_2 \frac{p\left(z_i, v_j\right)}{p(z_i) p(v_j)}. \tag{6.6}$$

Далее, если частный характер количества информации не будет оговариваться специально, то всегда будет подразумеваться количество информации в среднем на один элемент сообщения (6.6).

6.4 Свойства среднего количества информации в элементе сообщения

1. Неотрицательность. $I(Z,V) \geq 0$, так как всегда $H(Z) \geq H_V(Z)$.

2. $I(Z,V) = 0$ при отсутствии статистической связи между Z и V, так как при этом $H(Z) = H_V(Z)$.

3. $I(Z,V) = I(V,Z)$, то есть количество информации в V относительно Z равно количеству информации в Z относительно V. Действительно,

$$I(Z,V) - I(V,Z) = H(Z) - H_V(Z) - (H(V) - H_Z(V)) =$$
$$= H(Z) + H_Z(V) - (H(V) + H_V(Z)) = H(Z,V) - H(V,Z) = 0$$

4. При отсутствии помех $I(Z,V) = H(Z)$, поскольку при этом $H_V(Z) = 0$. Это максимальное количество информации, которое может быть получено от источника.

6.5 Количество информации при передаче сообщений от непрерывного источника

Соотношение для количества информации непрерывного источника получим из формулы (6.6) для дискретного случая. Обозначив переданный и принятый непрерывные сигналы соответственно Z и V разобьем область допустимых значений этих сигналов на равные интервалы и запишем приближенные вероятности (см. рис. 6.1):

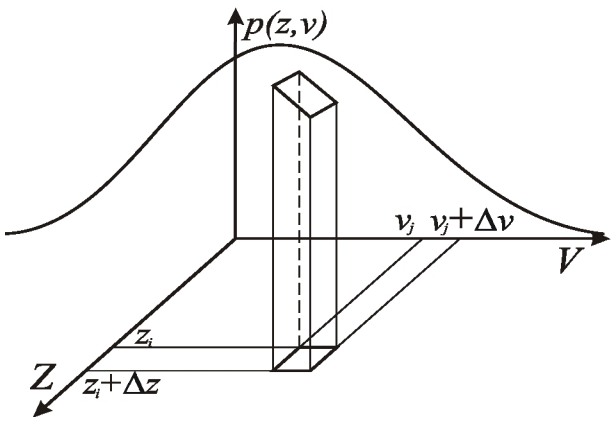

Рис. 6.1. Дискретизация области Z,V

$$\mathrm{P}\{z_i \leq Z < z_i + \Delta z, v_j \leq V < v_j + \Delta v\} \cong p(z_i^*, v_j^*)\Delta z \Delta v,$$

где $p(z_i^*, v_j^*)$ – ордината двумерной плотности распределения $p(z,v)$ в некоторой точке, принадлежащей прямоугольнику с номером i, j.

Для соответствующих заданной двумерной плотности $p(z,v)$ одномерных плотностей $p(z_i)$, $p(v_j)$, по аналогии с тем, как мы поступали при получении соотношения для дифференциальной энтропии, можно записать

$$\mathrm{P}\{z_i \le Z < z_i + \Delta z\} \cong p(z_i^*)\Delta z,$$

$$\mathrm{P}\{v_j \le V < v_j + \Delta v\} \cong p(v_j^*)\Delta v,$$

где $p(z_i^*)$, $p(v_j^*)$ – ординаты одномерных плотностей для значений z_i^* и v_j^*, взятых в интервалах $[z_i, z_i + \Delta z]$ и $[v_i, v_i + \Delta v]$ соответственно.

Заменяя в (6.6) $p(z_i, v_j)$, $p(z_i)$, $p(v_j)$ их приближенными значениями $p(z_i^*, v_j^*)\Delta z\Delta v$, $p(z_i^*)\Delta z$, $p(v_j^*)\Delta v$ соответственно, можно записать

$$I(Z,V) = \sum_i \sum_j p(z_i^*, v_j^*)\Delta z\Delta v \cdot \log_2 \frac{p(z_i^*, v_j^*)}{p(z_i^*)p(v_j^*)}. \tag{6.7}$$

Осуществляя в (6.7) предельный переход при $\Delta z \to 0$, $\Delta v \to 0$, получаем:

$$\begin{aligned} I(Z,V) &= \lim_{\substack{\Delta z \to 0 \\ \Delta v \to 0}} \sum_i \sum_j p^*(z_i, v_j) \log_2 \frac{p^*(z_i, v_j)}{p^*(z_i)p^*(v_j)} \Delta z\Delta v = \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(z,v) \log_2 \frac{p(z,v)}{p(z)p(v)} dzdv. \end{aligned} \tag{6.8}$$

Формула (6.8) может быть получена также с использования понятия дифференциальной энтропии. Действительно, по аналогии с дискретным случаем, определим количество информации как разность априорной и апостериорной (в данном случае дифференциальной) энтропии:

$$\begin{aligned} I(Z,V) &= h(Z) - h_v(Z) = \\ &= -\int_{-\infty}^{\infty} p(z) \log_2 p(z) dz + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(z,v) \log_2 p(z/v) dzdv. \end{aligned} \tag{6.9}$$

В (6.9) ничего не изменится, если первое слагаемое в правой части умножить на $\int_{-\infty}^{\infty} p(v/z)dv = 1$. Тогда, с учетом того, что $p(z,v) = p(v)p(z/v) = p(z)p(v/z)$, соотношение (6.9) можно переписать в следующем виде:

$$I(Z,V) = h(Z) - h_v(Z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(z,v)\log_2\frac{p(z,v)}{p(z)p(v)}dzdv\,. \qquad (6.10)$$

Поскольку $I(Z,V)$ в (6.10) определяется как разность $h(Z) - h_v(Z)$, количество информации при передаче от непрерывного источника, в отличие от дифференциальной энтропии, уже не зависит от масштаба случайной величины. Заметим, что соотношение между понятиями энтропии и количества информации для непрерывного источника информации подобно соотношению между потенциалом, определяемым как работа по перенесению заряда из бесконечности в данную точку поля, и напряжением, определяемым как разность потенциалов, которое рассматривается в физике.

6.6 Эпсилон-энтропия случайной величины

В этом разделе мы вернемся к рассмотрению понятия энтропии *непрерывной* случайной величины, воспользовавшись для этого теперь уже известным нам понятием количества информации.

В разделе 5.1 мы показали, что энтропия *непрерывной* случайной величины бесконечна, вследствие того, что реализации могут отличаться на сколь угодно малые величины. В действительности на практике, с одной стороны, нет возможности фиксировать сколь угодно малые отличия реализаций вследствие погрешности измерительной аппаратуры, с другой стороны, это обычно и не требуется. Поэтому разумной представляется идея: судить о непрерывной случайной величине Z по значениям другой статистически связанной с ней случайной величины V, если мера их различия не превышает заданной верности воспроизведения.

Для количественной оценки степени сходства вводят функцию $\rho(z,v)$, имеющую смысл «расстояния» между реализациями, а в качестве меры сходства – ее среднее значение по всему множеству значений z и v:

$$\mu(Z,V) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(z,v)\rho(z,v)dzdv\,. \qquad (6.11)$$

Здесь $p(z,v)$ – плотность совместного распределения вероятностей случайных величин Z и V.

Наиболее популярным является среднеквадратический критерий. При этом с учетом равенства $p(z,v) = p(z/v)p(v)$ критерий сходства может быть записан в виде

$$\mu(Z,V) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(z-v)^2 p(z/v)p(v)dzdv \le \varepsilon^2, \qquad (6.12)$$

где $p(z/v)$ – условная плотность распределения, характеризующая вероятность воспроизведения конкретного сигнала z сигналом v, а ε – заданное значение верности воспроизведения.

В соответствии с (6.10) количество информации о случайной величине Z, содержащейся в воспроизводящей величине V, равно

$$I(Z,V) = h(Z) - h_v(Z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(z/v)p(v)\log_2\frac{p(z/v)}{p(z)}dzdv. \qquad (6.13)$$

Заданную верность воспроизведения случайной величины Z желательно обеспечить при минимальном количестве получаемой информации. Поэтому условную плотность $p(z/v)$ вероятности того, что в тех случаях, когда был зафиксирован сигнал z, имел место сигнал v, следует подобрать так, чтобы в (6.13) имел место минимум информации $I(Z,V)$ по всем $p(z/v)$.

Величину $H_\varepsilon(Z)$, определяемую как

$$H_\varepsilon(Z) = \min_{p(z/v)} I(Z,V), \qquad (6.14)$$

при условии

$$\mu(Z,V) \le \varepsilon^2, \qquad (6.15)$$

называют эпсилон-энтропией (ε-энтропией) непрерывной случайной величины Z. В соответствии с (6.10) ее также можно определить как

$$H_\varepsilon(Z) = \min_{p(z/v)}\{h(Z) - h_V(Z)\} = h(Z) - \max_{p(z/v)} h_V(Z).$$

6.7 Избыточность сообщений

Сообщения, энтропия которых максимальна, являются оптимальными с точки зрения наибольшего количества передаваемой информации. Мерой отличия энтропии реального сообщения от оптимального является коэффициент сжатия:

$$\mu = \frac{H(Z)}{H_{opt}(Z)}. \tag{6.16}$$

Если оптимальное и неоптимальное сообщения характеризуются одинаковой общей энтропией, то имеет место равенство

$$nH(Z) = n'H_{opt}(Z), \tag{6.17}$$

где n – число элементов неоптимального сообщения, n' – число элементов оптимального сообщения.

С учетом (6.17) коэффициент сжатия (6.16) можно представить в виде

$$\mu = \frac{H(Z)}{H_{opt}(Z)} = \frac{n'}{n}.$$

Для характеристики близости энтропии реальных сообщений к оптимальному значению вводится также коэффициент избыточности:

$$K_z = 1 - \mu = \frac{n - n'}{n} = \frac{H_{opt}(Z) - H(Z)}{H_{opt}(Z)}.$$

Увеличение избыточности приводит к увеличению времени передачи сообщений. Однако некоторая избыточность может быть полезной с точки зрения повышения надежности системы.

Лекция 7

Оценка информационных характеристик источников сообщений

7.1 Понятие эргодического источника сообщений

Для построения модели источника дискретных сообщений достаточно задать объём алфавита и вероятности появления на выходе источника отдельных знаков. Наиболее широко используется модель Шеннона – эргодический источник сообщения. Эта модель предполагает, что источник представляется эргодической случайной последовательностью.

Свойства эргодической модели:

1) вероятности знаков не зависят от их места в последовательности;
2) статистические характеристики, полученные на одном длинном сообщении, справедливы для всех сообщений, создаваемых этим источником.

Если вероятности знаков не зависят от времени, то источник называется *стационарным.* Если вероятности не зависят и от предыдущих состояний, то источник называется *стационарным без памяти.* Стационарный источник без памяти, в котором каждый знак выбирается независимо от других, всегда эргодический.

Если имеет место корреляция между знаками, то в качестве модели используют цепь Маркова. Неопределенность этих источников описывается формулами (4.20), (4.21) (лекция 4). Порядок цепи зависит от того, сколько знаков связано корреляционной зависимостью.

Предположим, что вероятности знаков, формируемых источником с тремя возможными состояниями, следующие: $p(z_1)=0{,}1$, $p(z_2)=0{,}3$, $p(z_3)=0{,}6$. Ясно, что в этом случае знак z_2 в среднем должен встречаться в три раза чаще, чем z_1, но в два раза реже, чем z_3. Однако в конкретной последовательности, длина которой ограничена, знаки могут отсутствовать или появляться реже или чаще, чем это определено указанными вероятностями. Вероятности формиро-

вания различных последовательностей, связанные со свойствами эргодических последовательностей знаков, даются следующей теоремой.

7.2 Теорема о свойствах эргодических последовательностей знаков

Как бы ни были малы числа $\delta > 0$ и $\mu > 0$, при достаточно большом N все эргодические последовательности могут быть разбиты на две группы:

1. *Нетипичные последовательности.* Различных вариантов таких последовательностей большое число, однако любая из них имеет настолько ничтожную вероятность, что даже суммарная вероятность всех таких последовательностей очень мала и при достаточно большом N меньше сколь угодно малого числа δ.

2. *Типичные последовательности*, вероятности которых p при больших N одинаковы и удовлетворяют неравенству

$$\left| \frac{1}{N} \log_2 \frac{1}{p} - H(Z) \right| < \mu . \tag{7.1}$$

Соотношение (7.1) называют свойством *асимптотической равномерности.*

Доказательство. Для эргодического источника без памяти в длинной последовательности из N элементов алфавита объемом m $(z_1, z_2, ..., z_m)$ с вероятностями появления знаков $p_1, p_2, ..., p_m$ будет содержаться Np_1 элементов z_1, Np_2 элементов z_2 и т.д. Тогда вероятность p появления конкретной последовательности с учетом свойства независимости знаков

$$p = p_1^{Np_1} p_2^{Np_2} \dots p_m^{Np_m} = \prod_{i=1}^{m} p_i^{Np_i} . \tag{7.2}$$

Логарифмируя обе части равенства (7.2), получаем

$$\log_2 p = N \sum_{i=1}^{m} p_i \log_2 p_i . \tag{7.3}$$

Из (7.3) при $N \to \infty$ следует

$$\frac{1}{N} \log_2 \frac{1}{p} = H(Z), \tag{7.4}$$

что доказывает вторую часть теоремы.

Заметим, что это утверждение можно объяснить с несколько иных позиций. Поскольку по предположению источник выдает только эргодические последовательности, при $N \to \infty$ вероятности появления знаков в них будут соответствовать типичным для этих последовательностей значениям, следовательно, вероятности p появления этих последовательностей будут одинаковы. Общее число этих (типичных) последовательностей будет равным соответственно, $1/p$. Частная неопределенность каждой такой последовательности в соответствии с (4.4), (4.6): $-\log_2 p = \log_2(1/p)$, а неопределенность в среднем на один знак этой последовательности будет равна $\log_2(1/p)/N$, но эта величина по определению и является энтропией.

Покажем теперь, что при достаточно большом N типичные последовательности составляют незначительную долю от общего числа *возможных* вариантов различных последовательностей.

Общее число возможных вариантов последовательностей n_1, которое может быть сформировано из знаков алфавита объема m (с использованием основного логарифмического тождества), можно представить в виде

$$n_1 = m^N = 2^{\log_2 m^N} = 2^{N \log_2 m}.$$

С другой стороны, в соответствии с (7.4) число типичных последовательностей определяется как

$$n_T = \frac{1}{p} = 2^{NH(Z)}.$$

Запишем их отношение:

$$\frac{n_1}{n_T} = \frac{2^{N \log_2 m}}{2^{NH(Z)}} = 2^{N[\log_2 m - H(Z)]}.$$

В разделе 4.3 мы установили, что максимум энтропии $H(Z) = \log_2 m$ имеет место лишь в случае, когда знаки равновероятны. Это означает, что, если исключить случай равновероятного выбора элементов сообщений, в показателе степени двойки $H(Z) < \log_2 m$ и, следовательно, при $N \to \infty$ $n_1 >> n_T$.

7.3 Производительность источника дискретных сообщений

Производительность источника сообщений – это количество информации, вырабатываемое источником в единицу времени. Обычно помехи в источнике малы и их учитывают эквивалентным изменением модели канала связи. При этом производительность источника $\dot{I}_u(Z)$ численно равна величине энтропии в единицу времени и определяется соотношением

$$\dot{I}_u(Z) = \frac{H(Z)}{\tau_u}, \tag{7.5}$$

где τ_u – средняя длительность формирования одного знака.

Длительность выдачи каждого отдельного элемента сообщения в общем случае зависит не только от типа формируемого знака, но и от состояния источника. Поэтому средняя длительность τ_u выдачи источником одного знака в общем случае определяется как

$$\tau_u = \sum_q p(q) \sum_{i=1}^{N} p(z_i/q)\tau_{z_i q}, \tag{7.6}$$

где $\tau_{z_i q}$ – длительность выдачи знака z_i в состоянии q, $p(z_i/q)$ – вероятность появления знака z_i в состоянии q, а $p(q)$ – вероятность состояния q.

Из формулы (7.5) следует, что повысить производительность источника можно либо путем увеличения его энтропии, либо за счет уменьшения средней длительности формирования знаков. В соответствии с (7.6) уменьшение средней длительности τ_u наиболее эффективно за счет уменьшения длительности формирования тех знаков, которые имеют относительно высокие вероятности появления. Если длительности формирования знаков не зависят от состояний источника и одинаковы, повышение производительности возможно только за счет увеличения его энтропии.

7.4 Эпсилон-производительность источника непрерывных сообщений

Понятие эпсилон-производительности источника вводится подобно тому, как в разделе 6.6 было введено понятие эпсилон-энтропии непрерывной случайной величины.

Эпсилон-производительность (ε-*производительность*) источника непрерывных сообщений $\dot{H}_\varepsilon(Z)$ определяют как минимальное количество информации, которое необходимо создать источнику в единицу времени, чтобы любую реализацию $z_l(t)$ можно было воспроизвести с заданной верностью ε.

Предположим, что на достаточно длинном интервале T непрерывный сигнал $z_T(t)$ воспроизводится реализацией $v_T(t)$. Если указанные сигналы обладают ограниченным спектром F, то, в соответствии с теоремой Котельникова, каждую из этих реализаций можно представить составленными из отсчетов N-мерными ($N = T/\Delta t = 2FT$) векторами $(z_1, z_2, ..., z_N)$ и $(v_1, v_2, ..., v_N)$ соответственно. Соответствующие ансамбли сообщений можно представить N-мерными случайными векторами $\mathbf{Z}$, $\mathbf{V}$, компонентами которых являются случайные величины $Z_1, Z_2, ..., Z_N$, $V_1, V_2, ..., V_N$. Эти векторы могут быть статистически описаны с использованием N-мерных плотностей распределения – $p(\mathbf{Z})$, $p(\mathbf{V})$, $p(\mathbf{Z},\mathbf{V})$, $p(\mathbf{Z}/\mathbf{V})$, $p(\mathbf{V}/\mathbf{Z})$.

С использованием указанных N-мерных плотностей распределения, запишем соотношение (6.10) для количества информации, содержащегося в воспроизводящем векторе относительно исходного (здесь интегралы N-мерные):

$$I_N(\mathbf{Z},\mathbf{V}) = \int_{\mathbf{Z}}\int_{\mathbf{V}} p(\mathbf{Z},\mathbf{V}) \log_2 \frac{p(\mathbf{Z},\mathbf{V})}{p(\mathbf{Z})\,p(\mathbf{V})} d\mathbf{Z}\,d\mathbf{V}. \qquad (7.7)$$

Количество информации, приходящееся в среднем на один отсчет, определится как

$$I(\mathbf{Z},\mathbf{V}) = I_N(\mathbf{Z},\mathbf{V})/N. \qquad (7.8)$$

С использованием N-мерных плотностей распределения $p(\mathbf{Z/V})$ и $p(\mathbf{V})$, по аналогии с (6.11) можно также записать соотношение для количественной оценки степени сходства случайных векторов $\mathbf{Z}$, $\mathbf{V}$:

$$\mu(\mathbf{Z},\mathbf{V}) = \int_{\mathrm{Z}}\int_{\mathrm{V}} p(\mathbf{Z/V})\,p(\mathbf{V})\,\rho(\mathbf{Z},\mathbf{V})\,d\mathbf{Z}d\mathbf{V}, \tag{7.9}$$

где $\rho(\mathbf{Z},\mathbf{V})$ – функция, характеризующая близость случайных векторов $\mathbf{Z}$ и $\mathbf{V}$.

В соответствии с определением ε-производительности источника непрерывных сообщений можно записать

$$\dot{H}_{\varepsilon}(\mathbf{Z}) = \frac{1}{\tau_u}\min_{p(\mathbf{Z/V})} I(\mathbf{Z},\mathbf{V}), \tag{7.10}$$

при условии $\mu(\mathbf{Z},\mathbf{V}) \le \varepsilon^2$,

где $I(\mathbf{Z},\mathbf{V})$, $\mu(\mathbf{Z},\mathbf{V})$ определяются соотношениями (7.7)–(7.9) соответственно, а $\tau_u = 1/2F$ – время формирования одного отсчета источником.

Геометрически требование обеспечения заданной верности воспроизведения непрерывного сигнала можно представить как требование того, чтобы конец соответствующего сообщению $z_T(t)$ N-мерного вектора $(z_1, z_2, ..., z_N)$ попал в ε-область N-мерного вектора $(v_1, v_2, ..., v_N)$, соответствующего воспроизводящему непрерывному сигналу $v_T(t)$. Следует заметить, что заданная верность воспроизведения будет достигаться лишь при большой длительности сообщений, притом что $N = T/\Delta t = 2FT$, т.е. когда погрешностью от замены непрерывных сообщений совокупностью отсчетов можно пренебречь.

Лекция 8

Информационные характеристики каналов связи

8.1 Модели дискретных каналов

Канал связи – совокупность устройств, предназначенных для передачи сообщения от одного места к другому или от одного момента времени к другому. Канал, предназначенный для передачи дискретных сообщений, называют дискретным. Сигнал в таком канале при передаче от входа к выходу обычно подвергается преобразованиям в следующей последовательности устройств: источник сообщения – кодер источника – модулятор – передатчик – линия связи – приемник – демодулятор – декодер – приемник сообщения.

По линии связи, как правило, передается непрерывный сигнал. Считается, что именно в линии связи возникают наибольшие помехи. Поэтому при теоретическом исследовании модели канала с помехами полагают, что помехи в источнике отсутствуют, т.к. они малы по сравнению с помехами в канале. Если помехи в канале связи также невелики, то для теоретического анализа в первом приближении можно использовать идеализированную модель канала без помех.

Дискретный канал считается заданным, если известны множества символов (*алфавиты*) на входе и выходе, а также вероятностные свойства формирования (передачи) этих символов.

Для передачи по каналу сообщение из знаков алфавита источника $z_1, z_2, ...z_l$ преобразуется в дискретные последовательности символов из другого алфавита $v_1, v_2, ...v_m$, как правило, меньшего объёма.

В каждом состоянии канал характеризуется некоторой переходной вероятностью $p(v_j / z_i)$ того, что переданный символ z_i будет восприниматься на выходе как символ v_j. Если указанные вероятности не зависят от времени, то канал называют *стационарным*, если зависят от времени, то – *нестационарным.* Если эти вероятности зависят от предшествующего состояния, то имеет место канал *с памятью*, если не зависят, то это канал *без памяти*.

Если число символов на входе и на выходе канала одинаково и равно k, такой канал называют k-*ичным*. Стационарный двоичный канал без памяти характеризуется четырьмя переходными вероятностями (рис. 8.1). Если $p(0/0) = p(1/1)$ и $p(1/0) = p(0/1)$, то канал называется *симметричным*.

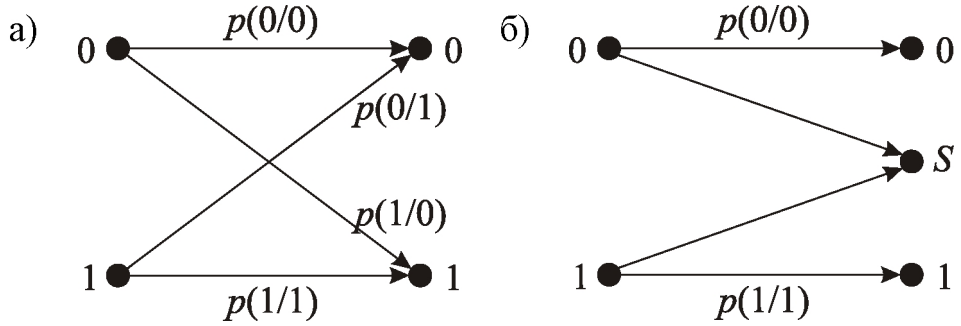

Рис. 8.1. Схемы каналов: двоичный (а); двоичный со стиранием (б)

Иногда также рассматривают модель канала *со стиранием*. На рис. 8.1, б приведена схема двоичного канала со стиранием. В данном случае на выходе канала фиксируются состояния S, которые с равной вероятностью могут быть отнесены как к единице, так и к нулю. При декодировании этот символ S расшифровывают с учетом дополнительной информации.

Если в канале имеется возможность формировать запрос на повторную передачу в случае обнаружения ошибки, такой канал называют каналом *с обратной связью*.

8.2 Скорость передачи информации по дискретному каналу

Различают техническую и информационную скорость передачи по дискретному каналу. Под *технической скоростью* понимают число элементов сообщения (символов), передаваемых в единицу времени:

$$V_\tau = \frac{1}{\tau_{ср}}, \qquad (8.1)$$

где $\tau_{ср}$ – средняя длительность передачи одного символа. Единицей технической скорости передачи является *бод* – один символ за одну секунду.

Под *информационной скоростью* понимают среднее количество информации, передаваемое по каналу в единицу времени. Она определяется как

$$\dot{I}(V,Z)=\frac{I(V,Z)}{\tau_{ср}}=V_{\tau}I(V,Z), \qquad (8.2)$$

где $I(V,Z)$ – среднее количество информации, переносимое одним символом.

8.3 Пропускная способность дискретного канала без помех

Пропускная способность дискретного канала без помех – C_{∂} определяется, как максимальная скорость передачи информации по данному каналу, которая в принципе может быть достигнута:

$$C_{\partial}=\max \dot{I}(V,Z)=\max V_{\tau}I(V,Z). \qquad (8.3)$$

В соответствии с (8.3) при фиксированной технической скорости передачи ($V_{\tau}=Const$) пропускная способность канала определяется максимумом среднего количества информации $I(V,Z)$, приходящейся на один символ принятого сигнала.

При отсутствии помех имеет место взаимно однозначное соответствие между символами на входе и выходе канала, а

$$I(V,Z)=H(Z).$$

С другой стороны, как было показано ранее, при фиксированном объеме алфавита m максимум $H(Z)$ имеет место при равновероятности символов и определяется как

$$\max H(Z)=\log_2 m.$$

Таким образом, для увеличения скорости передачи информации по дискретному каналу без помех необходимо осуществлять такое преобразование сообщений, при котором элементы сообщений оказываются независимыми и равновероятными. Из последнего равенства видно, что пропускная способность канала может быть повышена также путем увеличения объема алфавита m, однако это может быть связано с серьезными изменениями используемой элементной базы технических устройств.

8.4 Пропускная способность дискретного канала с помехами

В разделе 6.3 было показано, что количество информации в среднем на один элемент сообщения, поступающей от источника и передаваемой по каналу связи, определяется соотношением (6.6):

$$I(Z,V)=\sum_{ij}p(z_i,v_j)\log_2\frac{p\left(z_i,v_j\right)}{p(z_i)p(v_j)}. \tag{8.4}$$

Соответственно, скорость передачи информации по каналу с помехами в силу (8.2) дается равенством

$$\dot{I}(Z,V)=V_\tau\sum_{ij}p(z_i,v_j)\log_2\frac{p\left(z_i,v_j\right)}{p(z_i)p(v_j)}, \tag{8.5}$$

а пропускная способность дискретного канала с помехами определяется как предельное значение скорости передачи по каналу:

$$C_\partial=\max_{p(z)}\dot{I}(Z,V)=\max_{p(z)}V_\tau I(Z,V). \tag{8.6}$$

Здесь $p\left(z\right)$ – множество распределений вероятностей входных сигналов, формируемых источником. Если техническая скорость V_τ передачи элементов сообщений фиксирована, то пропускная способность может достигаться за счет изменения статистических свойств последовательностей символов посредством их преобразования (кодирования).

На практике предельные возможности канала обычно не достигаются. Степень загрузки канала характеризуется *коэффициентом использования*:

$$\lambda=\dot{I}(Z)/C_\partial\,,\ (0\le\lambda\le1),$$

где $\dot{I}(Z)$ – производительность источника сообщений.

8.5 Скорость передачи по непрерывному гауссову каналу связи

Под *гауссовым каналом связи* понимают математическую модель реального канала, удовлетворяющего следующим требованиям:

1) физические параметры канала известны и детерминированы;

2) полоса пропускания канала ограничена полосой F_k герц;

3) в канале действует аддитивный гауссов белый шум (с равномерным частотным спектром и нормальным распределением амплитуд);

4) статистическая связь между сигналом и шумом отсутствует, а ширина спектра сигнала и помехи ограничена полосой пропускания канала.

Предположим по указанному гауссову каналу (рис. 8.2) передается непрерывный сигнал $z_T(t)$ со средней мощностью $P_z = \sigma_z^2$. На выходе канала фиксируется сигнал $v_T(t)$, который искажен аддитивным гауссовым шумом $\xi(t)$ со средней мощностью $P_\xi = \sigma_\xi^2$.

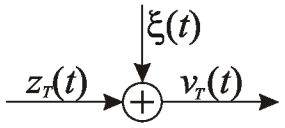

Рис. 8.2. Схема гауссова канала связи

Будем считать, что длительность T передаваемого сигнала достаточно велика, так что, в соответствии с теоремой Котельникова, можно заменить непрерывные реализации $z_T(t)$ и $v_T(t)$ последовательностями из $N = 2F_kT$ отсчетов, взятых через интервалы $\Delta t = (2F_k)^{-1}$, где F_k – полоса пропускания. Тогда среднее количество информации, передаваемой по каналу

$$I(\mathbf{Z},\mathbf{V}) = H(\mathbf{Z}) - H_V(\mathbf{Z}) = H(\mathbf{V}) - H_Z(\mathbf{V}), \tag{8.7}$$

где $H(\mathbf{V})$ и $H_Z(\mathbf{V})$ – априорная и апостериорная энтропии N-мерного случайного вектора $\mathbf{V}$.

По определению гауссова канала помеха аддитивна и статистически независима с входным сигналом, поэтому

$$H_Z(\mathbf{V}) = H_Z(\mathbf{Z}+\boldsymbol{\Xi}) = H_Z(\boldsymbol{\Xi}) = H(\boldsymbol{\Xi}), \tag{8.8}$$

где $H(\boldsymbol{\Xi})$ – энтропия N-мерного случайного вектора помехи, компонентами которого являются случайные величины в соответствующих сечениях непрерывного аддитивного гауссова белого шума $\xi(t)$.

Поскольку значения белого шума в моменты отсчетов некоррелированы,

$$H(\boldsymbol{\Xi}) = N \cdot h(\xi) = 2F_kT \cdot h(\xi), \tag{8.9}$$

где $h(\xi)$ – дифференциальная энтропия в среднем на один отсчет. В данном случае, поскольку шум распределен по нормальному закону:

$$p(\xi)=\left(\sigma_\xi\sqrt{2\pi}\right)^{-1}\exp\left(-\xi^2/2\sigma_\xi^2\right),$$

дифференциальная энтропия определяется как

$$\begin{aligned}h(\xi)&=-\int_{-\infty}^{+\infty}p(\xi)\log_2 p(\xi)d\xi=\\&=-\log_2\left(\sigma_\xi\sqrt{2\pi}\right)^{-1}\int_{-\infty}^{+\infty}p(\xi)d\xi+\frac{\log_2 e}{2\sigma_\xi^2}\int_{-\infty}^{+\infty}\xi^2 p(\xi)d\xi=\\&=\log_2\left(\sigma_\xi\sqrt{2\pi}\right)+\frac{1}{2}\log_2 e=\frac{1}{2}\log_2 2\pi e\sigma_\xi^2.\end{aligned}\qquad(8.10)$$

Энтропия $H(\mathbf{V})$ выражается аналогично (8.9) через дифференциальную энтропию $h(v)$ одного отсчета выходного сигнала:

$$H(\mathbf{V})=2F_kT\cdot h(v).\qquad(8.11)$$

Далее, подставляя энтропии, определяемые равенствами (8.9) и (8.11), в (8.7) с учетом (8.10) получаем следующее выражение для среднего количества информации, передаваемой по каналу:

$$I(\mathbf{Z},\mathbf{V})=2F_kT\left[h(v)-\frac{1}{2}\cdot\log_2 2\pi e\sigma_\xi^2\right].\qquad(8.12)$$

Соответственно, скорость передачи информации по непрерывному гауссову каналу связи определяется как

$$\dot{I}(\mathbf{Z},\mathbf{V})=2F_k\left[h(v)-\frac{1}{2}\cdot\log_2 2\pi e\sigma_\xi^2\right].\qquad(8.13)$$

8.6 Пропускная способность непрерывного гауссова канала связи

Пропускная способность непрерывного канала $C_н$ определяется как

$$C_н=\max_{p(\mathbf{Z})}\dot{I}(\mathbf{Z},\mathbf{V}).\qquad(8.14)$$

Следовательно, в соответствии с соотношением (8.13), для ее определения необходимо искать ансамбль входных сигналов, при котором дифференциальная энтропия $h(v)$ максимальна.

По предположению, в гауссовом канале связи средняя мощность сигнала и помехи ограничены. Ранее было показано, что при ограничении на величину дисперсии наибольшее значение $h(v)$ достигается в случае нормального распределения. Шум $\xi(t)$, по предположению, имеет нормальное распределение, следовательно, для того, чтобы выходной сигнал $v(t)$ имел нормальное распределение, необходимо, чтобы входной сигнал $z(t)$ также был нормальным и центрированным (поскольку центрированность сигнала при заданной средней мощности соответствует максимальному значению дисперсии).

Кроме того, входной сигнал, в пределах заданной достаточно широкой полосы частот F_k, должен иметь равномерный энергетический спектр. Только в этом случае можно говорить о независимости отсчетов. Заметим, что при этом средняя мощность выходного сигнала равна сумме средних мощностей входного сигнала и помехи:

$$P_v = \sigma_v^2 = \sigma_z^2 + \sigma_\xi^2 = P_z + P_\xi . \tag{8.15}$$

Если все указанные предположения выполняются, то с учетом (8.15)

$$\max h(v) = \frac{1}{2}\log_2 2\pi e P_v = \frac{1}{2}\log_2 2\pi e\left(\sigma_z^2 + \sigma_\xi^2\right) = \frac{1}{2}\log_2 2\pi e\left(P_z + P_\xi\right),$$

а пропускная способность непрерывного гауссова канала

$$C_H = F_k\left[\log_2 2\pi e\left(P_z + P_\xi\right) - \log_2 2\pi e P_\xi\right] = F_k \log_2\left(1 + \frac{P_z}{P_\xi}\right). \tag{8.16}$$

Представляет интерес установить, как зависит пропускная способность гауссова канала от ширины полосы пропускания. Произведем замену $P_\xi = P_o F_k$, где P_o – (удельная) мощность шума, приходящаяся на единицу частоты, и представим (8.16) в виде:

$$C_H = \left[\log_2\left(1 + P_z\gamma / P_o\right)\right]/\gamma , \tag{8.17}$$

где $\gamma = 1/F_k$. Вычисляя предел при $\gamma \to 0$ ($F_k \to \infty$) имеем

$$\lim_{F_k\to\infty} C_H = \lim_{\gamma\to 0}\frac{\log_2\left(1 + P_z\gamma / P_o\right)}{\gamma} = \frac{1{,}443 \cdot P_z}{P_o}. \tag{8.18}$$

График зависимости пропускной способности непрерывного гауссова канала связи от ширины полосы пропускания в соответствии с (8.17), (8.18) имеет вид, показанный на рис. 8.3.

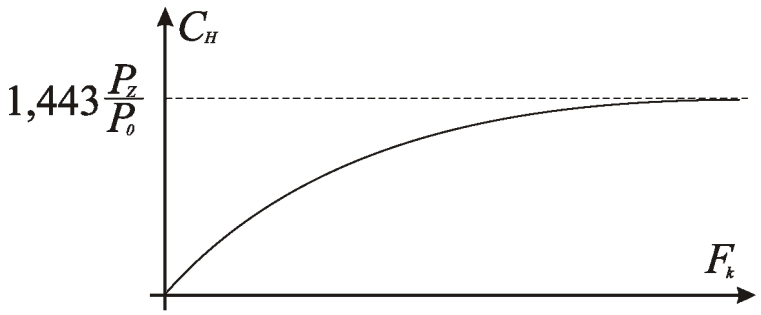

Рис. 8.3. Зависимость пропускной способности от полосы пропускания канала

8.7 Согласование физических характеристик сигнала и канала

Непрерывный канал характеризуется тремя параметрами:

1) ширина полосы пропускания сигнала F_k;
2) время T_k предоставления канала для передачи сигнала;
3) допустимое превышение сигнала над помехой, определяемое как $H_k = \log\left(P_{z,\max}/P_\xi\right)$,

где $P_{z,\max}$ – максимально допустимая мощность сигнала в канале.

Произведение указанных параметров называют объемом канала: $V_k = T_k F_k H_k$.

Аналогичными параметрами можно характеризовать сигнал:

1) ширина спектра частот сигнала F_c;
2) длительность сигнала T_c;
3) превышение сигнала над помехой $H_c = \log\left(P_z/P_\xi\right)$.

Соответственно, объём сигнала определяется как $V_c = T_c F_c H_c$, а необходимое условие его неискаженной передачи – $V_k \geq V_c$. Достаточные условия неискаженной передачи:

$T_k \geq T_c$, $F_k \geq F_c$, $H_k \geq H_c$.

Если выполнено необходимое условие неискаженной передачи, то достаточные условия всегда могут быть выполнены путем соответствующих преобразований сигнала.

Например, сигнал может быть записан с высокой скоростью, а затем передаваться по каналу с более низкой. При этом F_c уменьшится, а T_c увеличится соответствующим образом. Если превышение сигнала над помехой не удовлетворяет заданным ограничениям, то его снижают до допустимого уровня, но при этом должно произойти соответствующее увеличение времени передачи сигнала для обеспечения заданной верности.

Заметим, что в соответствии с (8.16) предельное количество информации, которое может быть передано по гауссовому каналу связи за время T_k:

$$I_{\max}(V,Z) = T_k F_k \log_2\left(1 + P_Z / P_\xi\right).$$

Замечательным является тот факт, что при $P_z / P_\xi >> 1$ это количество информации совпадает с объемом канала.

Лекция 9

Эффективное кодирование

9.1 Цель кодирования, основные понятия и определения

Как отмечалось в разделе 1.1, кодирование в широком смысле – преобразование сообщений в сигнал. Кодирование в узком смысле – представление дискретных сообщений определенными сочетаниями символов. Далее мы будем рассматривать кодирование только в узком смысле.

Кодирование осуществляется, с одной стороны, для того, чтобы обеспечить наилучшее согласование характеристик источника сообщений и канала, с другой стороны, для повышения достоверности передачи информации при наличии помех. Кроме того, при выборе системы кодирования (представления сообщений) стремятся обеспечить простоту и надежность аппаратной реализации устройств.

В процессе кодирования сообщений длинная последовательность (например, из N символов) обычно формируется из кодовых комбинаций, каждая из которых соответствует одному знаку (букве). Число n символов, из которых составлена такая кодовая комбинация, называется *значностью* или длиной кода. Количество разных символов m, использованных для построения кодовой комбинации, называется основанием кода. Физически символы реализуются в виде сигналов, несущих некоторые признаки. В качестве признаков могут использоваться, например, амплитуда, длительность импульсов и др.

Каждому кодируемому знаку можно приписать какой-либо порядковый номер. При этом задача кодирования сводится к представлению кодовых комбинаций числами в какой-либо системе счисления. Наиболее употребительной является позиционная система счисления, в которой значение цифры (символа) зависит от ее места (позиции).

Любое число $\mathbf{A}_n$ в позиционной системе счисления можно представить в виде:

$$\mathbf{A}_n = \sum_{i=1}^{n} a_{i-1} m^{i-1} = a_{n-1} m^{n-1} + a_{n-2} m^{n-2} + \ldots + a_1 m + a_0, \qquad (9.1)$$

где m – основание системы счисления, i – номер разряда, $i = \overline{1,n}$, a_i – коэффициент i-го разряда, принимающий целочисленные значения от 0 до $m-1$.

С точки зрения экономии времени передачи сообщений выгодно иметь меньше цифр в представлении числа. Однако увеличение m с целью уменьшения n приводит к усложнению устройств, реализующих m признаков (устойчивых состояний). Поэтому для характеристики эффективности систем используют произведение $n \times m$. Можно показать, что по этому критерию наиболее эффективной является троичная система. Тем не менее наиболее широко используются незначительно уступающие троичной системе двоичные коды.

Математическая запись двоичного кода, в соответствии с (9.1), имеет вид

$$\mathbf{A}_n = \sum_{i=1}^{n} a_{i-1} 2^{i-1}.$$

Максимально возможное число кодовых комбинаций простого двоичного кода $N_{\max} = 2^n$. Ниже приводятся используемые далее по тексту правила сложения, умножения и сложения по модулю ($\oplus$) в двоичной системе.

Сложение

	0	1
0	0	1
1	1	10

Умножение

	0	1
0	0	0
1	0	1

Сложение по модулю

	0	1
0	0	1
1	1	0

Помимо двоичной системы получили распространение системы с основанием, равным целой степени двойки (восьмеричная, шестнадцатиричная), которые легко сводятся как к двоичной, так и к десятичной, но дают более компактную запись. Например, в восьмеричной системе каждой из восьми цифр (0-7) ставится в соответствие трехразрядное двоичное число. В шестнадцатиричной системе перевод чисел в двоичную осуществляется путем замены каждой шестнадцатиричной цифры четырехразрядным двоичным числом.

Используются также двоично-десятичные коды, в которых каждую цифру десятичного числа записывают в виде четырехразрядного двоичного числа. Этот код относится к числу взвешенных кодов. Для фиксации цифр десятичного числа наибольшее практическое применение нашли коды 8-4-2-1; 7-4-2-1; 5-1-2-1 и 2-4-2-1. Цифры в названии кода выражают веса единиц в соответствующих разрядах.

При некоторых способах кодирования непрерывных сообщений (например, при преобразовании угла поворота диска с нанесенной на него маской в двоичный код) источником больших ошибок может быть одновременное изменение цифр в нескольких разрядах. Например, в простом двоичном коде одновременное изменение цифр в четырех разрядах имеет место при переходе от изображения (маски) цифры 7 к маске цифры 8. Для устранения этого явления используют специальные двоичные коды, у которых при переходе от изображения одного числа к изображению следующего соседнего числа изменяется значение цифры только одного разряда. При этом ошибка неоднозначности считывания не превышает единицы младшего разряда. К числу таких кодов относится код Грея.

9.2 Основная теорема Шеннона о кодировании для канала без помех

Эффективное кодирование сообщений, минимизирующее среднее число символов, требуемых для представления одного знака сообщения, опирается на следующую теорему (Шеннона):

1) при любой производительности источника сообщений, меньшей пропускной способности канала: $\dot{I}(Z) < C_д$, существует способ кодирования, позволяющий передавать по каналу все сообщения, вырабатываемые источником;
2) не существует способа кодирования, обеспечивающего передачу сообщений без их неограниченного накопления, если $\dot{I}(Z) > C_д$.

Справедливость теоремы покажем, опираясь на свойство асимптотической равномерности.

Пусть количество знаков в последовательности равно N, а энтропия источника $H(Z)$. Предположим также, что длина сообщения T велика и все сообщения являются типичными. Тогда для этих последовательностей выполняется неравенство (7.1):

$$\left|\frac{1}{N}\cdot\log\left(\frac{1}{p}\right)-H(Z)\right|<\mu,\ 0\leftarrow\mu>0,$$

а число типичных последовательностей $N_T=1/p$ в соответствии с ним будет

$$N_T=2^{NH(Z)}=2^{\frac{T}{\tau_u}H(Z)}=2^{T\dot{I}(Z)}. \qquad (9.2)$$

Здесь предполагается, что средняя длительность знака τ_u известна, поэтому $N=T/\tau_u$ и по определению $\dot{I}(Z)=H(Z)/\tau_u$.

Предположим, что кодирование осуществляется с использованием алфавита объемом m. Тогда с учетом того, что пропускная способность дискретного канала $C_\partial=(\log_2 m)/\tau_к$, число последовательностей длительности T (с числом знаков $N=T/\tau_к$), пропускаемых каналом, определится как:

$$N_k=m^N=m^{\frac{T}{\tau_k}}=2^{\frac{T}{\tau_k}\log_2 m}=2^{T\frac{\log_2 m}{\tau_k}}=2^{TC_\partial}. \qquad (9.3)$$

Сравнивая (9.2) и (9.3), нетрудно заметить, что если $\dot{I}(Z)<C_\partial$, то имеет место неравенство $N_k>N_T$. Это означает, что число последовательностей, пропускаемых каналом, достаточно, чтобы закодировать все типичные последовательности знаков. Вероятность появления нетипичных последовательностей при $N\to\infty$ стремится к 0, что и доказывает первую часть теоремы.

Справедливость второй части теоремы, указывающей на невозможность осуществить передачу при $\dot{I}(Z)>C_\partial$, следует из определения пропускной способности канала, как максимально достижимой скорости передачи информации. Поэтому в данном случае неизбежно накопление на передающей стороне.

9.3 Методы эффективного кодирования некоррелированной последовательности знаков, код Шеннона-Фано

Теорема Шеннона отвечает на вопрос: при каких условиях возможно, в принципе, построение кода, обеспечивающего передачу всех сообщений, формируемых источником. Естественно стремление строить наилучшие с точки зрения максимума передаваемой информации коды. Для того чтобы каждый символ (например, двоичного) кода нес максимум информации, символы кодовой комбинации должны быть *независимы* и принимать значения (0 и 1) с равными вероятностями. Построение эффективных кодов в случае *статистической независимости* символов сообщений опирается на методики Шеннона и Фано (код Шеннона-Фано).

Код строится следующим образом. Кодируемые знаки выписывают в таблицу в порядке убывания их вероятностей в сообщениях. Затем их разделяют на две группы так, чтобы значения сумм вероятностей в каждой группе были близкими. Все знаки одной из групп в соответствующем разряде кодируются, например, единицей, тогда знаки второй группы кодируются нулем. Каждую полученную в процессе деления группу подвергают вышеописанной операции до тех пор, пока в результате очередного деления в каждой группе не останется по одному знаку (табл. 9.1).

Таблица 9.1

Знаки	Вероятности	Коды
z_1	1/2	1
z_2	1/4	01
z_3	1/8	001
z_4	1/16	0001
z_5	1/32	00001
z_6	1/64	000001
z_7	1/128	0000001
z_8	1/128	0000000

В приведенном примере среднее число символов на один знак

$$l_{ср} = \sum_{i=1}^{8} p_i l_i = \sum_{i=1}^{7} \frac{i}{2^i} + \frac{7}{2^7} = \frac{127}{64},$$

где l_i – число символов в i-м разряде, имеет такую же величину, как и энтропия, рассчитанная в среднем на один знак:

$$H(z) = -\sum_{i=1}^{8} p_i \log_2 p_i = -\left(\sum_{i=1}^{7} \frac{1}{2^i} \log_2 2^{-i} + \frac{1}{2^7} \log_2 2^{-7} \right) = \frac{127}{64}.$$

Совпадение результатов связано с тем, что вероятности знаков являются целочисленными отрицательными степенями двойки. В общем случае

$$l_{ср} \geq H(z).$$

Если величина среднего числа символов на знак оказывается значительно большей, чем энтропия, то это говорит об избыточности кода. Эту избыточность можно устранить, если перейти к кодированию блоками. Рассмотрим простой пример кодирования двумя знаками z_1, z_2 с вероятностями их появления в сообщениях 0,1 и 0,9 соответственно.

Если один из этих знаков кодировать, например, нулем, а другой единицей, т.е. по одному символу на знак, имеем соответственно

$$l_{ср} = 0{,}1 \cdot 1 + 0{,}9 \cdot 1 = 1{,}0,$$

$$H(z) = -0{,}1 \cdot \log_2 0{,}1 - 0{,}9 \cdot \log_2 0{,}9 = 0{,}47.$$

При переходе к кодированию блоками по два знака (табл. 9.2)

$$l_{ср} = \frac{l_{ср,бл}}{2} = \frac{1}{2}\left(0{,}81 \cdot 1 + 0{,}09 \cdot 2 + 0{,}09 \cdot 3 + 0{,}01 \cdot 3\right) = 0{,}645.$$

Таблица 9.2

Блоки	Вероятности	Коды
$z_1\ z_1$	0,81	1
$z_2\ z_1$	0,09	01
$z_1\ z_2$	0,09	001
$z_2\ z_2$	0,01	000

Можно проверить, что при кодировании блоками по три символа среднее число символов на знак уменьшается и оказывается равным около 0,53. Эффект

достигается за счет того, что при укрупнении блоков, группы можно делить на более близкие по значениям суммарных вероятностей подгруппы. Вообще, $\lim_{n\to\infty} l_{ср} = H(z)$, где n – число символов в блоке.

9.4 Методика кодирования Хаффмана

Рассмотренная выше методика кодирования не всегда приводит к хорошему результату, вследствие отсутствия четких рекомендаций относительно того, как делить множество кодируемых знаков на подгруппы. Рассмотрим методику кодирования Хаффмана, которая свободна от этого недостатка.

Кодируемые знаки, также как при использовании метода Шеннона-Фано, располагают в порядке убывания их вероятностей (табл. 9.3). Далее на каждом этапе две последние позиции списка заменяются одной и ей приписывают вероятность, равную сумме вероятностей заменяемых позиций. После этого производится пересортировка списка по убыванию вероятностей, с сохранением информации о том, какие именно знаки объединялись на каждом этапе. Процесс продолжается до тех пор, пока не останется единственная позиция с вероятностью, равной 1.

Таблица 9.3

Знаки	p_i	Вспомогательные столбцы						
		1	2	3	4	5	6	7
z_1	0,22	0,22	0,22	0,26	0,32	0,42	**0,58**	**0,1**
z_2	0,2	0,2	0,2	0,22	0,26	**0,32**	**0,42**	
z_3	0,16	0,16	0,16	0,2	**0,22**	**0,26**		
z_4	0,16	0,16	0,16	**0,16**	**0,2**			
z_5	0,1	0,1	**0,16**	**0,16**				
z_6	0,1	**0,1**	**0,1**					
z_7	**0,04**	**0,06**						
z_8	**0,02**							

После этого строится кодовое дерево. Корню дерева ставится в соответствие узел с вероятностью, равной 1. Далее каждому узлу приписываются два потомка с вероятностями, которые участвовали в формировании значения вероят-

ности обрабатываемого узла. Так продолжают до достижения узлов, соответствующих вероятностям исходных знаков.

Процесс кодирования по кодовому дереву осуществляется следующим образом. Одной из ветвей, выходящей из каждого узла, например, с более высокой вероятностью, ставится в соответствие символ 1, а с меньшей – 0. Спуск от корня к нужному знаку дает код этого знака. Правило кодирования в случае равных вероятностей оговаривается особо. Таблицы 9.3, 9.4 и рис. 9.1 иллюстрируют применение методики Хаффмана. Жирным шрифтом в табл. 9.3 выделены объединяемые позиции, подчеркиванием – получаемые при объединении позиции.

Таблица 9.4

Знаки	Коды
z_1	01
z_2	00
z_3	111
z_4	110
z_5	100
z_6	1011
z_7	10101
z_8	10100

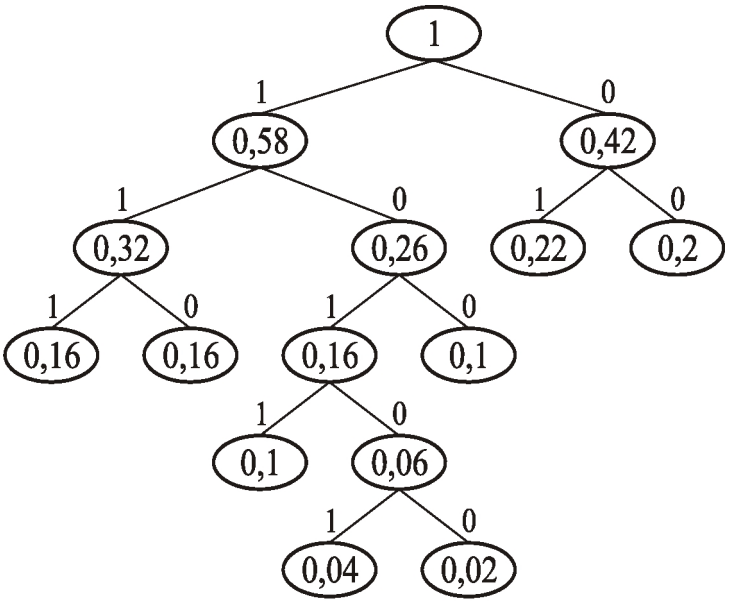

Рис. 9.1. Кодовое дерево

Замечательным свойством кодов, построенных с применением методик Шеннона-Фано или Хаффмана, является их префиксность. Оно заключается в том, что ни одна комбинация кода не является началом другой, более длинной комбинации. Это позволяет при отсутствии ошибок осуществлять однозначное декодирование ряда следующих друг за другом кодовых комбинаций, между которыми отсутствуют разделительные символы.

9.5 Методы эффективного кодирования коррелированной последовательности знаков

Ранее было показано, что повышение производительности источников и каналов достигается путем формирования и передачи шумоподобных сигналов (символы независимы друг от друга и равномерно распределены). Это свойство может не соблюдаться, если знаки в сообщениях коррелированны. Для повышения эффективности кодирования коррелированной последовательности искусственно производят декорреляцию.

Один из способов заключается в укрупнении алфавита знаков. При этом передаваемые сообщения разбиваются на двух-, трех- или n-знаковые сочетания (непересекающиеся блоки), вероятности которых известны. Каждое сочетание кодируется одним из описанных выше способов:

$$\underbrace{z_1 z_1 z_3 \ldots . z_4 z_1 z_2}_{n}\ \underbrace{z_3 z_1 z_2 \ldots z_4 z_1 z_3}_{n}\ .$$

При увеличении числа знаков в сочетаниях корреляция знаков в сообщении уменьшается. Однако при этом возрастает задержка в передаче сигналов на время формирования сочетаний.

От этого недостатка в некоторой степени свободен метод, в котором каждое сочетание из l знаков (l-грамма) формируется путем добавления текущего знака сообщения и отбрасывания последнего знака l-граммы:

$$\overbrace{}^{\text{2-я } l\text{-грамма}}$$
$$z_1 z_3 z_2 \ldots z_4 z_2 z_1 z_3 z_1 z_2 \ldots$$
$$\underbrace{}_{\text{1-я } l\text{-грамма}}$$

Сочетание из двух знаков называют диграммой, из трех – триграммой и т.д.

В процессе кодирования l-грамма непрерывно перемещается по тексту сообщения, а кодовое обозначение каждого знака сообщения зависит от $l-1$ предшествующих знаков и может быть определено с использованием методик Шеннона-Фано или Хаффмана. Задержка сигнала в данном случае имеет место лишь на начальном этапе формирования первой l-граммы.

9.6 Недостатки методов эффективного кодирования

1. *Различия в длине кодовых комбинаций.* Обычно знаки на вход устройства кодирования поступают через равные промежутки времени. Если им соответствуют комбинации различной длины, то для обеспечения полной загрузки канала при передаче без потерь необходимо предусмотреть буферное устройство, как на передающей, так и на приемной стороне.

2. *Задержка в передаче информации.* Как было показано, достоинства эффективного кодирования проявляются в полной мере при кодировании длинными блоками. Для этого необходимо накапливать знаки, как при кодировании, так и при декодировании, что приводит к значительным задержкам.

3. *Низкая помехозащищенность.* Даже одиночная ошибка, возникшая в процессе передачи, может нарушить свойство префиксности кода и повлечь за собой неправильное декодирование ряда последующих комбинаций. Это явление называют *треком* ошибки.

4. *Сложность технической реализации.* Использование буферных устройств, для обеспечения равномерной загрузки канала, при разной длине кодовых комбинаций и организация кодирования блоками для повышения эффективности приводят к усложнению реализации систем эффективного кодирования. Если вдобавок применяются некоторые аппаратные решения, обеспечивающие повышение помехозащищенности, то все это в совокупности может свести на нет основное достоинство систем эффективного кодирования, связанное с тем, что знаки, имеющие большую вероятность, кодируются более короткими кодовыми словами.

Лекция 10

Введение в теорию помехоустойчивого кодирования

10.1 Теорема Шеннона о кодировании для канала с помехами

Теоретической основой помехоустойчивого кодирования является следующая теорема (Шеннона):

1) при любой производительности источника меньшей, чем пропускная способность канала, существует способ кодирования, который позволяет обеспечить передачу всей информации от источника со сколь угодно малой вероятностью ошибки;
2) не существует способа кодирования, позволяющего вести передачу информации со сколь угодно малой вероятностью ошибки, если производительность источника больше пропускной способности канала.

Доказательство. Пусть источник генерирует типичные (разрешенные) последовательности большой длительности T, с числом символов $N = T/\tau_u$, где τ_u – среднее время формирования одного символа. Тогда справедливо неравенство (7.1), а число типичных последовательностей в соответствии с (9.2)

$$N_T(Z) = 2^{NH(Z)} = 2^{\frac{T}{\tau_u}H(Z)}. \qquad (10.1)$$

Если предположить, что последовательности формируются из символов алфавита объемом m так, что символы статистически независимы, то общее число возможных последовательностей длительности T, которые могут быть в принципе сформированы на входе канала:

$$N(Z) = m^N = 2^{N\log_2 m} = 2^{\frac{T}{\tau_k}\log_2 m}, \qquad (10.2)$$

где τ_k – среднее время передачи одного символа по каналу связи.

Пусть выполняется условие первой части теоремы – пропускная способность канала больше производительности источника:

$$C_д > \dot{I}(Z) = H(Z)/\tau_u. \qquad (10.3)$$

В соответствии с (8.6), пропускная способность дискретного канала

$$C_{\partial} = \max_{p(z)} \frac{I(Z,V)}{\tau_k} = \frac{\max_{p(z)} H(Z) - H_V(Z)}{\tau_k} = \frac{\log_2 m - H_V(Z)}{\tau_k}. \qquad (10.4)$$

Подставляя правую часть (10.4) в левую часть неравенства (10.3), имеем

$$\frac{\log_2 m - H_V(Z)}{\tau_к} > \frac{H(Z)}{\tau_и}.$$

Ничего не изменится, если умножить обе части последнего равенства на T:

$$\frac{T}{\tau_к}\left(\log_2 m - H_V(Z)\right) > \frac{T}{\tau_и} H(Z). \qquad (10.5)$$

Поскольку условная энтропия $H_V(Z) \geq 0$, при ее отбрасывании неравенство (10.5) только усилится:

$$\frac{T}{\tau_к}\log_2 m >> \frac{T}{\tau_и} H(Z).$$

Нетрудно заметить, что левая и правая части последнего неравенства суть показатели степени в (10.1) и (10.2), следовательно, в силу свойства степеней

$$2^{\frac{T}{\tau_к}\log_2 m} >> 2^{\frac{T}{\tau_и}H(Z)},$$

откуда следует

$$N(Z) >> N_T(Z). \qquad (10.6)$$

Это означает, что существует $C_{N(Z)}^{N_T(Z)}$ различных способов кодирования, позволяющих каждой типичной последовательности поставить в соответствие последовательность из множества $N(Z)$. При равновероятном выборе последовательностей из этого множества вероятность p того, что данная последовательность окажется разрешенной:

$$p = \frac{N_T(Z)}{N(Z)} = \frac{2^{T\frac{H(Z)}{\tau_и}}}{2^{T\frac{\log_2 m}{\tau_к}}} = \frac{1}{2^{T\left(\frac{\log_2 m}{\tau_к} - \frac{H(Z)}{\tau_и}\right)}}. \qquad (10.7)$$

При получении на выходе канала конкретной последовательности v остается неопределенность относительно переданной последовательности z, свя-

занная с $H_V(Z)$, которая определяется уровнем шумов в канале. Эта неопределенность эквивалентна неопределенности выбора из

$$N_V(Z) = 2^{(T/\tau_k)H_V(Z)} \tag{10.8}$$

последовательностей. Заметим, что соотношение (10.8) может быть получено по аналогии с (10.1).

Конкретная последовательность может быть идентифицирована со сколь угодно малой вероятностью ошибки, если среди $N_V(Z)$ последовательностей она оказалась единственной разрешенной. Отсюда, в частности, следует, что любой способ кодирования и декодирования должен заключаться в разбиении всего множества последовательностей на подмножества, каждое из которых содержит лишь одну разрешенную.

Оценим среднюю по всем возможным способам кодирования вероятность $\overline{p}$ того, что ни одна из $N_V(Z)-1$ последовательностей не является разрешенной:

$$\overline{p} = (1-p)^{N_V(Z)-1}. \tag{10.9}$$

Здесь p – вероятность (10.7) того, что данная последовательность разрешенная.

Поскольку $(1-p)<1$, вместо равенства (10.9) можно записать неравенство

$$\overline{p} > (1-p)^{N_V(Z)}. \tag{10.10}$$

Разложим правую часть (10.10) в ряд Тейлора в окрестности $p=0$:

$$(1-p)^{N_V(Z)} = 1 - N_V(Z)p + \frac{1}{2}N_V(Z)\left(N_V(Z)-1\right)p^2 - \ldots$$

Можно показать, что члены этого ряда убывают по абсолютной величине. По признаку Лейбница, если ряд знакопеременный и члены убывают по абсолютной величине, то величина остатка не превышает величину первого отбрасываемого члена и имеет с ним одинаковый знак.

Таким образом, если ограничиться двумя первыми членами, неравенство (10.10) только усилится:

$$\overline{p} > 1 - N_V(Z)p, \tag{10.11}$$

где p – вероятность, определяемая в (10.7). Прежде чем осуществить ее замену в (10.11), несколько преобразуем (10.7). Для этого воспользуемся неравенством (10.5), которое представим в виде

$$\frac{T}{\tau_к}\log_2 m - \frac{T}{\tau_и}H(Z) > \frac{T}{\tau_к}H_V(Z). \qquad (10.12)$$

Добавив в правую часть (10.12) некоторое положительное число η, это неравенство превратим в равенство:

$$T\left(\frac{\log_2 m}{\tau_к} - \frac{H(Z)}{\tau_и}\right) = T\left(\frac{H_V(Z)}{\tau_к} + \eta\right). \qquad (10.13)$$

Наконец, заменив показатель степени двойки в (10.7) правой частью из (10.13), получим:

$$p = \frac{1}{2^{T\left(\frac{H_V(Z)}{\tau_к}+\eta\right)}} = \frac{1}{2^{\frac{TH_V(Z)}{\tau_к}}2^{T\eta}} = \frac{1}{N_V(Z)\cdot 2^{T\eta}}. \qquad (10.14)$$

Подставив полученное значение вероятности p в (10.11), получаем

$$\overline{p} > 1 - \frac{1}{2^{T\eta}}. \qquad (10.15)$$

Напомним, что $\overline{p}$ – вероятность того, что ни одна из $N_V(Z)-1$ не является разрешенной (следовательно, одна из $N_V(Z)$ последовательностей является разрешенной). Тогда вероятность ошибки:

$$\overline{p}_{ош} = 1 - \overline{p} < 2^{-T\eta}. \qquad (10.16)$$

Из (10.16) видно, что $\overline{p}_{ош} \to 0$ при $T \to \infty$. Таким образом, всегда можно подобрать длину последовательности такую, что средняя вероятность ошибки окажется сколь угодно малой по всем способам кодирования. Вторую часть теоремы примем без доказательства.

Теорема имеет важное теоретическое значение. Хотя в ней не объясняется, как строить коды, она обосновывает принципиальную возможность построения кодов, обеспечивающих передачу со сколь угодно высокой точностью. Теорема опровергает интуитивно казавшееся правильным предположение, что безошибочная передача в канале с помехами невозможна. Из (10.16) следует, что при

безграничном увеличении длительности T сообщений может быть достигнута как угодно высокая точность передачи. Конечно, безошибочная передача при наличии помех возможна лишь теоретически, т.к. нельзя безгранично увеличивать длительность кодируемой последовательности.

10.2 Общие принципы построения помехоустойчивых кодов

Повышение достоверности передачи и хранения информации достигается введением избыточности (дополнительных символов). При выборе этих символов используются условия, проверка которых при декодировании дает возможность обнаруживать и исправлять ошибки. Коды, обладающие этим свойством, называют помехоустойчивыми.

Обычно указанные условия связаны с алгебраической структурой кода, при этом соответствующий код называют *алгебраическим*. Алгебраические коды могут строиться как блоковые или непрерывные. В случае блоковых кодов процедура кодирования заключается в сопоставлении k информационным символам, соответствующих кодируемому знаку, блока из n символов. Если n постоянно для всех знаков кодируемого сообщения, блоковый код называют равномерным.

Предположим, что на вход кодирующего устройства поступает последовательность из k (соответствующих кодируемому знаку) информационных символов, которые преобразуются в кодовую комбинацию из n символов, причем $n > k$. Всего возможно 2^k различных входных и 2^n выходных последовательностей. Среди указанных выходных последовательностей только 2^k так называемых *разрешенных* последовательностей, соответствующих входным информационным последовательностям. Остальные $2^n - 2^k$ комбинаций являются *запрещенными*. Ясно, что любая из 2^k разрешенных комбинаций может быть трансформирована помехой в любую из 2^n комбинаций. При этом возможны следующие случаи;

1) 2^k случаев безошибочной (неискаженной) передачи;

2) $2^k(2^n - 2^k)$ случаев, когда разрешенные комбинации помехой трансформируются в запрещенные, но обнаруживаемые;

3) $2^k(2^k - 1)$ случаев перехода в другие разрешенные комбинации. Такие ошибки не могут быть обнаружены.

Поскольку всего случаев передачи $2^k \cdot 2^n$, относительное число обнаруживаемых ошибок (вероятность обнаружения ошибки) составит

$$p = \frac{2^k(2^n - 2^k)}{2^k 2^n} = 1 - \frac{1}{2^{n-k}}.$$

Нетрудно заметить, что при $n \to \infty$ вероятность обнаружения ошибки стремится к единице. Из соображений простоты реализации число $n - k$ проверочных разрядов, характеризующих избыточность кода, ограничивают.

Избыточность является одной из основных характеристик помехоустойчивого кода. Относительную избыточность определяют как

$$R_1 = \frac{n-k}{n} \text{ или } R_\infty = \frac{n-k}{k}.$$

При $n \to \infty$ предельное значение R_1 равно 1, а R_∞ – бесконечности.

Процедуры определения проверочных символов обычно строятся как линейные операции над определенными информационными символами. Поэтому эти коды называют линейными.

10.3 Математическое введение к линейным кодам

Кодовые комбинации можно рассматривать как элементы некоторого множества. Множество элементов, в котором определена одна основная операция, выполняются аксиомы замкнутости и ассоциативности, имеется нулевой (если основная операция – сложение) или единичный (если основная операция – умножение) и для всякого элемента существует противоположный (обратный) элемент называется *группой*.

Если основная операция коммутативна, группа называется *коммутативной* или *абелевой*. Число элементов в конечной группе называют *порядком* группы. Для построения двоичных кодов используется коммутативная опера-

ция *сложения по модулю* 2, при выполнении которой число разрядов кода не увеличивается. Поэтому множество n-разрядных комбинаций двоичного кода является конечной абелевой группой.

Подмножество группы, само являющееся группой относительно операции, заданной в группе, называют *подгруппой*. Пусть в абелевой группе $\mathbf{G}$ задана подгруппа $\mathbf{A}$ и элемент $b_j \notin \mathbf{A}$. Множество элементов, образованное как суммы (по модулю 2) элемента b_j с каждым из элементов подгруппы $\mathbf{A}$: $b_j \oplus \mathbf{A} = \{b_j \oplus a,\ a \in \mathbf{A}\}$ называется *смежным классом*, а сам элемент $b_j \notin \mathbf{A}$ – *образующим элементом*. Задавая образующие элементы группы так, чтобы они не входили в уже образованные классы, можно разложить всю группу на смежные классы по подгруппе $\mathbf{A}$.

Заметим, что в соответствии с теоремой Шеннона любой метод кодирования можно рассматривать, как правило разбиения множества запрещенных кодовых комбинаций на 2^k непересекающихся подмножества, в каждом из которых лишь одна разрешенная комбинация. Операция разложения на классы смежности указывает формальное правило такого разбиения. Далее нам понадобятся также понятия кольца и поля.

Лекция 11

Построение групповых кодов

11.1 Понятие корректирующей способности кода

Кодовое расстояние d выражается числом символов, в которых последовательности отличаются друг от друга. Для определения кодового расстояния между двумя комбинациями двоичного кода достаточно сложить их по модулю 2 и подсчитать число единиц в полученном результате. Минимальное расстояние, подсчитанное по всем парам разрешенных кодовых комбинаций, называют *минимальным кодовым расстоянием* данного кода.

Вес (*Хэмминга*) кодовой последовательности определяется как число ненулевых компонент этой последовательности. Ясно, что кодовое расстояние между двумя последовательностями равно весу некоторой третьей последовательности, являющейся их суммой, которая (в силу свойства операции сложения по модулю два) также обязана быть последовательностью данного кода. Следовательно, минимальное кодовое расстояние для линейного кода равно минимальному весу его ненулевых векторов.

Вектором ошибок называют n-разрядную двоичную последовательность, содержащую единицы в разрядах, подверженных ошибкам, и нули в остальных разрядах. Любая искаженная комбинация может рассматриваться как результат сложения по модулю 2 исходной разрешенной комбинации и вектора ошибки.

Число r искаженных символов кодовой комбинации называют *кратностью ошибки*. При кратности ошибок r всего может быть C_n^r n-разрядных двоичных векторов ошибок. Ошибки символов, при которых вероятность появления любой комбинации зависит только от числа r искаженных символов и вероятности p искажения одного символа, называют взаимно независимыми. При взаимно независимых ошибках вероятность искажения любых r символов в n-разрядной кодовой комбинации

$$p_r = C_n^r p^r (1-p)^{n-r}.$$

Корректирующая способность кода характеризуется значениями кратности r ошибок, которые обнаруживаются, и кратностью s ошибок, которые могут исправляться корректирующим кодом. Подчеркнем, что конкретный корректирующий код не обязан исправлять любую комбинацию ошибок. Он может обнаруживать и исправлять лишь ошибки заданной кратности, которые принимались в расчет при его построении.

11.2 Общая схема построения группового кода

Исходными данными для построения группового кода являются: объем кода Q (количество передаваемых дискретных сообщений) и заданная корректирующая способность. Задача заключается в определении числа разрядов n кода и правила формирования проверочных разрядов.

Количество информационных разрядов k по заданному Q определяется из условия

$$2^k - 1 \geq Q \qquad (11.1)$$

(здесь учтено, что нулевая комбинация обычно не используется, т.к. не изменяет состояния канала связи). Далее каждой из этих $2^k - 1$ ненулевых информационных последовательностей необходимо поставить в соответствие n-разрядный избыточный код (разрешенную комбинацию).

Множество 2^k n-разрядных разрешенных комбинаций (вместе с нулевой) образует подгруппу группы всех 2^n n-разрядных комбинаций. Разложим группу на смежные классы по этой подгруппе. В качестве образующих элементов смежных классов примем векторы ошибок, которые мы намерены исправлять.

Если, например, ставится задача исправлять все одиночные ошибки (кратность $s = 1$), то в качестве образующих элементов должны быть взяты n разных векторов, содержащих по одной единице в одном из n разрядов. Если кроме одиночных необходимо исправлять также все двойные ошибки (кратность $s = 2$), то добавится $C_n^s = C_n^2$ векторов ошибок (образующих элементов классов) и т.д.

Кроме самой подгруппы разрешенных комбинаций в результате разложения группы всех n-разрядных комбинаций может быть образовано $2^{n-k}-1$ непересекающихся смежных классов. Если число подлежащих исправлению векторов ошибок не превышает числа смежных классов, каждому из них можно поставить в соответствие некоторый класс смежности. Таким образом, для того чтобы обеспечивалась возможность определения и исправления ошибок кратности до s включительно, в общем случае должно выполняться неравенство

$$2^{n-k}-1 \geq C_n^1 + C_n^2 + \ldots + C_n^s$$

или

$$2^{n-k} \geq \sum_{i=0}^{s} C_n^i . \qquad (11.2)$$

В соответствии с (11.2) число разрядов корректирующего кода, предназначенного для исправления ошибок кратности s, определяется неравенством:

$$n \geq k + \log_2 \sum_{i=0}^{s} C_n^i . \qquad (11.3)$$

Для исправления ошибок необходимо определить, какому классу смежности принадлежит принятая кодовая последовательность, а затем соответствующий этому классу образующий элемент (вектор ошибки) сложить (по модулю два) с принятой последовательностью. Для определения класса смежности каждому из них ставится в соответствие последовательность $n-k$ символов, называемая опознавателем или синдромом. Исправление ошибок возможно лишь при взаимнооднозначном соответствии между множеством смежных классов (векторов ошибок) и множеством опознавателей.

11.3 Связь корректирующей способности с кодовым расстоянием

Обычно декодирование осуществляется таким образом, что любая принятая запрещенная кодовая комбинация отождествляется с разрешенной комбинацией, находящейся от неё на минимальном кодовом расстоянии. Если минимальное кодовое расстояние данного кода $d=1$, т.е. все комбинации кода являются разрешенными, то обнаружить ошибку не удастся. Если $d=2$, то удастся обнаружить единичную ошибку и т.д. В общем случае при необходимости

обнаружения ошибки кратности до r включительно минимальное кодовое расстояние должно удовлетворять условию

$$d_{\min} \geq r+1. \qquad (11.4)$$

Для исправления ошибок кратности s, в соответствии с описанной в разделе 11.2 общей схемой построения группового кода, каждой разрешенной кодовой комбинации необходимо поставить в соответствие подмножество запрещенных комбинаций так, чтобы эти подмножества не пересекались. Для этого должно выполняться неравенство

$$d_{\min} \geq 2s+1. \qquad (11.5)$$

Число комбинаций, расположенных на расстоянии i от заданной разрешенной, равно C_n^i. Следовательно, при выполнении условия (11.5) число исправляемых ошибок будет равно числу запрещенных комбинаций, находящихся в подмножестве, соответствующем разрешенной комбинации: $\sum_{i=1}^{s} C_n^i$.

Для исправления ошибок кратности s и одновременного обнаружения всех ошибок кратности r $(r \geq s)$ минимальное кодовое (*хэммингово*) расстояние должно удовлетворять неравенству

$$d_{\min} \geq r+s+1. \qquad (11.6)$$

Дадим геометрическую трактовку приведенным выше соотношениям.

Любая n-разрядная двоичная кодовая комбинация может быть интерпретирована как вершина n-мерного гиперкуба с длиной ребра равной 1. Например, при $n=2$ это квадрат, при $n=3$ – единичный куб. В общем случае n-мерный гиперкуб содержит 2^n вершин, что совпадает с возможным числом n-разрядных двоичных кодовых комбинаций.

Кодовое расстояние можно интерпретировать как наименьшее число ребер, которое надо пройти, чтобы попасть из одной разрешенной комбинации в другую. В подмножество каждой разрешенной комбинации, в соответствии с (11.5), относят все вершины, оказавшиеся в сфере радиуса

$$s \leq (d-1)/2. \qquad (11.7)$$

Если в результате действия шума разрешенная комбинация переходит в точку, принадлежащую сфере, то она может быть исправлена.

11.4 Построение опознавателей ошибок

В соответствии с общей схемой построения группового кода, каждой из $2^k - 1$ ненулевых информационных последовательностей ставится в соответствие n-разрядная разрешенная кодовая комбинация, в которой $n - k$ символов проверочные. Они должны быть заполнены опознавателями так, чтобы имело место взаимнооднозначное соответствие множеств исправляемых ошибок (классов смежности) и опознавателей.

Предположим, что двоичный код, предназначенный для исправления всех ошибок кратности до s включительно, построен так, что в (11.2), (11.3) имеет место равенство:

$$2^{n-k} - 1 = \sum_{i=1}^{s} C_n^i .$$

В частности, если исправлению подлежат только одиночные ошибки, имеем

$$2^{n-k} - 1 = n .$$

Этому равенству удовлетворяют, например, $n = 7$ и $k = 4$. Для указанных значений можно построить $2^7 / 2^4 - 1 = 2^3 - 1 = 7$ классов смежности. Каждому из этих семи классов смежности можно поставить в соответствие трехразрядный опознаватель вектора ошибки. В данном случае в качестве опознавателей можно взять двоичные числа, указывающие номер разряда, в котором произошла ошибка (табл. 11.1).

Таблица 11.1

Вектор ошибки	№ разряда	Опознаватель
0000001	1	001
0000010	2	010
0000100	3	011
0001000	4	100
0010000	5	101
0100000	6	110
1000000	7	111

При построении опознавателей ошибок более высокой кратности (векторы ошибок имеют единицы в нескольких разрядах) их можно строить как суммы по модулю два опознавателей одиночных ошибок. При этом (выбирая

очередной опознаватель одиночной ошибки в следующем разряде), необходимо следить за тем, что очередная кодовая комбинация и формируемые с ее использованием опознаватели векторов ошибок более высокой кратности еще не использованы в качестве опознавателей одиночных и кратных ошибок в предшествующих разрядах. Такую проверку необходимо делать при переходе к одиночной ошибке каждого следующего разряда.

Заметим, что при использовании указанной процедуры формирования опознавателей для составления проверочных равенств, о которых пойдет речь в следующем разделе, достаточно знать лишь опознаватели одиночных ошибок в каждом разряде.

11.5 Определение проверочных равенств и уравнений кодирования

Как указывалось выше, для обеспечения возможности исправления ошибок на этапе построения кода необходимо обеспечить взаимнооднозначное соответствие между множеством векторов ошибок, смежных классов и множеством опознавателей.

На этапе декодирования процедура определения символов опознавателя реализуется с использованием так называемых проверочных равенств как проверка на четность. При отсутствии ошибок в декодируемой последовательности, в результате всех проверок на четность должен получиться опознаватель из одних нулей. При наличии ошибок в соответствующих разрядах опознавателя появляются единицы. Рассмотрим общие принципы построения проверочных равенств и уравнений кодирования. Для наглядности изложение проведем на примере построенного выше кода (7, 4), который носит исключительно иллюстративный характер.

Разряды, которые должны входить в каждую из проверок на четность, определяются по таблице опознавателей (табл. 11.1). В коде (7, 4) число проверочных разрядов, а следовательно, и проверочных равенств должно быть три: $n-k=7-4=3$.

В данном случае в качестве опознавателей взяты двоичные коды номеров разрядов, в которых произошла ошибка. Каждое проверочное равенство строится по значениям символов в соответствующем этому равенству разряде опознавателей. Единица в первом (младшем) разряде опознавателя является следствием ошибки в одном из следующих разрядов: 1, 3, 5, 7; поэтому в качестве первого проверочного равенства можно взять

$$a_1 \oplus a_3 \oplus a_5 \oplus a_7 = 0. \qquad (11.8)$$

Единица во втором разряде опознавателей является следствием ошибки в одном из следующих разрядов: 2, 3, 6, 7. Поэтому второе проверочное равенство определяется как

$$a_2 \oplus a_3 \oplus a_6 \oplus a_7 = 0. \qquad (11.9)$$

Аналогично, единица в третьем разряде опознавателей является следствием ошибки в одном из следующих разрядов: 4, 5, 6, 7; т.е. третье проверочное равенство можно записать в виде:

$$a_4 \oplus a_5 \oplus a_6 \oplus a_7 = 0. \qquad (11.10)$$

Номера проверочных разрядов целесообразно выбирать так, чтобы каждый из них входил только в одно *проверочное равенство* (11.8)-(11.10). Это обеспечит однозначное определение значений символов в проверочных разрядах при кодировании:

$$\left.\begin{aligned} a_1 &= a_3 \oplus a_5 \oplus a_7 \\ a_2 &= a_3 \oplus a_6 \oplus a_7 \\ a_4 &= a_5 \oplus a_6 \oplus a_7 \end{aligned}\right\}. \qquad (11.11)$$

В данном случае проверочными будут первый, второй и четвертый разряд, которые заполняются на этапе формирования разрешенных комбинаций в соответствии с *уравнениями кодирования* (11.11).

В рассматриваемом примере $d_{\min} = 3$, поэтому, в соответствии с (11.4), (11.5), данный код может использоваться либо для обнаружения единичных и двойных ошибок, либо для исправления одиночных ошибок. Из табл.11.1. видно, что сумма любых двух опознавателей единичных ошибок дает ненулевой

опознаватель, который и может использоваться для обнаружения двойной ошибки.

Для того чтобы одновременно исправлять одиночные и обнаруживать двойные ошибки, необходимо, в соответствии с (11.6), построить код с $d_{\min} \geq 4$, например, путем добавления еще одного (8-го разряда) для дополнительной проверки на четность.

Лекция 12

Циклические коды

12.1 Математическое введение к циклическим кодам

Математическим аппаратом циклических кодов является теория колец. Множество $\mathbf{F}$ называется *кольцом*, если для любой пары элементов из $\mathbf{F}$ определены операции сложения и умножения, множество $\mathbf{F}$ является аддитивной абелевой группой, а также выполняются аксиомы замкнутости, ассоциативности и дистрибутивности. Подмножество элементов кольца $\mathbf{F}$, само являющееся кольцом относительно операций в $\mathbf{F}$, называют подкольцом.

Подкольцо $\mathbf{I}$ аддитивной группы $\mathbf{F}$ называется *идеалом*, если для любого α из $\mathbf{R}$ и любого β из $\mathbf{I}$ элемент $\alpha\beta$ принадлежит $\mathbf{I}$. Если все элементы $\mathbf{I}$ кратны некоторому элементу кольца α, он называется главным идеалом, а α – образующим элементом идеала.

Кольцо коммутативно, если $\alpha\beta = \beta\alpha$. Коммутативное кольцо $\mathbf{F}$ называется полем, если выполняются аксиомы:

1) кольцо $\mathbf{F}$ содержит элемент 1 такой, что для любого α из $\mathbf{F}$ $1 \cdot \alpha = \alpha \cdot 1 = \alpha$;
2) для любого $\alpha \in \mathbf{F}$ существует $\alpha^{-1} \in \mathbf{F}$ такой, что $\alpha \cdot \alpha^{-1} = \alpha^{-1}\alpha = 1$.

Таким образом, поле $\mathbf{F}$ является абелевой группой. Подмножество $\mathbf{F} \setminus \{0\}$ является мультипликативной абелевой группой.

Пусть на множестве $\mathbf{R}_m$ целых чисел сложение и умножение определены по модулю m. Множество $\mathbf{R}_m$ называется кольцом классов вычетов по модулю m. Оно является коммутативным кольцом, а также кольцом главных идеалов.

Если p – простое число, то кольцо чисел по модулю p является полем. Это поле далее будем обозначать $\mathbf{GF}(p)$. Поле не может иметь менее двух элементов, т.к. в нем должны быть единичные элементы как относительно сложения, так и умножения. Поле, включающее только 0 и 1, далее будем обозна-

чать $\mathbf{GF}(2)$, а вместо специального знака $\oplus$, обозначающего операцию сложения по модулю два, для простоты будем использовать обычный знак сложения.

Многочленом относительно x *над полем* $\mathbf{F}$ называется выражение

$$f(x)=\alpha_0+\alpha_1 x+\ldots+\alpha_n x^n,$$

где α_i, $i=\overline{0,n}$ – принадлежат полю $\mathbf{F}$.

Степенью $\deg(f)$ многочлена $f(x)$ называется наибольшее число i, такое, что $\alpha_i\neq 0$. Многочлен нулевой степени называется константой. Если $\deg(f)=n$, то α_n – старший коэффициент. Многочлен, у которого $\alpha_n=1$, называется *нормированным*.

Множество всех многочленов над полем $\mathbf{F}$ с определенными в поле операциями сложения и умножения составляют кольцо $\mathbf{F}(x)$.

Для любых многочленов $a(x)$ и $b(x)$ из кольца $\mathbf{F}(x)$ имеем и притом единственным образом

$$a(x)=b(x)q(x)+r(x),\ \deg(r(x))<\deg(q(x)). \qquad (12.1)$$

Если $r(x)=0$, то $b(x)$ является *делителем* $a(x)$, а сам $a(x)$ является многочленом, *кратным* $b(x)$. Если единственными делителями $a(x)$ являются α или $\alpha\cdot a(x)$, где α – некоторый элемент из $\mathbf{F}$, то $a(x)$ называется *неприводимым* многочленом над полем $\mathbf{F}$.

Любой многочлен $f(x)\neq Const$ может быть представлен в виде

$$f(x)=\alpha\left[p_1(x)\right]^{l_1}\left[p_2(x)\right]^{l_2}\ldots\left[p_r(x)\right]^{l_r},\quad l_i>0,$$

где $p_i(x)$, $i=\overline{1,r}$ – неприводимые нормированные многочлены, а $\left[p_i(x)\right]^{l_i}$, $i=\overline{1,r}$ – простые делители многочлена $f(x)$.

Любой многочлен $f(x)$ над полем $\mathbf{GF}(p)$, где p – простое число, не делящийся на x, является делителем многочлена $1-x^i$ для некоторого целого i. Наименьшее положительное число $i=T$ называется *показателем*, которому принадлежит многочлен $f(x)$. Если многочлен n-й степени принадлежит показа-

телю T, то число $p^n - 1$ делится на T. Для любого n и любого простого p существует по крайней мере один неприводимый многочлен n-й степени, принадлежащий показателю $p^n - 1$, который называется *многочленом, принадлежащим максимальному показателю*.

Простыми делителями многочлена $x - x^{p^k}$ являются неприводимые многочлены над полем $\mathbf{GF}(p)$, на степени которых делится k. Многочлен $1 - x^k$ делится на многочлен $1 - x^h$ тогда и только тогда, когда k делится на h.

12.2 Понятие и общая схема построения циклического кода

Циклическим называется код, каждая комбинация которого может быть получена путем циклического сдвига комбинации, принадлежащей этому же коду. Если сдвиг осуществляется справа налево, крайний левый символ переносится в конец кодовой комбинации (табл. 12.1).

Описание циклических кодов удобно проводить с помощью многочленов. Для этого вводят фиктивную переменную x, степени которой соответствуют номерам разрядов, начиная с 0. В качестве коэффициентов многочленов берут цифры 0 и 1, т.е. вводятся в рассмотрение многочлены над полем $\mathbf{GF}(2)$. Например, первая строка из примера (табл. 12.1) описывается многочленом

Таблица 12.1

0	0	1	0	1	1
0	1	0	1	1	0
1	0	1	1	0	0
0	1	1	0	0	1

$$0 \cdot x^5 + 0 \cdot x^4 + 1 \cdot x^3 + 0 \cdot x^2 + 1 \cdot x^1 + 1 \cdot x^0 = x^3 + x + 1.$$

Многочлен для каждой следующей строки образуется из предыдущего путем умножения на x. При этом, если крайний левый символ отличается от нуля, для реализации операции переноса единицы в конец комбинации из результата необходимо вычесть (сложить по модулю 2) многочлен $x^n + 1$.

Все комбинации циклического кода могут быть построены на кольце многочленов путем задания на множестве n-разрядных кодовых комбинаций двух операций – сложения и умножения. Операция сложения многочленов в данном

случае реализуется как сложение соответствующих коэффициентов по модулю 2.

Операция умножения реализуется в следующей последовательности. Многочлены перемножаются как обычно с последующим приведением коэффициентов по модулю 2. Если в результате умножения получается многочлен степени n и выше, то осуществляется его деление на заданный многочлен степени n, а результатом умножения считают остаток от деления. Ясно, что старшая степень этого остатка не будет превышать величины $n-1$, а полученный остаток будет соответствовать некоторой n-разрядной кодовой комбинации, т.е. обеспечивается замкнутость.

Для реализации циклического сдвига с использованием описанной операции умножения необходимо после умножения на x выполнить деление на двучлен x^n+1. Эта операция называется *взятием остатка* или *приведением по модулю* x^n+1, а сам остаток называют вычетом:

$$\begin{array}{rl} (x^{n-1}+x^{n-2}+\ldots+x+1)\cdot x \;=\; x^n+x^{n-1}+\ldots+x^2+x & \left|\underline{x^n+1}\right. \\ \oplus\ \underline{x^n+1} & \left|1\right. \\ 0+x^{n-1}+\ldots+x^2+x+1 & \end{array}.$$

Нетрудно заметить, что в данном случае остаток (вычет) формируется путем сложения по модулю 2 двучлена x^n+1 с результатом умножения на x.

12.3 Построение циклического кода на кольце многочленов

Выделим в кольце подмножество всех многочленов, кратных некоторому многочлену $g(x)$. Ясно, что это подмножество будет идеалом, а многочлен $g(x)$ – *порождающим* или *образующим* многочленом идеала. Если $g(x)=0$, то весь идеал состоит из одного этого многочлена. Если $g(x)=1$, то в идеал войдут все многочлены кольца.

В кольце 2^n всех возможных многочленов степени n-1 над полем $\mathbf{GF}(2)$ неприводимый многочлен $g(x)$ степени $m=n-k$ порождает 2^k элементов идеала. Следовательно, можно определить циклический двоичный код как иде-

ал, каждому многочлену которого ставится в соответствие n-разрядная разрешенная кодовая комбинация. Установим, каким требованиям при этом должен удовлетворять образующий многочлен идеала – $g(x)$.

По определению идеала все его многочлены $g_1(x)$, $g_2(x)$,... должны делиться без остатка на $g(x)$. На множестве многочленов идеала выделим подмножество так называемых *базовых полиномов* $g_1(x)$, $g_2(x)$,...,$g_k(x)$, суммированием которых во всех возможных комбинациях могут быть построены все многочлены идеала.

В соответствии с описанной выше схемой циклического сдвига, базовые полиномы могут быть образованы последовательным умножением на x с последующим приведением по модулю x^n+1:

$$\begin{aligned} &g_1(x)=g(x),\\ &g_2(x)=g_1(x)x+c(x^n+1),\\ &\ldots \quad \ldots \quad \ldots,\\ &g_k(x)=g_{k-1}(x)x+c(x^n+1), \end{aligned} \tag{12.2}$$

где $c=1$, если степень $g_i(x)x$ превышает $n-1$ и $c=0$, если степень $g_i(x)x$ не превышает $n-1$.

Для того чтобы все многочлены, соответствующие комбинациям циклического кода, делились без остатка на $g(x)$, достаточно, чтобы на него делились без остатка указанные выше базовые полиномы. Из (12.2) следует, что для этого должен делиться без остатка на $g(x)$ многочлен x^n+1. Таким образом, чтобы порождающий идеал многочлен $g(x)$ являлся образующим элементом циклического кода, он должен быть делителем многочлена x^n+1.

Если $g(x)$ удовлетворяет этому требованию, то кольцо многочленов можно разложить на классы вычетов по идеалу. Для наглядности схема разложения представлена в табл. 12.2. Первой строкой в этой таблице является сам идеал вместе с нулевым многочленом. В качестве образующих элементов классов бе-

рутся (соответствующие векторам ошибок) многочлены $r(x)$, не принадлежащие идеалу, а классы вычетов по идеалу образуются путем сложения элементов идеала с образующими многочленами.

Таблица 12.2

0	$g(x)$	$xg(x)$	$(x+1)g(x)$	…	$f(x)\cdot g(x)$
$r_1(x)$	$g(x)+r_1(x)$	$xg(x)+r_1(x)$	$(x+1)g(x)+r_1(x)$	…	$f(x)\cdot g(x)+r_1(x)$
$r_2(x)$	$g(x)+r_2(x)$	$xg(x)+r_2(x)$	$(x+1)g(x)+r_2(x)$	…	$f(x)\cdot g(x)+r_2(x)$
…				…	
$r_z(x)$	$g(x)+r_z(x)$	$xg(x)+r_z(x)$	$(x+1)g(x)+r_z(x)$	…	$f(x)\cdot g(x)+r_z(x)$

Если реализована указанная схема образования классов вычетов, а многочлен $g(x)$ степени $m=n-k$ является делителем двучлена x^n+1, то каждый элемент кольца либо делится на $g(x)$ без остатка (тогда он элемент идеала), либо появляется остаток от деления $r(x)$ – это многочлен степени не выше $m-1$. Элементы кольца, дающие один и тот же остаток $r(x)$, относят к одному классу вычетов.

Корректирующая способность кода тем выше, чем больше классов вычетов, т.е. остатков $r(x)$. Наибольшее число остатков 2^m-1 дает неприводимый многочлен. В качестве примера в табл. 12.3 приведены неприводимые многочлены до третьей степени включительно. Таблицы, включающие большое число неприводимых многочленов, можно найти, например, в [2], [3].

Таблица 12.3

M	Код	$g(x)$	Обозначение
1	11	$x+1$	$P(x^1)$
2	111	x^2+x+1	$P(x^2)$
3	1011	x^3+x+1	$P_1(x^3)$
3	1101	x^3+x^2+1	$P_2(x^3)$

12.4 Выбор образующих многочленов для обнаружения и исправления одиночных ошибок

Обнаружение одиночных ошибок. В данном случае искаженная кодовая комбинация может быть представлена в виде $q(x)=a(x)+\xi_i(x)$, где

$\xi_i(x) = x^i$, $i = \overline{0,n-1}$ – соответствуют множеству одиночных ошибок. Если $\xi_i(x) = 0$, то $q(x)$ должен делиться без остатка на $g(x)$. Если $\xi_i(x) \neq 0$, то появляется остаток – признак ошибки, это означает, что x^i не должен делится на $g(x)$.

Среди неприводимых многочленов, входящих в разложение $x^n + 1$, многочленом наименьшей степени, удовлетворяющим этому требованию, является $x + 1$. Остатком от деления любого многочлена на $x + 1$ является многочлен нулевой степени, принимающий два значения: либо 0, либо 1. Поэтому все кольцо в данном случае состоит из идеала и одного класса вычетов, соответствующего единственному остатку, равному 1.

Таким образом, для обнаружения одиночных и любого нечетного количества ошибок необходим один проверочный разряд. Проверочный символ в этом разряде выбирается так, чтобы число единиц в любой разрешенной комбинации было четным.

Исправление одиночных ошибок. Каждой одиночной ошибке в одном из n разрядов должен соответствовать свой класс вычетов и свой опознаватель – остаток от деления на образующий многочлен $g(x)$. Как указывалось выше, наибольшее число остатков дает неприводимый многочлен. Если $m = n - k$ степень этого многочлена, число ненулевых остатков будет $2^{n-k} - 1$. Таким образом, для исправления всех n одиночных ошибок необходимо, чтобы выполнялось $2^{n-k} - 1 \geq C_n^1 = n$. Откуда степень образующего многочлена

$$m = n - k \geq \log_2(n+1).$$

Выше было показано, что образующий многочлен должен быть делителем $x^n + 1$. С другой стороны, известно, что любой двучлен вида $x^{2^m - 1} + 1 = x^n + 1$ всегда может быть представлен в виде произведения всех неприводимых многочленов, степени которых являются делителями числа m от 1 до m включительно. Следовательно, для любого n существует хотя бы один неприводимый

многочлен степени m, входящий сомножителем в разложение двучлена x^n+1. Этот многочлен и может быть принят в качестве образующего.

Например, для рассматривавшегося в разделах 11.4, 11.5 случая построения кода (7,4), т.е. для n=7 и m=3, двучлен

$$x^7+1=x^{2^3-1}+1$$

можно записать в виде произведения следующих неприводимых многочленов (см. табл. 12.3):

$$(x+1)\cdot(x^3+x+1)\cdot(x^3+x^2+1),$$

степени которых являются делителями числа 3. Любой из сомножителей третьей степени в данном случае может быть принят в качестве образующего многочлена.

12.5 Методы формирования комбинаций и декодирования циклического кода

Способ 1. Для построения n-разрядной разрешенной комбинации многочлен $a(x)$, соответствующий кодируемой последовательности информационных символов, умножается на образующий многочлен:

$$q(x)=a(x)g(x). \qquad (12.3)$$

При декодировании (возможно отличающийся от $q(x)$) многочлен $\tilde{q}(x)$, соответствующий принятой комбинации, делят на $g(x)$. Ясно, что в случае отсутствия ошибок сразу получится исходный многочлен $a(x)$. Если в принятой комбинации содержится ошибка, при делении образуется остаток $r(x)$, т.е.

$$\tilde{q}(x)/g(x)=f(x)+r(x)/g(x).$$

По остатку определяется класс вычетов и производится исправление ошибки.

Недостаток данного способа кодирования заключается в том, что после обнаружения и исправления ошибки необходимо снова делить на $g(x)$ для того, чтобы выделить информационные символы.

Способ 2. Многочлен, соответствующий исходной информационной посылке $a(x)$, умножается на x^m. Образовавшиеся после умножения свободные младшие разряды заполняются остатком от деления данного выражения на образующий многочлен:

$$q(x) = a(x) \cdot x^m + r(x). \qquad (12.4)$$

Многочлен $q(x)$ обязан делиться на $g(x)$ без остатка. Покажем это.

При делении $a(x)x^m$ на $g(x)$ в общем случае имеем

$$a(x) \cdot x^m / g(x) = c(x) + r(x)/g(x),$$

где $c(x)$ – целый полином. Это равенство (с учетом того, что операции вычитания и сложения по модулю два совпадают) можно переписать в виде

$$a(x) \cdot x^m / g(x) + r(x)/g(x) = c(x),$$

или

$$q(x) = a(x) \cdot x^m + r(x) = c(x)g(x).$$

Из (12.4) видно, что в данном случае информационные символы всегда остаются на первых k позициях. Такой код называют *систематическим*. При таком способе кодирования после исправления ошибок сразу становится известной исходная кодовая последовательность, занимающая первые k позиций.

Существует также способ формирования циклического кода, реализуемый в виде рекуррентных соотношений с использованием так называемого генераторного многочлена. Этот способ мы рассмотрим в разделе 14.6, посвященном линейным последовательным машинам.

Лекция 13

Матричные представления в теории кодирования

13.1 Групповой код как подпространство линейного пространства

Линейным (*векторным*) *пространством* V над полем $\mathbf{F}$ называют множество элементов (векторов), для которого выполняются аксиомы:

1) множество V является коммутативной группой по сложению;
2) для любого $\mathbf{v} \in V$ и скаляра c определено $c\mathbf{v} \in V$ (замкнутость);
3) для любых $\mathbf{v}$, $\mathbf{x}$ из V и α, β из $\mathbf{F}$ $(\alpha+\beta)\mathbf{v} = \alpha\mathbf{v} + \beta\mathbf{v}$, $\alpha(\mathbf{v}+\mathbf{x}) = \alpha\mathbf{v} + \alpha\mathbf{x}$ (дистрибутивность);
4) если $\mathbf{v}$ – вектор из V, а α, β – скаляры, то $(\alpha\beta)\mathbf{v} = \alpha(\beta\mathbf{v})$ (ассоциативность к умножению на скаляр) и $1 \cdot \mathbf{v} = \mathbf{v}$.

Множество n-разрядных двоичных комбинаций помехоустойчивого кода можно рассматривать как векторное линейное пространство над полем $\mathbf{GF}(2)$ с операцией сложения по модулю 2, а кодовые комбинации – как его векторы. Действительно, если определить операцию умножения последовательности из n элементов поля $\mathbf{GF}(2)$ (кодовой комбинации) на элемент a_i поля $\mathbf{GF}(2)$ аналогично правилу умножения вектора на скаляр:

$$a_i(a_1, a_2, ..., a_n) = (a_i a_1, a_i a_2, ..., a_i a_n),$$

то все указанные выше аксиомы выполняются.

Подмножество элементов векторного пространства, удовлетворяющее аксиомам векторного пространства, называют *подпространством*. По-видимому, множество векторов, соответствующих разрешенным комбинациям, образует подпространство векторного пространства всех n-разрядных кодовых комбинаций над полем $\mathbf{GF}(2)$.

Заметим, что такое подпространство комбинаций над полем $\mathbf{GF}(2)$, вообще говоря, образует любая совокупность двоичных кодовых комбинаций, яв-

ляющаяся подгруппой группы всех n-разрядных двоичных кодовых комбинаций.

13.2 Понятие образующей матрицы, построение разрешенных кодовых комбинаций с использованием образующей матрицы

Расположим $2^k - 1$ разрешенных n-разрядных кодовых комбинаций друг под другом в виде строк матрицы $\mathbf{M}$ размерности $\left(2^k - 1\right) \times n$. Поскольку $n - k$ проверочных символов каждой строки этой матрицы формируются в виде линейных комбинаций информационных символов, только k столбцов этой матрицы будут линейно независимыми, т.е. $\text{rank}\mathbf{M} = k$. Это означает, что среди строк (кодовых комбинаций) матрицы $\mathbf{M}$ только k линейно независимых.

Образующей (*порождающей*) называется матрица, состоящая из любых k линейно независимых векторов (строк). Совокупность этих векторов образует базис пространства. Все остальные разрешенные комбинации могут быть представлены в виде линейной комбинации базисных векторов. Если образующая матрица содержит k строк по n элементов поля $\mathbf{GF}(2)$, соответствующий код называют (n,k)-кодом.

Если известна образующая матица $\mathbf{M}_{n,k}$, любая n-разрядная разрешенная комбинация ($n \times 1$-вектор $\mathbf{A}_n$) может быть получена путем умножения k-разрядной комбинации, составленной из информационных символов ($k \times 1$-вектора $\mathbf{A}_k$) на образующую матрицу:

$$\mathbf{A}_n = \mathbf{A}_k \cdot \mathbf{M}_{n,k}. \qquad (13.1)$$

Перестановка строк (столбцов) образующей матрицы приводит к эквивалентному коду с той же корректирующей способностью.

Если формируемый код должен быть систематическим, образующая матрица представляется в виде двух блоков: единичной $k \times k$-матрицы $\mathbf{E}_k$ и так называемой *матрицы-дополнения* $\mathbf{P}_{k,n-k}$ размерности $k \times (n-k)$:

$$\mathbf{M}_{n,k} = \left[\mathbf{E}_k \vdots \mathbf{P}_{k,n-k}\right] = \left[\begin{array}{ccc|ccc} 1 & \ldots & 0 & p_{1,k+1} & \ldots & p_{1,n} \\ \vdots & \ddots & \vdots & \vdots & p_{i,j} & \vdots \\ 0 & \ldots & 1 & p_{k,k+1} & \ldots & p_{k,n} \end{array}\right], \tag{13.2}$$

где $p_{i,j}$ – проверочные символы.

При умножении в соответствии с (13.1) вектор-строки $\mathbf{A}_k = \left[a_1,\ldots,a_k\right]$ на матрицу $\mathbf{M}_{n,k}$ (13.2) получаем

$$\mathbf{A}_n = \mathbf{A}_k \mathbf{M}_{n,k} = \left[\mathbf{A}_k \mathbf{E}_k \vdots \mathbf{A}_k \mathbf{P}_{k,n-k}\right] = \left[\mathbf{A}_k \vdots \mathbf{A}_{n-k}\right]. \tag{13.3}$$

В данном случае первые k символов вектор-строки $\mathbf{A}_n$ всегда информационные, а последние $n-k$ – так называемые проверочные символы являются их линейными комбинациями:

$$a_j = \sum_{i=1}^{k} a_i p_{i,j}, \quad j = \overline{k+1,n}. \tag{13.4}$$

Заметим, что формирование кодовой комбинации по правилу (13.3) сводится к поразрядному сложению строк образующей матрицы с номерами, соответствующими номерам ненулевых информационных символов вектора $\mathbf{A}_k$.

13.3 Построение матрицы-дополнения

Из (13.2)–(13.4) видно, что матрица-дополнение содержит всю информацию о схеме построения кода. Например, $p_{i,j} = 1$ говорит о том, что в образовании j-го проверочного разряда $\left(j = \overline{k+1,n}\right)$ участвовал i-й $\left(i = \overline{1,k}\right)$ информационный разряд. Следовательно, по матрице-дополнению всегда можно записать уравнения кодирования в виде (11.11) или (13.4).

Наоборот, если заданы уравнения кодирования, то значение любого символа $p_{i,j}$ матрицы-дополнения может быть определено путем применения соответствующего уравнения для формирования j-го проверочного разряда к i-й строке единичной матрицы.

Существует формальный способ построения матрицы дополнения, основанный на следующем требовании. Вектор-строка, получающаяся в результате

суммирования любых l, $(1 \le l \le k)$ строк матрицы дополнения, должна содержать не менее $d_{\min} - l$ отличных от нуля символов, где $d_{\min}$ – минимальное кодовое расстояние. В соответствии с указанным требованием, матрица-дополнение может строиться с соблюдением следующих правил:

1) количество единиц в строке должно быть не менее $d_{\min} - 1$;
2) сумма по модулю два двух любых строк должна содержать не менее $d_{\min} - 2$ единиц.

При соблюдении указанных требований комбинация, полученная суммированием любых двух строк образующей матрицы, будет содержать не менее $d_{\min}$ ненулевых символов.

13.4 Понятие и построение проверочной (контрольной) матрицы

Код представляет собой n-мерное векторное пространство. Образующая матрица $\mathbf{M}_{n,k}$ определяет k-мерное подпространство. Следовательно, существует ортогональное подпространство размерности $n-k$. Пусть

$$\mathbf{H} = \begin{bmatrix} h_{1,1} & \dots & h_{1,n} \\ \vdots & \ddots & \vdots \\ h_{n-k,1} & \dots & h_{n-k,n} \end{bmatrix} \tag{13.5}$$

– матрица, векторы-строки которой задают это подпространство.

В силу ортогональности указанных подпространств $\mathbf{M}_{n,k}\mathbf{H}^T = 0$. Следовательно, для разрешенного кодового слова $\mathbf{A}_n$ будем иметь:

$$\mathbf{A}_n \mathbf{H}^T = \mathbf{A}_k \mathbf{M}_{n,k} \mathbf{H}^T = 0. \tag{13.6}$$

Матрица $\mathbf{H}$, для которой имеет место равенство (13.6), всегда существует и называется *проверочной* (*контрольной*) матрицей, а указанное выражение используется для определения ошибок в кодовой комбинации. Подчеркнем, что, в соответствии с (13.6), векторы, соответствующие разрешенным кодовым комбинациям, принадлежат нуль-пространству матрицы $\mathbf{H}^T$.

Для систематического кода проверочная матрица имеет вид

$$\mathbf{H} = \left[\mathbf{P}^T_{k,n-k} \vdots \mathbf{E}_{n-k} \right]. \tag{13.7}$$

Нетрудно заметить, что в данном случае

$$\mathbf{A}_n \mathbf{H}^T = \left[\mathbf{A}_k \vdots \mathbf{A}_{n-k} \right] \begin{bmatrix} \mathbf{P}_{k,n-k} \\ --- \\ \mathbf{E}_{n-k} \end{bmatrix} = \mathbf{S} = \left[0,0,...,0 \right],$$

где $\mathbf{S}$ – вектор, компоненты которого определяются как

$$\sum_{i=1}^{k} a_i p_{i,j} + a_j = 0, \quad j = \overline{k+1,n}.$$

Если кодовый вектор $\tilde{\mathbf{A}}_{n,i}$ содержит ошибки: $\tilde{\mathbf{A}}_n = \mathbf{A}_n + \xi_n$, $(\xi_n \neq 0)$,

$$\left[\mathbf{A}_n + \xi_n \right] \mathbf{H}^T = \left[\mathbf{A}_n \mathbf{H}^T \right] + \left[\xi_n \mathbf{H}^T \right] = \left[\xi_n \mathbf{H}^T \right].$$

При этом компоненты S_j:

$$S_j = \sum_{i=1}^{k} \xi_i p_{i,j} + \xi_j, \quad j = \overline{k+1,n}$$

вектора $\mathbf{S}$ могут отличаться от нуля. Они зависят только от вектора ошибок, а составленный из них вектор $\mathbf{S}$ является опознавателем ошибки (*синдромом*).

13.5 Границы для числа разрешенных комбинаций

Опираясь на понятие проверочной матрицы, можно построить так называемую *границу Варшамова-Гилберта* для числа проверочных символов кода длины n с заданным минимальным кодовым расстоянием d.

В соответствии с (13.6) код является разрешенным тогда и только тогда, когда

$$\sum_{i=1}^{n} a_i \mathbf{h}_i = 0, \tag{13.8}$$

где $\mathbf{h}_i$ – i-й столбец $m \times n$ матрицы $\mathbf{H}$. Ясно, что число столбцов матрицы $\mathbf{H}$, которые входят в (13.8) с ненулевыми коэффициентами, равно весу кодового слова, а вектор, соответствующий этому кодовому слову, принадлежит нуль-пространству матрицы $\mathbf{H}^T$.

Отсюда, в частности, следует, что любой код, принадлежащий нуль-пространству матрицы $\mathbf{H}$, имеет минимальный вес, а следовательно, и минимальное кодовое расстояние, равное, самое меньшее, d, тогда и только тогда, когда любые $d-1$ или меньше столбцов матрицы $\mathbf{H}$ линейно-независимы.

Матрица $\mathbf{H}$, обладающая указанным свойством, может быть построена путем последовательного добавления столбцов по следующему правилу. В качестве первого столбца берется любая ненулевая последовательность длины $m = n - k$. Вторым столбцом может быть любая некратная первой ненулевая последовательность длины m. Третий столбец – любая последовательность длины m, не являющаяся линейной комбинацией первых двух. Вообще, в качестве i-го столбца берется любая последовательность длины m, не являющаяся линейной комбинацией никаких $d-2$ или меньше предыдущих столбцов. При этом никакая линейная комбинация из $d-1$ или меньше столбцов матрицы не обращается в нуль.

Число всех возможных двоичных линейных комбинаций из $d-2$ или меньше столбцов, выбранных из общего числа n столбцов, в наихудшем случае (когда все они различны) равно

$$C_n^1 + C_n^2 + \ldots + C_n^{d-2} = \sum_{i=1}^{d-2} C_n^i . \qquad (13.9)$$

Очередной столбец может быть присоединен к матрице в том случае, если число комбинаций определяемых суммой (13.9) меньше, чем общее число отличных от нуля последовательностей длины m:

$$\sum_{i=1}^{d-2} C_n^i < 2^m - 1. \qquad (13.10)$$

Таким образом, возможно построение кода длины n с минимальным расстоянием d и m проверочными символами, где m – наименьшее целое число, удовлетворяющее неравенству (13.10).

Соответствующая неравенству (13.10) граница (Варшамова-Гилберта) получена в расчете на наихудший случай. Она указывает лишь на принципиальную возможность реализации n-разрядного кода с заданной корректирующей

способностью. Представляет интерес установить также верхнюю границу для *оптимального* кода, обеспечивающего заданную корректирующую способность при минимальной избыточности. Определим наибольшее число разрешенных кодовых комбинаций для n-значного помехоустойчивого кода, обладающего способностью исправлять ошибки до кратности s включительно.

Подмножество запрещенных комбинаций для каждой разрешенной содержит $\sum C_n^i, \quad i=\overline{1,s}$ элементов. Вместе с разрешенной общее число комбинаций в подмножестве составляет $\sum C_n^i, \quad i=\overline{0,s}$. Следовательно, при разложении группы на непересекающиеся классы число разрешенных комбинаций не может превышать величину, определяемую неравенством

$$2^k \le 2^n \Big/ \sum_{i=0}^{s} C_n^i . \qquad (13.11)$$

Приведенное соотношение (13.11) называют оценкой *Хэмминга*. Если в этом выражении имеет место равенство, код называют *плотно упакованным*.

13.6 Матричное представление циклических кодов

Циклический код является групповым кодом, поэтому он может строиться с использованием матричных представлений так, как описано выше. Однако в данном случае появляются также некоторые дополнительные возможности, связанные со свойством цикличности. Рассмотрим способы построения образующей матрицы циклического кода.

Способ 1. Пусть образующий многочлен задан в виде

$$g(x) = g_m x^m + \ldots + g_1 x + g_0 .$$

Тогда образующая матрица может быть построена путем умножения $g(x)$ на одночлен $x^{k-1}, k = n - m$ и последующим циклическим сдвигом так, что каждая i-я строка образующей матрицы составляется из коэффициентов многочлена $g(x) \cdot x^{k-i}$ $(i=\overline{1,k})$:

$$\mathbf{M}_{n,k}=\begin{bmatrix} g_m & g_{m-1} & \cdots & g_0 & 0 & \cdots & 0 \\ 0 & g_m & g_{m-1} & \cdots & g_0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & g_m & g_{m-1} & \cdots & g_0 \end{bmatrix}. \qquad (13.12)$$

Способ 2. Рассматриваются многочлены $Q_i(x)$, соответствующие коду, содержащему только один ненулевой разряд: $Q_i(x)=x^{n-i}, i=\overline{1,k}$. Для них вычисляются остатки $r_i(x)=Q_i(x)/g(x)$. Каждая i-я строка образующей матрицы формируется путем сложения по модулю два указанных многочленов и соответствующих им остатков. При этом образующая матрица (в данном случае систематического кода) представляется двумя подматрицами:

$$\mathbf{M}_{n,k}=\left[\mathbf{E}_k \vdots \mathbf{P}_{k,n-k}\right],$$

где $\mathbf{E}_k$ – единичная $k\times k$ - матрица, а строками матрицы дополнения $\mathbf{P}_{k,n-k}$ являются остатки $r_i(x), i=\overline{1,k}$.

13.7 Построение проверочной матрицы циклического кода

Проверочная матрица в данном случае может строиться так же, как в случае обычного группового кода, например с использованием проверочных равенств и/или матрицы-дополнения. Однако для циклического кода существует еще один способ построения проверочной матрицы, заключающийся в делении многочлена x^n+1 на многочлен $g^{-1}(x)$, являющийся дополнением к образующему. Многочлен дополнения соответствует кодовой комбинации, которая получается из комбинации, соответствующей образующему многочлену путем перестановки символов в обратном порядке.

Предположим, что в результате деления двучлена x^n+1 на многочлен дополнения получен некоторый многочлен:

$$\frac{x^n+1}{g^{-1}(x)}=b_k x^k+\ldots b_1 x+b_0. \qquad (13.13)$$

Из коэффициентов этого многочлена составляется первая строка проверочной матрицы, а остальные строки образуются циклическим сдвигом:

$$\mathbf{H}=\begin{bmatrix} b_k & \cdots & b_1 & b_0 & 0 & \cdots & 0 \\ 0 & b_k & \cdots & b_1 & b_0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & b_k & \cdots & b_1 & b_0 \end{bmatrix}. \qquad (13.14)$$

В качестве примера построим проверочную матрицу для кода (7, 4), порождаемого образующим многочленом $g(x)=x^3+x+1$. Соответствующий многочлен дополнения $g^{-1}(x)=x^3+x^2+1$. В результате деления на него двучлена x^7+1 получаем многочлен $x^4+x^3+x^2+1$. Соответствующая этому многочлену проверочная матрица имеет вид

$$\mathbf{H}=\begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{bmatrix}$$

Нетрудно убедиться, что любая разрешенная комбинация $\mathbf{A}_n$, полученная путем умножения некоторого заданного информационного многочлена $a(x)$ на указанный выше образующий многочлен: $g(x)$ – в результате умножения на транспонированную проверочную матрицу: $\mathbf{A}_n\mathbf{H}^T$ дает синдром, состоящий из одних нулей.

Лекция 14

Кодирование линейными последовательными машинами

14.1 Понятие линейной последовательной машины

Линейная последовательная машина (ЛПМ) – это система с конечным числом входов $u_i, i=\overline{1,l}$ и выходов $y_j, j=\overline{1,m}$, сигналы на которых наблюдаются в дискретные моменты времени, и выполняющая следующие элементарные функции (рис. 14.1) [2]:

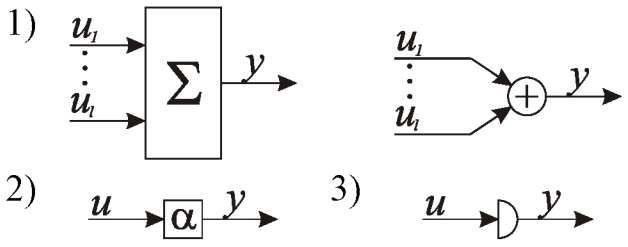

Рис. 4.1. Элементы ЛПМ

1) сложение: $y=\sum_{i=1}^{l} u_i$;

2) умножение на постоянную:

$y=\alpha u$;

3) задержка: $y(t)=u(t-1)$.

Здесь и далее под аргументом сигнала подразумевается номер момента времени.

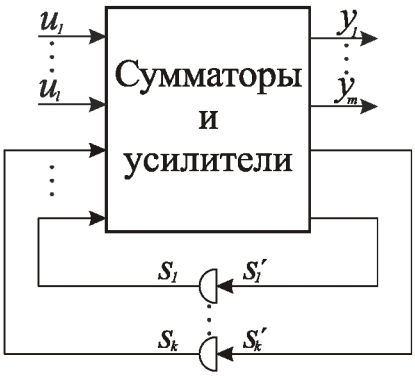

Рис. 14.2. Общая схема ЛПМ

Общая схема ЛПМ может быть представлена в виде, показанном на рис. 14.2. Число задержек определяет *размерность ЛПМ*. Запрещаются петли, не содержащие ни одной задержки, т.к. это приводит к неопределенности в описании состояний $s_i(t)$, $i=\overline{1,k}$. Для ЛПМ размерности k имеют место равенства $s_i'(t-1)=s_i(t)$, $i=\overline{1,k}$ или в векторном виде

$$\mathbf{s}'(t-1)=\mathbf{s}(t),$$

где $\mathbf{s}(t)$ – $k\times 1$-вектор состояний. Множество векторов $\mathbf{s}(t)$ образует пространство состояний ЛПМ.

14.2 Матричное описание ЛПМ

В соответствии с общей схемой (рис. 14.2), работу ЛПМ можно описать следующими соотношениями

$$s_i(t+1)=\sum_{j=1}^{k}a_{ij}s_j(t)+\sum_{j=1}^{l}b_{ij}u_j(t),\ i=\overline{1,k}, \tag{14.1}$$

$$y_i(t)=\sum_{j=1}^{k}c_{ij}s_j(t)+\sum_{j=1}^{l}d_{ij}u_j(t),\ i=\overline{1,m}. \tag{14.2}$$

Равенства (14.1), (14.2) можно представить компактно в векторно-матричной форме:

$$\mathbf{s}(t+1)=\mathbf{A}\mathbf{s}(t)+\mathbf{B}\mathbf{u}(t), \tag{14.3}$$

$$\mathbf{y}(t)=\mathbf{C}\mathbf{s}(t)+\mathbf{D}\mathbf{u}(t), \tag{14.4}$$

где $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ – $k\times k$, $k\times l$, $m\times k$, $m\times l$-матрицы, а $\mathbf{u}$, $\mathbf{y}$ – $l\times 1$, $m\times 1$-векторы соответственно.

По соотношениям (14.3), (14.4) нетрудно выписать реакцию системы на любом шаге. В частности, при отсутствии входного сигнала ($\mathbf{u}(t)\equiv 0$), выходной сигнал на шаге t связан с начальным состоянием ЛПМ соотношением вида

$$\mathbf{y}(t)=\mathbf{C}\mathbf{s}(t)=\mathbf{C}\mathbf{A}\mathbf{s}(t-1)=....=\mathbf{C}\mathbf{A}^t\mathbf{s}(0). \tag{14.5}$$

14.3 Каноническая и естественная нормальная форма ЛПМ

Аннулирующим многочленом для матрицы $\mathbf{A}$ является многочлен $\varphi(x)$, такой, что

$$\varphi(\mathbf{A})=0.$$

Аннулирующий многочлен минимальной степени со старшим коэффициентом, равным единице, называется *минимальным*.

Многочлен

$$\varphi(x)=\det(\mathbf{A}-\mathbf{E}x)$$

называется *характеристическим*. По теореме Гамильтона-Кэли всякая матрица удовлетворяет своему характеристическому многочлену, т.е.

$$\varphi(\mathbf{A})=0.$$

Следовательно, характеристический полином всегда является аннулирующим, но не обязательно минимальным.

Матрица

$$\mathbf{A}_{\varphi(x)} = \begin{bmatrix} 0 & \vdots & & \mathbf{E} \\ \cdots & \vdots & \cdots & \cdots \cdots \\ -\alpha_0 & \vdots & -\alpha_1 & \cdots -\alpha_{k-1} \end{bmatrix} \tag{14.6}$$

называется *канонической* (*сопровождающей*) матрицей для многочлена

$$\varphi(x) = x^k + \alpha_{k-1}x^{k-1} + \ldots + \alpha_1 x + \alpha_0 .$$

Многочлен $\varphi(x)$ может быть разложен на элементарные множители:

$$\varphi(x) = \varphi_1(x)\varphi_2(x)\ldots\varphi_r(x) = \left[p_1(x)\right]^{l_1}\left[p_2(x)\right]^{l_2}\ldots\left[p_r(x)\right]^{l_r}.$$

Многочлены $\varphi_i(x)$, $i = \overline{1,r}$ называют элементарными делителями матрицы $\mathbf{A}_{\varphi(x)}$. С использованием указанного разложения на элементарные делители может быть построена *естественная нормальная форма* матрицы:

$$\mathbf{A}^*_{\varphi(x)} = \begin{bmatrix} \mathbf{A}_{\varphi_1(x)} & 0 & \cdots & 0 \\ 0 & \mathbf{A}_{\varphi_2(x)} & \cdots & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & \cdots & \mathbf{A}_{\varphi_r(x)} \end{bmatrix}, \tag{14.7}$$

где $\mathbf{A}_{\varphi_i(x)}$, $i = \overline{1,r}$ – матрицы вида (14.6).

14.4 Подобные и минимальные ЛПМ

Преобразование $\hat{\mathbf{A}} = \mathbf{PAP}^{-1}$, где $\mathbf{P}$ – невырожденная матрица, называется *преобразованием подобия*. Преобразование подобия не изменяет собственные значения матрицы, следовательно, подобные матрицы имеют одинаковые элементарные делители. В частности, если $\mathbf{A}_{\varphi(x)}$ подобна некоторой матрице $\hat{\mathbf{A}}$ с элементарными делителями $\varphi_1(x)$, …, $\varphi_r(x)$, то она также подобна естественной нормальной форме (14.7).

Введя в пространстве состояний преобразование координат $\overline{\mathbf{s}}(t)=\mathbf{P}\mathbf{s}(t)$ и умножив (14.3) слева на $\mathbf{P}$, систему (14.3) (14.4) представим в виде

$$\mathbf{P}\mathbf{s}(t+1)=\mathbf{P}\mathbf{A}\mathbf{P}^{-1}\overline{\mathbf{s}}(t)+\mathbf{P}\mathbf{B}\mathbf{u}(t), \tag{14.8}$$

$$\mathbf{y}(t)=\mathbf{C}\mathbf{P}^{-1}\overline{\mathbf{s}}(t)+\mathbf{D}\mathbf{u}(t), \tag{14.9}$$

Далее, введя обозначения

$$\hat{\mathbf{A}}=\mathbf{P}\mathbf{A}\mathbf{P}^{-1},\ \hat{\mathbf{B}}=\mathbf{P}\mathbf{B},\ \hat{\mathbf{C}}=\mathbf{C}\mathbf{P}^{-1},\ \hat{\mathbf{D}}=\mathbf{D},$$

с учетом того, что, в соответствии с используемым преобразованием координат

$$\mathbf{P}\mathbf{s}(t+1)=\overline{\mathbf{s}}(t+1),$$

уравнения (14.8), (14.9) можно переписать в виде

$$\overline{\mathbf{s}}(t+1)=\hat{\mathbf{A}}\overline{\mathbf{s}}(t)+\hat{\mathbf{B}}\mathbf{u}(t), \tag{14.10}$$

$$\mathbf{y}(t)=\hat{\mathbf{C}}\overline{\mathbf{s}}(t)+\hat{\mathbf{D}}\mathbf{u}(t). \tag{14.11}$$

Системы (14.3), (14.4) и (14.10), (14.11) описывают различные, но совпадающие по входу и выходу ЛПМ. Такие ЛПМ называют *подобными*. Путем преобразований подобия может быть построена ЛПМ, имеющая минимальное число задержек. Такая ЛПМ называется минимальной.

Минимальная ЛПМ может быть определена в результате выполнения следующей последовательности шагов [2].

1. Строится так называемая диагностическая матрица (наблюдаемости)

$$\mathbf{K}=\left[\mathbf{C}\vdots\mathbf{C}\mathbf{A}\cdots\mathbf{C}\mathbf{A}^{k-1}\right]^{T}.$$

2. Из линейно независимых строк диагностической матрицы формируется матрица $\mathbf{T}$ и осуществляется преобразование подобия: $\hat{\mathbf{A}}=\mathbf{T}\mathbf{A}\mathbf{T}^{-1}$, $\hat{\mathbf{B}}=\mathbf{T}\mathbf{B}$, $\hat{\mathbf{C}}=\mathbf{C}\mathbf{T}^{-1}$, $\hat{\mathbf{D}}=\mathbf{D}$.

Результатом преобразования будет минимальная ЛПМ.

Если ЛПМ с матрицей $\mathbf{A}$ имеет подобную ЛПМ с матрицей $\hat{\mathbf{A}}$, то она имеет и естественную нормальную форму $\mathbf{A}^{*}$. Каждая подматрица $\mathbf{A}_{\varphi_i(x)}$ матрицы $\mathbf{A}^{*}$, имеющая вид (14.6), соответствует некоторой канонической ЛПМ.

Каноническая форма является минимальной ЛПМ. Следовательно, в результате преобразования подобия исходная ЛПМ всегда может быть представлена в виде совокупности ЛПМ, каждая из которых соответствует элементарному делителю $\varphi_i(x)$, $i=\overline{1,r}$ в разложении многочлена $\varphi(x)$.

14.5 Понятие простой автономной ЛПМ

Рассмотрим каноническую (минимальную) ЛПМ, имеющую сопровождающую матрицу вида (14.6) при $\mathbf{u}(t)\equiv 0$. ЛПМ с нулевым входным воздействием называются *автономными*. Выходные последовательности на всех выходах ЛПМ, являющихся компонентами вектора $\mathbf{y}$, в этом случае формируются по соотношению (14.5) под действием начальных условий.

Для автономной ЛПМ можно выполнить преобразование подобия для каждого отдельного выхода (компонента вектора $\mathbf{y}$) исходной ЛПМ. При этом из ЛПМ с m выходами будет получено m различных ЛПМ с одинаковыми матрицами $\mathbf{A}$ и различными матрицами $\mathbf{C}$, представляющими собой отдельные строки исходной $m\times n$ – матрицы $\mathbf{C}$.

Каждая из построенных таким образом m схем называется *простой автономной* ЛПМ (простой АЛПМ), а матрица $\mathbf{A}_{\varphi(x)}$ каждой простой АЛПМ имеет вид (14.6) и является сопровождающей для многочлена обратной связи

$$\varphi(x)=x^k+\alpha_{k-1}x^{k-1}+\ldots+\alpha_1 x+\alpha_0 .$$

Матричные соотношения, описывающие соответствующую матрице $\mathbf{A}_{\varphi(x)}$ и указанному многочлену $\varphi(x)$ простую автономную ЛПМ при $\mathbf{C}=[1,0,\ldots,0]$, имеют вид:

$$\begin{bmatrix} s_1(t+1)\\ \vdots \\ s_k(t+1)\end{bmatrix}=\left[\begin{array}{c|ccc} 0 & & \mathbf{E} & \\ \cdots & \vdots & \cdots\ \cdots\ \cdots & \\ -\alpha_0 & -\alpha_1 & \cdots & -\alpha_{k-1}\end{array}\right]\begin{bmatrix} s_1(t)\\ \vdots \\ s_k(t)\end{bmatrix},$$

$$y(t)=[1,0,\ldots,0]\begin{bmatrix} s_1(t)\\ \vdots \\ s_k(t)\end{bmatrix}.$$

Приведенные равенства можно представить в виде схемы, показанной на рис. 14.3.

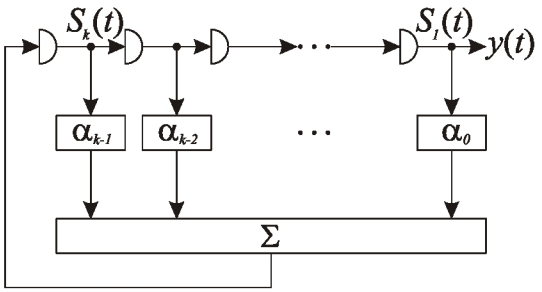

Рис. 14.3. Схема простой АЛПМ

Непосредственно по схеме можно записать соотношение для формирования выходной последовательности простой АЛПМ:

$$y_{t+k} = \alpha_{k-1} y_{t+k-1} + \ldots + \alpha_1 y_{t+1} + \alpha_0 y_t . \qquad (14.12)$$

Нетрудно заметить, что символы выходной последовательности являются линейной комбинацией начального состояния АЛПМ.

14.6 Формирование разрешенных комбинаций циклического кода с помощью АЛПМ

В разделе 12.5 мы рассмотрели два способа формирования комбинаций и декодирования циклических кодов. Рассмотрим еще один способ, который наиболее удобно реализуется с помощью АЛПМ.

Определим многочлен обратной связи $\varphi(x)$ как частное от деления $x^n + 1$ на образующий многочлен. В силу свойств $g(x)$ такой целый полином существует:

$$\varphi(x) = \frac{x^n + 1}{g(x)} = x^k + \alpha_{k-1} x^{k-1} + \ldots + \alpha_1 x + \alpha_0 . \qquad (14.13)$$

Многочлен (14.13) называют также *генераторным* полиномом. Для этого полинома можно построить сопровождающую матрицу $\mathbf{A}_{\varphi(x)}$ вида (14.6) и соответствующую ей АЛПМ.

Если начальное состояние АЛПМ (рис. 14.3) соответствует исходной информационной последовательности, на выходе будет сформирована комбинация, первые k символов которой информационные, а следующие за ними $n-k$ являются линейной комбинацией предыдущих символов:

$$a_{j+k} = \sum_{i=0}^{k-1} \alpha_i a_{j+i}, \quad j = \overline{1, n-k} . \qquad (14.14)$$

где α_i – двоичные коэффициенты многочлена обратной связи АЛПМ (14.13) (генераторного многочлена). Таким образом, с использованием АЛПМ может быть построен систематический циклический код.

14.7 Образующая матрица АЛПМ

Если $\varphi(x)$ – многочлен обратной связи (генераторный многочлен), удовлетворяющий (14.13), то образующий многочлен степени $m = n - k$ определяется как

$$g(x) = \frac{x^n + 1}{\varphi(x)} = g_m x^m + \ldots + g_1 x + g_0 .$$

Тогда, в соответствии с описанным в разделе 13.6 первым способом, может быть построена образующая матрица (13.12) соответствующего циклического кода:

$$\mathbf{M}_{n,k} = \begin{bmatrix} g_m & g_{m-1} & \cdots & g_0 & 0 & \ldots & 0 \\ 0 & g_m & g_{m-1} & \cdots & g_0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & g_m & g_{m-1} & \cdots & g_0 \end{bmatrix}.$$

Разделим образующую матрицу $\mathbf{M}_{n,k}$ на два блока $\mathbf{M} = [\mathbf{M}_1 \vdots \mathbf{M}_2]$ так, чтобы $\mathbf{M}_1$ была квадратной. В силу неприводимости многочлена $g(x)$ ее диагональные элементы отличны от нуля, следовательно, матрица $\mathbf{M}_1$ является невырожденной.

Последовательность информационных символов $\mathbf{A}_k$ можно представить как линейную комбинацию строк матрицы $\mathbf{M}_1$: $\mathbf{A}_k = \mathbf{v}^T \cdot \mathbf{M}_1$, откуда

$$\mathbf{v}^T = \mathbf{A}_k \mathbf{M}_1^{-1} . \qquad (14.15)$$

С другой стороны, избыточный код является той же линейной комбинацией строк матрицы $\mathbf{M}$:

$$\mathbf{A}_n = \mathbf{v}^T \cdot \mathbf{M} = \mathbf{v}^T [\mathbf{M}_1 \vdots \mathbf{M}_2].$$

Подставляя в это равенство $\mathbf{v}^T$ из (14.15), имеем

$$\mathbf{A}_n = \mathbf{A}_k \mathbf{M}_1^{-1} \left[\mathbf{M}_1 \vdots \mathbf{M}_2 \right] = \mathbf{A}_k \left[\mathbf{E} \vdots \mathbf{M}_1^{-1} \mathbf{M}_2 \right].$$

Матрица $\mathbf{M} = \left[\mathbf{E} \vdots \mathbf{M}_1^{-1} \mathbf{M}_2 \right] = \left[\mathbf{E} \vdots \mathbf{P}_{k,n-k} \right]$ является образующей матрицей АЛПМ с многочленом обратной связи $\varphi(x)$. Очевидно, что с ее использованием может быть сформирован систематический код.

Подводя итог, следует заметить, что в настоящей лекции, посвященной изучению линейных последовательных машин, мы привели мало новых сведений, посвященных собственно теории кодирования. Цель этого раздела состояла в том, чтобы показать связь теории кодирования с общей теорией линейных систем. Нам представляется это чрезвычайно важным для понимания общих принципов построения кибернетических систем.

Лекция 15

Обнаружение и различение сигналов

15.1 Постановка задачи обнаружения сигналов при наличии помех

Задача приемного устройства – извлечение из принятого сигнала максимума полезной информации. Для этого последовательно решаются по крайней мере две задачи [9]:

1) обнаружение (принятие решения о наличии сигнала);

2) восстановление (определение параметров сигнала).

Задача определения параметров сигналов рассматривается в следующей лекции. Здесь рассмотрим методы обнаружения сигналов.

Принимаемый сигнал будем представлять вектором $\mathbf{Y}$, компоненты которого являются отсчетами, каждый из которых представляет собой сумму отсчетов компонентов векторов полезного сигнала $\mathbf{X}$ и помехи Ξ. Ясно, что по принятому вектору $\mathbf{Y}$ мы не можем однозначно судить о векторе $\mathbf{X}$. О переданном в действительности сигнале $\mathbf{X}$ можно судить лишь с некоторой вероятностью $p(\mathbf{X}/\mathbf{Y})$.

В общем случае, в соответствии с формулой Байеса, апостериорная плотность вероятности вектора $\mathbf{X}$ определяется как

$$w(\mathbf{X}/\mathbf{Y}) = \frac{w(\mathbf{X})w(\mathbf{Y}/\mathbf{X})}{w(\mathbf{Y})}, \qquad (15.1)$$

где $w(\mathbf{X})$ – априорная плотность вероятности вектора $\mathbf{X}$, $w(\mathbf{Y}/\mathbf{X})$ – условная плотность вероятности вектора $\mathbf{Y}$ при условии, что вектор $\mathbf{X}$ известен, а $w(\mathbf{Y}) = \int_{V_x} w(\mathbf{X})w(\mathbf{Y}/\mathbf{X})d\mathbf{X}$ – безусловная плотность вероятности вектора $\mathbf{Y}$, где V_X – пространство передаваемого сигнала.

Если вектор $\mathbf{X}$ имеет конечное число значений, по аналогии с (15.1)

$$p(\mathbf{X}/\mathbf{Y}) = \frac{p(\mathbf{X})w(\mathbf{Y}/\mathbf{X})}{w(\mathbf{Y})} = \frac{p(\mathbf{X})w(\mathbf{Y}/\mathbf{X})}{\sum_{j=1}^{N} p(x_j)w(\mathbf{Y}/x_j)}, \tag{15.2}$$

где $p(\mathbf{X})$ – априорная, а $p(\mathbf{X}/\mathbf{Y})$ – апостериорная вероятности вектора X.

Таким образом, для определения апостериорной плотности $w(\mathbf{X}/\mathbf{Y})$ и/или вероятности $p(\mathbf{X}/\mathbf{Y})$ необходимо знать априорные плотность $w(\mathbf{X})$ и/или вероятность $p(\mathbf{X})$, а также условную плотность $w(\mathbf{Y}/\mathbf{X})$, которая при известном (измеренном) $\mathbf{Y}$ зависит только от $\mathbf{X}$ и обозначается $L(\mathbf{X})$:

$$w(\mathbf{Y}/\mathbf{X}) = L(\mathbf{X}). \tag{15.3}$$

Функция $L(\mathbf{X})$ называется *функцией правдоподобия*. Эта функция может иметь конечное (в случае дискретного $\mathbf{X}$) или бесконечное (в случае непрерывного $\mathbf{X}$) число значений.

Задача обнаружения сигнала заключается в принятии одной из возможных взаимно исключающих альтернатив (гипотез): гипотезы H_1 о том, что $\mathbf{X} = x_1$ – сигнал есть, или гипотезы H_0 о том, что $\mathbf{X} = x_0$ – сигнал отсутствует. В математическом отношении эта задача эквивалентна задаче оптимального разбиения пространства принимаемых сигналов $\mathbf{V}$ на области v_1 и v_0. Если принятый вектор $\mathbf{Y}$ окажется в области v_1, принимается гипотеза H_1, если же он окажется в области v_0, принимается гипотеза H_0.

Для построения правила принятия решения о выборе гипотезы (разбиения пространства принимаемых сигналов) в рассмотрение вводится так называемая функция (отношение) правдоподобия:

$$\lambda = \frac{L(x_1)}{L(x_0)} = \frac{w(\mathbf{Y}/x_1)}{w(\mathbf{Y}/x_0)}. \tag{15.4}$$

Рассмотрим различные критерии принятия решений, формулируемые в терминах отношения правдоподобия (15.4).

15.2 Обнаружение по критерию максимального правдоподобия

По этому критерию наиболее правдоподобным считается то значение $\mathbf{X}$, для которого функция правдоподобия максимальна. Поскольку в задаче обнаружения рассматривается две альтернативы, существо дела сводится к сравнению $L(x_1)$ и $L(x_0)$. При этом решающее правило в терминах отношения правдоподобия принимает вид:

$$\text{если } \lambda = \frac{L(x_1)}{L(x_0)} > 1, \text{ то } \mathbf{X} = x_1, \tag{15.5}$$

$$\text{если } \lambda = \frac{L(x_1)}{L(x_0)} \le 1, \text{ то } \mathbf{X} = x_0, \tag{15.6}$$

Важное достоинство критерия максимума правдоподобия состоит в том, что в данном случае не требуется знание априорных вероятностей $p(x_1)$, $p(x_0)$ сигнала $\mathbf{X}$.

15.3 Обнаружение сигналов по критерию максимума апостериорной вероятности

В соответствии с этим критерием сравниваются значения апостериорных вероятностей $p(x_1/\mathbf{Y})$ и $p(x_0/\mathbf{Y})$:

$$\text{если } \frac{p(x_1/\mathbf{Y})}{p(x_0/\mathbf{Y})} > 1, \text{ то } \mathbf{X} = x_1, \tag{15.7}$$

$$\text{если } \frac{p(x_1/\mathbf{Y})}{p(x_0/\mathbf{Y})} \le 1, \text{ то } \mathbf{X} = x_0. \tag{15.8}$$

С использованием формулы Байеса (15.2) и равенства (15.3) отношение апостериорных вероятностей выражается через отношение правдоподобия:

$$\frac{p(x_1/\mathbf{Y})}{p(x_0/\mathbf{Y})} = \frac{p(x_1)L(x_1)}{p(x_0)L(x_0)} = \frac{p(x_1)}{p(x_0)}\lambda.$$

При этом критерий можно записать следующим образом:

$$\text{если } \frac{p(x_1)}{p(x_0)}\lambda > 1, \text{ то } \mathbf{X} = x_1, \tag{15.9}$$

$$\text{если } \frac{p(x_1)}{p(x_0)}\lambda \le 1, \text{ то } \mathbf{X} = x_0. \tag{15.10}$$

Решающее правило можно также представить в виде:

$$\text{если } \lambda > \frac{p(x_0)}{p(x_1)} = \lambda_0, \text{ то } \mathbf{X} = x_1, \tag{15.11}$$

$$\text{если } \lambda \le \frac{p(x_0)}{p(x_1)} = \lambda_0, \text{ то } \mathbf{X} = x_0, \tag{15.12}$$

где λ_0 – пороговое значение отношения правдоподобия. Критерий максимума апостериорной вероятности применяется в случае, когда известны априорные вероятности $p(x_1)$, $p(x_0)$ сигнала $\mathbf{X}$.

15.4 Информационный критерий обнаружения

С точки зрения теории информации наиболее предпочтительно то значение $\mathbf{X}$, относительно которого в $\mathbf{Y}$ содержится больше информации:

$$\begin{aligned} &I(\mathbf{Y}, x_1) - I(\mathbf{Y}, x_0) = \left[-\log_2 p(x_1) + \log_2 p(x_1/\mathbf{Y})\right] - \\ &-\left[-\log_2 p(x_0) + \log_2 p(x_0/\mathbf{Y})\right] = \\ &= \log_2 \frac{p(x_1/\mathbf{Y})p(x_0)}{p(x_0/\mathbf{Y})p(x_1)} = \log_2 \frac{p(\mathbf{Y}/x_1)}{p(\mathbf{Y}/x_0)} = \log_2 \lambda. \end{aligned} \tag{15.13}$$

В соответствии с информационным критерием (15.13), если логарифм отношения правдоподобия положителен, следует принять гипотезу H_1 ($\mathbf{X} = x_1$), если отрицателен или равен нулю – H_0 ($\mathbf{X} = x_0$).

Нетрудно заметить, что этот критерий совпадает с критерием максимального правдоподобия (15.5), (15.6).

15.5 Обнаружение по критерию Неймана-Пирсона

При решении задачи обнаружения сигналов могут иметь место ошибки двух типов:

1) ошибка *первого рода* – «ложная тревога» (при отсутствии сигнала принята гипотеза H_1 – $\mathbf{X} = x_1$), вероятность которой определяется как

$$\alpha = \int_{v_1} w(\mathbf{Y}/x_0)d\mathbf{Y}; \qquad (15.14)$$

2) ошибка *второго рода* «пропуск сигнала» (при наличии сигнала принята гипотеза $H_0 - \mathbf{X} = x_0$), вероятность которой

$$\beta = \int_{v_0} w(\mathbf{Y}/x_1)d\mathbf{Y}. \qquad (15.15)$$

При этом общая вероятность ошибочного решения

$$p_{ош} = p(x_0)\alpha + p(x_1)\beta. \qquad (15.16)$$

В соответствии с критерием Неймана–Пирсона наилучшим считается решение, при котором

$$\beta = \int_{v_0} w(\mathbf{Y}/x_1)d\mathbf{Y} \to \min,$$

при условии, что

$$\alpha = \int_{v_1} w(\mathbf{Y}/x_0)d\mathbf{Y} = \varepsilon,$$

где ε – заданная величина.

Рассмотрим решение указанной задачи для простейшего случая, когда $\mathbf{Y} = y$ – скаляр. При этом

$$\alpha = \int_{\lambda_0}^{\infty} w(y/x_0)dy, \ \beta = \int_{0}^{\lambda_0} w(y/x_1)dy,$$

а функция Лагранжа принимает вид

$$F = \int_{0}^{\lambda_0} w(y/x_1)dy + \lambda\left[\int_{\lambda_0}^{\infty} w(y/x_0)dy - \varepsilon\right].$$

Необходимые условия экстремума $(\partial F/\partial\lambda_0 = 0, \quad \partial F/\partial\lambda = 0)$

$$w(\lambda_0/x_1) - \lambda \cdot w(\lambda_0/x_0) = 0, \qquad (15.17)$$

$$\int_{\lambda_0}^{\infty} w(y/x_0)dy = \varepsilon. \qquad (15.18)$$

В соответствии с (15.17)

$$w(\lambda_0/x_1)/w(\lambda_0/x_0) = \lambda. \qquad (15.19)$$

С другой стороны, в соответствии с (15.4)

$$\frac{w(y/x_1)}{w(y/x_0)}=\frac{w(\mathbf{Y}/x_1)}{w(\mathbf{Y}/x_0)}=\lambda,$$

следовательно,

$$\frac{w(\lambda_0/x_1)}{w(\lambda_0/x_0)}=\lambda_0,$$

где пороговое значение λ_0 определяется из необходимого условия (15.18):

$$\int_{\lambda_0}^{\infty} w(y/x_0)dy=\varepsilon.$$

Таким образом, решающее правило можно записать в виде:

если $\frac{w(\mathbf{Y}/x_1)}{w(\mathbf{Y}/x_0)}=\lambda>\lambda_0$, то $\mathbf{X}=x_1$,

если $\frac{w(\mathbf{Y}/x_1)}{w(\mathbf{Y}/x_0)}=\lambda\le\lambda_0$, то $\mathbf{X}=x_0$.

15.6 Обнаружение сигналов по критерию минимального риска

Этот критерий является обобщением критерия Неймана-Пирсона. Он учитывает также потери, к которым могут привести ошибки первого и второго рода. Для этого ошибкам первого и второго рода ставятся в соответствие веса r_α, r_β, характеризующие цены ошибок, а величину r, определяемую как

$$r=r_\alpha p(x_0)\alpha+r_\beta p(x_1)\beta, \qquad (15.20)$$

называют *риском*. В соответствии с критерием принимается гипотеза, при которой обеспечивается минимум риска.

Подставляя в (15.20) выражения для ошибок первого и второго рода, можно записать

$$\begin{aligned} r&=r_\alpha p(x_0)\int_{V_1} w(\mathbf{Y}/x_0)d\mathbf{Y}+r_\beta p(x_1)\int_{V_0} w(\mathbf{Y}/x_1)d\mathbf{Y}=\\ &=r_\beta p(x_1)-\int_{V_1}\left[r_\beta p(x_1)w(\mathbf{Y}/x_1)-r_\alpha p(x_0)w(\mathbf{Y}/x_0)\right]d\mathbf{Y}. \end{aligned} \qquad (15.21)$$

Минимум в (15.21) будет достигаться только при условии положительности подынтегральной функции:

$$r_\beta p(x_1) w(\mathbf{Y}/x_1) - r_\alpha p(x_0) w(\mathbf{Y}/x_0) > 0. \qquad (15.22)$$

В соответствии с (15.22) решающее правило принимает вид

$$\text{если } \frac{w(\mathbf{Y}/x_1)}{w(\mathbf{Y}/x_0)} = \lambda > \frac{r_\alpha p(x_0)}{r_\beta p(x_1)} = \lambda_0, \text{ то } \mathbf{X} = x_1, \qquad (15.23)$$

$$\text{если } \frac{w(\mathbf{Y}/x_1)}{w(\mathbf{Y}/x_0)} = \lambda \le \frac{r_\alpha p(x_0)}{r_\beta p(x_1)} = \lambda_0, \text{ то } \mathbf{X} = x_0. \qquad (15.24)$$

Критерий минимального риска обеспечивает принятие наиболее обоснованного решения, учитывающего также и экономические потери. Достигается это за счет использования более богатой априорной информации. Помимо функций распределения $w(\mathbf{Y}/\mathbf{X})$ и априорных вероятностей $p(\mathbf{X})$ в данном случае необходимо знать цены потерь r_α, r_β.

15.7 Различение сигналов

В данном случае сигнал $\mathbf{X}$ может иметь m возможных значений x_1, x_2, …, x_m с априорными вероятностями $p(x_1)$, $p(x_2)$, …, $p(x_m)$:

$$\mathbf{X} = \begin{cases} x_1 \to p(x_1); \\ x_2 \to p(x_2); \\ \quad \dots \quad \dots \\ x_m \to p(x_m). \end{cases}$$

При этом пространство принимаемых сигналов разбивается на m областей: $v_1, v_2, ..., v_m$. Соответственно, выдвигается m гипотез: $H_1, H_2, ..., H_m$ о том, что $\mathbf{X} = x_1$, $\mathbf{X} = x_2$, …, $\mathbf{X} = x_m$.

Процедура различения гипотез строится как дерево решений. По принятому вектору $\mathbf{Y}$ определяются функции правдоподобия:

$$L(x_1) = p(\mathbf{Y}/x_1), \quad L(x_2) = p(\mathbf{Y}/x_2), \quad \dots \quad , L(x_m) = p(\mathbf{Y}/x_m)$$

и вычисляются отношения правдоподобия

$$\lambda_{i,j} = \frac{p(\mathbf{Y}/x_j)}{p(\mathbf{Y}/x_i)}$$

для всех возможных сочетаний пар x_i, x_j.

Полученные значения $\lambda_{i,j}$ сравниваются с заданными пороговыми и принимается гипотеза, для которой все $\lambda_{i,j} > \lambda_{0,}$ $j = \overline{1,m}$. Описанная выше процедура может быть реализована в сочетании с любым из рассмотренных выше критериев.

Лекция 16

Оценка параметров сигналов

16.1 Общая формулировка задачи восстановления сигналов

Восстановление сигналов сводится к оценке некоторого числа параметров. Задача ставится следующим образом [11]. Пусть сигнал является функцией некоторого аргумента, например, времени t:

$$y(t) = f(c_1,...,c_M,t) = f(\mathbf{c},t). \qquad (16.1)$$

Задача состоит в том, чтобы по принятой последовательности (вектору $\mathbf{Y} = [y_1, y_2,..., y_N]^T$) определить вектор параметров $\mathbf{c} = [c_1,...,c_M]^T$.

Другими словами, ищется

$$\hat{\mathbf{c}}: \qquad Q(\hat{\mathbf{c}}) = \min_{\mathbf{c}} Q(\mathbf{c}), \qquad (16.2)$$

где $Q(\mathbf{c})$ – некоторый критерий, характеризующий качество восстановления сигнала. Вид критерия качества определяется доступной априорной информацией.

Наиболее широко в задачах восстановления используются линейные зависимости сигнала от искомых параметров. При оценке параметров динамических моделей это достигается линеаризацией в окрестности рабочей точки. При этом искомые параметры имеют смысл коэффициентов влияния малых отклонений сигналов от некоторого заданного (установившегося) рабочего режима.

Часто функциональную зависимость общего вида (16.1) специально представляют в виде, допускающем преобразование ее к линейной модели, например, экспоненциальными зависимостями. При этом преобразование к линейной относительно искомых параметров модели осуществляется путем логарифмирования.

В качестве зависимостей (16.1) широко используются также ортогональные представления сигналов (см. раздел 1.2):

$$y(t) = \sum_{k=1}^{M} c_k \varphi_k(t),$$

где $\varphi_k(t)$ – заданные ортогональные или ортонормированные базисные функции, а c_k – искомые коэффициенты. Нетрудно заметить, что эти модели также линейные по искомым параметрам.

16.2 Задача оценки параметров линейных моделей

В случае дискретного аргумента и аддитивных ошибок измерений ξ_k, $k=1,2,\ldots$ линейную модель сигнала можно представить в виде

$$y_k = \mathbf{x}_k^T \mathbf{c} + \zeta_k, \quad k=1,2,\ldots \tag{16.3}$$

Если вектор искомых параметров $\mathbf{c}$ в пределах допустимой точности модели считается неизменным для различных k, после проведения N измерений y_k, $\mathbf{x}_k$, $k=\overline{1,N}$, в соответствии с (16.3), можно записать векторно-матричное соотношение [9]

$$\mathbf{Y} = \mathbf{X}\mathbf{c} + \boldsymbol{\xi}, \tag{16.4}$$

где $\mathbf{Y}$, $\boldsymbol{\xi}$ – $N\times 1$-векторы, а $\mathbf{X}$ – $N\times M$-матрица.

Задача оценки $M\times 1$-вектора параметров $\mathbf{c}$ состоит в построении приближенных соотношений

$$\hat{\mathbf{c}} \cong h(\boldsymbol{\xi}).$$

Естественно стремление строить оценки, обладающие «хорошими» свойствами. Обычно рассматривают следующие свойства оценок.

1. Несмещенность. Оценка $\hat{\mathbf{c}}$ векторного параметра $\mathbf{c}$ называется несмещенной, если

$$M\{\hat{\mathbf{c}}\} = \mathbf{c}. \tag{16.5}$$

2. Состоятельность. Последовательность оценок $\hat{\mathbf{c}}_k$ называется состоятельной, если для сколь угодно малого $\varepsilon > 0$ с ростом k

$$\lim_{k\to\infty} \mathrm{P}\{|\hat{\mathbf{c}}_k - \mathbf{c}| > \varepsilon\} = 0, \tag{16.6}$$

т.е. $\hat{\mathbf{c}}_k$ сходится по вероятности к истинному значению $\mathbf{c}$.

3. Эффективность. Оценка $\hat{\mathbf{c}}$ называется эффективной, если для любой несмещенной оценки $\hat{\mathbf{b}}$

$$M\left\{(\hat{\mathbf{c}}-\mathbf{c})(\hat{\mathbf{c}}-\mathbf{c})^T\right\} \le M\left\{(\hat{\mathbf{b}}-\mathbf{c})(\hat{\mathbf{b}}-\mathbf{c})^T\right\}. \tag{16.7}$$

Неравенство $\mathbf{A} \le \mathbf{B}$ здесь понимается в том смысле, что матрица $\mathbf{B}-\mathbf{A}$ неотрицательно-определенная.

16.3 Достижимая точность, неравенство Крамера-Рао

При построении оценок одним из основных является следующий вопрос: какова наивысшая (предельная) точность возможна на имеющихся наблюдениях и на каких оценках она достигается. Важнейшей характеристикой точности оценивания векторного параметра является ковариационная матрица

$$\mathbf{D}(\hat{\mathbf{c}}) = M\left\{(\hat{\mathbf{c}}-\mathbf{c})(\hat{\mathbf{c}}-\mathbf{c})^T\right\}. \tag{16.8}$$

Построим неравенство (Крамера-Рао), характеризующее ее нижнюю границу.

Пусть выборочный вектор ξ:

$$\xi = \mathbf{Y} - \mathbf{X}\mathbf{c} \tag{16.9}$$

обладает плотностью распределения $w(\xi)$. Введем в рассмотрение так называемую информационную матрицу Фишера:

$$\mathbf{I}(\mathbf{c}) = M\left\{\nabla_{\mathbf{c}} \ln\, w(\xi) \nabla_{\mathbf{c}^T} \ln\, w(\xi)\right\} \tag{16.10}$$

с элементами $\mathbf{I}_{i,j}(\mathbf{c}) = M\left\{\dfrac{\partial}{\partial c_i}\ln\, w(\xi)\dfrac{\partial}{\partial c_j}\ln\, w(\xi)\right\}$.

Теперь запишем заведомо неотрицательно-определенную матрицу:

$$\begin{aligned}\mathbf{B} = M\Big\{&\left[\mathbf{I}^{-1}(\mathbf{c})\nabla_{\mathbf{c}} \ln\, w(\xi) - (\hat{\mathbf{c}}-\mathbf{c})\right]\times \\ &\times\left[\mathbf{I}^{-1}(\mathbf{c})\nabla_{\mathbf{c}} \ln\, w(\xi) - (\hat{\mathbf{c}}-\mathbf{c})\right]^T\Big\} \ge 0.\end{aligned} \tag{16.11}$$

После перемножения и взятия операции математического ожидания, с учетом (16.8), (16.10), имеем (для краткости, вместо $\mathbf{I}(\mathbf{c})$ здесь и далее используется обозначение $\mathbf{I}$)

$$\begin{aligned}\mathbf{B} = {}&\mathbf{I}^{-1}\mathbf{I}\mathbf{I}^{-1} - \mathbf{I}^{-1}M\left\{\nabla_{\mathbf{c}} \ln w(\xi)\cdot(\hat{\mathbf{c}}-\mathbf{c})^T\right\} - \\ &-M\left\{(\hat{\mathbf{c}}-\mathbf{c})\nabla_{\mathbf{c}^T} \ln w(\xi)\right\}\mathbf{I}^{-1} + \mathbf{D}(\hat{\mathbf{c}}) \ge 0.\end{aligned} \tag{16.12}$$

Предполагая, что функция плотности вероятности $w(\xi)$ допускает дифференцирование под знаком интеграла, вычислим градиент от обеих частей равенства нормировки $\int w(\xi)d\xi = 1$:

$$\nabla_{\mathbf{c}} \int w(\xi)d\xi = \int \nabla_{\mathbf{c}} w(\xi) \cdot \frac{1}{w(\xi)} w(\xi)d\xi = M\left\{\nabla_{\mathbf{c}} \ln w(\xi)\right\} = 0. \qquad (16.13)$$

Аналогично из условия несмещенности оценок параметров

$$\int \hat{\mathbf{c}} w(\xi)d\xi = \mathbf{c}$$

с учетом того, что $\nabla_{\mathbf{c}^T}\mathbf{c} = \nabla_{\mathbf{c}}\mathbf{c}^T = \mathbf{E}$, где $\mathbf{E}$ – единичная матрица, имеем

$$\int \hat{\mathbf{c}} \cdot \nabla_{\mathbf{c}^T} w(\xi)d\xi = \int \hat{\mathbf{c}} \cdot \frac{\nabla_{\mathbf{c}^T} w(\xi)}{w(\xi)} w(\xi)d\xi = \\ = M\left\{\hat{\mathbf{c}} \cdot \nabla_{\mathbf{c}^T} \ln w(\xi)\right\} = M\left\{\nabla_{\mathbf{c}} \ln w(\xi)\hat{\mathbf{c}}^T\right\} = \mathbf{E}. \qquad (16.14)$$

С учетом (16.13), (16.14) и очевидного равенства $\mathbf{I}\mathbf{I}^{-1} = \mathbf{E}$ неравенство (16.12) можно переписать в виде

$$\mathbf{I}^{-1} - \mathbf{I}^{-1} - \mathbf{I}^{-1} + D(\hat{\mathbf{c}}) \geq 0$$

или

$$\mathbf{D}(\hat{\mathbf{c}}) \geq \mathbf{I}^{-1}\left[\xi(\mathbf{c})\right]. \qquad (16.15)$$

Мы получили неравенство Крамера-Рао, которое устанавливает нижнюю границу дисперсий оценок в классе всех несмещенных оценок. Заметим, что это неравенство получено при самых общих предположениях о выполнении условия нормировки и свойства несмещенности оценок, не связанных с методом оценивания. Оно позволяет судить, насколько данная оценка близка к оптимальной.

16.4 Оценки, минимизирующие среднеквадратическую ошибку

Они используются в условиях статистической неопределенности, когда нет сведений о распределении ошибок. В этом случае, опираясь на восходящее к Гауссу мнение, считают, что наилучшей является оценка, минимизирующая средневзвешенную квадратическую ошибку:

$$Q(\mathbf{c}) = \frac{1}{2}\sum_{i,j=1}^{N} g_{i,j}\xi_i\xi_j \,.$$

В векторно-матричной форме критерий запишется в виде

$$Q(\mathbf{c}) = \frac{1}{2}\xi^T\mathbf{G}\cdot\xi, \tag{16.16}$$

где $\mathbf{G}$ – заданная положительно-определенная $N\times N$-матрица.

Если известна ковариационная матрица $\mathbf{K} = M\left\{\xi\cdot\xi^T\right\}$ коррелированной помехи с нулевым средним, то матрицу $\mathbf{G}$ обычно задают в виде $\mathbf{G} = \mathbf{K}^{-1}$:

$$Q(\mathbf{c}) = \frac{1}{2}\xi^T\mathbf{K}^{-1}\xi\,. \tag{16.17}$$

Оценку (16.17) называют оценкой обобщенного метода наименьших квадратов (ОМНК) или оценкой Гаусса-Маркова.

Если об ошибках измерений ничего не известно и нет никаких оснований, отдать предпочтение каким-либо измерениям, полагают $\mathbf{G} = \mathbf{E}$:

$$Q(\mathbf{c}) = \frac{1}{2}\xi^T\xi\,. \tag{16.18}$$

Соответствующая этому критерию оценка наиболее широко используется на практике и называется оценкой метода наименьших квадратов (МНК).

16.5 Оценка максимального правдоподобия

Метод максимального правдоподобия используется в случае, когда априори известна плотность распределения $w(\xi)$. Он основан на интуитивном представлении, что наиболее правдоподобна оценка, соответствующая максимальному значению плотности распределения.

Поскольку функция $\ln w(\xi)$ достигает максимума в тех же точках, что и $w(\xi)$, в качестве функции потерь обычно применяют

$$Q(\mathbf{c}) = -\ln w\left[\xi(\mathbf{c})\right]. \tag{16.19}$$

В случае гауссовых помех совместная плотность вероятности

$$w(\xi) = (2\pi)^{-\frac{N}{2}}(\det\mathbf{K})^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\xi^T\mathbf{K}^{-1}\xi\right). \tag{16.20}$$

При этом, в соответствии с (16.19), получаем

$$Q(\mathbf{c}) = -\ln w(\xi) = -\ln\left[(2\pi)^{-\frac{N}{2}}(\det \mathbf{K})^{-\frac{1}{2}}\right] + \frac{1}{2}\xi^T \mathbf{K}^{-1}\xi. \qquad (16.21)$$

Нетрудно заметить, что первое слагаемое в правой части не зависит от искомых параметров, а второе слагаемое совпадает (16.17). Следовательно, критерий максимального правдоподобия совпадает с ОМНК при гауссовых помехах.

16.6 Оптимальность оценок МНК и максимального правдоподобия

Покажем, что в случае нормального распределения ошибок ОМНК-оценка и совпадающая с ней оценка максимального правдоподобия оптимальны в смысле минимума дисперсии. Для этого достаточно показать, что ковариационная матрица ошибок оценивания совпадает с обратной информационной матрицей Фишера.

Выпишем ковариационную матрицу ошибок оценивания. В соответствии с (16.17) с учетом того, что $\xi = \mathbf{Y} - \mathbf{X}\mathbf{c}$, искомая ОМНК-оценка является решением уравнения

$$\nabla_{\mathbf{c}} Q(\mathbf{c}) = \nabla_{\mathbf{c}} \frac{1}{2}\xi^T \mathbf{K}^{-1}\xi = \mathbf{X}^T \mathbf{K}^{-1}\xi = \mathbf{X}^T \mathbf{K}^{-1}\mathbf{Y} - \mathbf{X}^T \mathbf{K}^{-1}\mathbf{X}\hat{\mathbf{c}} = 0,$$

т.е. $$\hat{\mathbf{c}} = \mathbf{R}\mathbf{Y}, \qquad (16.22)$$

где $$\mathbf{R} = \left[\mathbf{X}^T \mathbf{K}^{-1}\mathbf{X}\right]^{-1} \mathbf{X}^T \mathbf{K}^{-1}. \qquad (16.23)$$

Подставляя в (16.22) $\mathbf{Y} = \mathbf{X}\mathbf{c} + \xi$ из (16.4), с учетом того, что в соответствии с (16.23) $\mathbf{R}\mathbf{X} = \left[\mathbf{X}^T \mathbf{K}^{-1}\mathbf{X}\right]^{-1} \mathbf{X}^T \mathbf{K}^{-1}\mathbf{X} = \mathbf{E}$, имеем

$$\hat{\mathbf{c}} = \mathbf{R}\mathbf{X}\mathbf{c} + \mathbf{R}\xi = \mathbf{c} + \mathbf{R}\xi. \qquad (16.24)$$

Теперь, с использованием (16.24) запишем ковариационную матрицу ошибок оценивания:

$$\mathbf{D}(\hat{\mathbf{c}}) = M\left\{(\hat{\mathbf{c}} - \mathbf{c})(\hat{\mathbf{c}} - \mathbf{c})^T\right\} = M\left\{(\mathbf{R}\xi)(\mathbf{R}\xi)^T\right\} = \mathbf{R}M\left\{\xi\xi^T\right\}\mathbf{R}^T = \mathbf{R}\mathbf{K}\mathbf{R}^T.$$

Наконец, подставляя в последнее равенство матрицу $\mathbf{R}$ из (16.23), окончательно получаем

$$\mathbf{D}(\hat{\mathbf{c}}) = \left[\mathbf{X}^T\mathbf{K}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{K}^{-1}\mathbf{K}\mathbf{K}^{-1}\mathbf{X}\left[\mathbf{X}^T\mathbf{K}^{-1}\mathbf{X}\right]^{-1} = \left[\mathbf{X}^T\mathbf{K}^{-1}\mathbf{X}\right]^{-1}. \qquad (16.25)$$

Теперь запишем информационную матрицу Фишера (16.10) для гауссовой плотности (16.20). С учетом (16.21)

$$\nabla_{\mathbf{c}} \ln\, w(\boldsymbol{\xi}) = \nabla_{\mathbf{c}} \frac{1}{2}\boldsymbol{\xi}^T\mathbf{K}^{-1}\boldsymbol{\xi} = \mathbf{X}^T\mathbf{K}^{-1}\boldsymbol{\xi}.$$

Отсюда в соответствии с определением (16.10) сразу получаем

$$\mathbf{I}(\mathbf{c}) = M\left\{\mathbf{X}^T\mathbf{K}^{-1}\boldsymbol{\xi}\boldsymbol{\xi}^T\mathbf{K}^{-1}\mathbf{X}\right\} = \mathbf{X}^T\mathbf{K}^{-1}M\left\{\boldsymbol{\xi}\boldsymbol{\xi}^T\right\}\mathbf{K}^{-1}\mathbf{X} = \mathbf{X}^T\mathbf{K}^{-1}\mathbf{X}. \qquad (16.26)$$

Подставляя полученные выражения для $\mathbf{D}(\hat{\mathbf{c}})$ и $\mathbf{I}(\mathbf{c})$ из (16.25) (16.26) в неравенство (16.15) (Крамера-Рао), убеждаемся, что оно превращается в равенство, следовательно, оценки максимального правдоподобия и ОМНК-оценки оптимальны и достигается нижняя граница дисперсий.

16.7 Байесовские оценки

Два метода: максимальной апостериорной вероятности и минимального среднего риска обычно называют байесовскими, т.к. для их построения используется формула Байеса (15.1):

$$w(\mathbf{c}/\mathbf{Y}) = \frac{w(\mathbf{c})w(\mathbf{Y}/\mathbf{c})}{w(\mathbf{Y})}, \qquad \text{где} \qquad w(\mathbf{Y}) = \int_{\mathbf{c}} w(\mathbf{c})w(\mathbf{Y}/\mathbf{c})d\mathbf{c}.$$

Апостериорная плотность вероятности описывает частоты появления значений параметров после того, как к априорной информации добавлена информация, извлеченная из наблюдений. Поэтому естественно в качестве оценок принять значения, соответствующие наибольшим апостериорным вероятностям или минимуму взятого со знаком минус логарифма плотности:

$$\hat{\mathbf{c}}: \qquad Q(\hat{\mathbf{c}}) = \min_{\mathbf{c}}\left[\ln w(\mathbf{Y}) - \ln w(\mathbf{c}) - \ln w(\mathbf{Y}/\mathbf{c})\right]. \qquad (16.27)$$

Первый член в квадратных скобках не зависит от $\mathbf{c}$, поэтому в качестве функции потерь можно принять

$$Q(\mathbf{c}) = -\left[\ln w(\mathbf{c}) + \ln w(\mathbf{Y}/\mathbf{c})\right].$$

Если плотности вероятностей гауссовы, критерий принимает вид

$$Q(\mathbf{c}) = \boldsymbol{\xi}^T\mathbf{K}^{-1}\boldsymbol{\xi} + (\mathbf{c}-\overline{\mathbf{c}})^T\mathbf{K}_{\mathbf{c}}^{-1}(\mathbf{c}-\overline{\mathbf{c}}), \qquad (16.28)$$

где $\mathbf{K}_c$, $\overline{\mathbf{c}}$ – ковариационная матрица и априорное среднее вектора $\mathbf{c}$ соответственно. Сравнивая (16.28) с (16.17), (16.21), легко заметить отличие метода максимальной апостериорной вероятности от ОМНК и метода максимального правдоподобия.

Пусть теперь вдобавок к априорной информации, которая использовалась при построении оценок максимальной апостериорной вероятности, известны также потери $\Pi(\mathbf{c},\hat{\mathbf{c}})$, связанные с численной величиной оценки $\hat{\mathbf{c}}$ при истинном значении вектора $\mathbf{c}$. Тогда мерой качества целесообразно выбрать функцию средних потерь по всевозможным наблюдениям для каждого фиксированного значения вектора параметров $\mathbf{c}$:

$$L(\mathbf{c},\hat{\mathbf{c}}) = M\{\Pi(\mathbf{c},\hat{\mathbf{c}})\} = \int_{\mathbf{Y}} \Pi(\mathbf{c},\hat{\mathbf{c}})\, w(\mathbf{Y}/\mathbf{c})dy,$$

которая называется функцией *условного риска*.

Функция среднего риска получается усреднением условного риска по всем возможным значениям случайных параметров $\mathbf{c}$:

$$\begin{aligned} L(\mathbf{c}) &= M\{L(\mathbf{c},\hat{\mathbf{c}})\} = \int_{\mathbf{c}} L(\mathbf{c},\hat{\mathbf{c}})\, w(\mathbf{c})d\mathbf{c} = \\ &= \int_{\mathbf{c}}\int_{\mathbf{Y}} \Pi(\mathbf{c},\hat{\mathbf{c}}) w(\mathbf{Y}/\mathbf{c})\, w(\mathbf{c})\, d\mathbf{Y}\, d\mathbf{c} = \int_{\mathbf{Y}}\int_{\mathbf{c}} \Pi(\mathbf{c},\hat{\mathbf{c}}) w(\mathbf{Y})\, w(\mathbf{c}/\mathbf{Y})\, d\mathbf{c}\, d\mathbf{Y} = \\ &= \int_{\mathbf{Y}} w(\mathbf{Y}) \left[\int_{\mathbf{c}} \Pi(\mathbf{c},\hat{\mathbf{c}}) w(\mathbf{c}/\mathbf{Y})\, d\mathbf{c} \right] d\mathbf{Y}. \end{aligned} \qquad (16.29)$$

Функция (16.29) минимальна, когда достигает минимума внутренний интеграл. Следовательно, искомый критерий представляется в виде

$$Q(\mathbf{c}) = \int_{\mathbf{c}} \Pi(\mathbf{c},\hat{\mathbf{c}}) w(\mathbf{c}/\mathbf{Y})\, d\mathbf{c}.$$

Отметим, что использование байесовских оценок на практике часто затруднено, из-за невозможности получить необходимую априорную информацию.

Список использованных источников

1. Биркгоф Г., Барти Т. Современная прикладная алгебра: Пер. с англ. – М.: Мир. – 1976. – 400 с.
2. Гилл А. Линейные последовательные машины: Пер. с англ. – М.: Наука. –1974. – 287 с.
3. Дмитриев В.И. Прикладная теория информации: Учеб. пособие. – М.: Высш. шк., 1989. – 320 с.
4. Кловский Д.Д. Теория передачи сигналов. – М.: Связь, 1973. – 376 с.
5. Колмогоров А.Н. Теория передачи информации // Сессия Академии наук СССР по научным проблемам автоматизации производства, 15-20 окт. 1956 г.: Пленар. заседания. – М.: Изд-во АН СССР, 1957. – С. 66-99.
6. Кузнецов Н.А. Информационное взаимодействие в технических и живых системах [Электронный ресурс]. – 2001. Т. 1. – № 1. С. 1-9.
7. Кузьмин И.В., Кедрус В.А. Основы теории информации и кодирования. 2-е изд. перераб. и доп. – Киев.: Вища шк. 1986. – 238 с.
8. Лифшиц Н.А., Пугачев В.Н. Вероятностный анализ систем автоматического управления. – М.: Сов. радио. – 1963. – 896 с.
9. Питерсон У. Коды, исправляющие ошибки: Пер. с англ. – М.: Мир, 1964. – 340 с.
10. Сойфер В.А. Теория информации: Учеб. пособие. – Куйбышев: КуАИ, 1977. – 80 с.
11. Фурсов В.А. Определение характеристик объектов в адаптивных системах управления: Учеб. пособие / Под ред. Б.М. Шамрикова – М.: МАИ, 1983. – 46 с.
12. Харкевич А.А. Спектры и анализ. – М.: Физматгиз, 1962. – 236 с.
13. Цыпкин Я.З. Основы информационной теории идентификации. – М.: Наука. Гл. ред. физ.-мат. лит. 1984. – 320 с.

ОГЛАВЛЕНИЕ

Учебное издание

Фурсов Владимир Алексеевич

ЛЕКЦИИ ПО ТЕОРИИ ИНФОРМАЦИИ

Учебное пособие

Технический редактор *С. Б. Попов*

Редакторская обработка *О. Ю. Дьяченко*

Корректорская обработка *А. В. Ярославцева, О. Ю. Дьяченко*

Верстка *Н. Е. Козин*

Доверстка *А. А. Нечитайло*

Подписано в печать 1.12.06. Формат 60×84 1/16

Бумага офсетная. Печать офсетная.

Усл. печ. л. 8,6. Усл. кр.-отт. 8,72. Печ. л. 9,25

Тираж 50 экз. Заказ . ИП-86/2006

Самарский государственный аэрокосмический университет.

443086, Самара, Московское шоссе, 34

Изд-во Самарского государственного аэрокосмического университета.

443086, Самара, Московское шоссе, 34