КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

ВЫСШАЯ ШКОЛА ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И ИНФОРМАЦИОННЫХ СИСТЕМ

Кафедра автономных робототехнических систем

Д.Е. Чикрин

ТЕОРИЯ ИНФОРМАЦИИ И КОДИРОВАНИЯ

Курс лекций

Казань — 2013

Принято на заседании Высшей школы информационных технологий и информационных систем

Протокол №8 от 20.08.2013

Научный редактор
кандидат техн. наук, старший научный сотрудник
ООО КБ "НТ" А.П. Овчаров

Рецензент
кандидат техн. наук, старший научный сотрудник
ООО КБ "НТ" О.С. Вершинин

Д.Е. Чикрин

Теория информации и кодирования: курс лекций / Д.Е. Чикрин. - Казань: Казанский университет, 2013. - 116 с.

Дисциплина "Теория информации и кодирования"является одной из основополагающих для инженерных и IT-дисциплин, специализирующихся на построении аппаратно-программных автоматизированных систем и комплексов, решающих задачи хранения, передачи и преобразования информации. Представленный курс представляет собой сжатое, но достаточно полное изложение наиболее важных аспектов теории информации и теории кодирования; состоит из четырех основных тем-разделов, посвященных основам теории информации, теоретическим основам построения каналов связи; эффективному кодированию (сжатию информационных сообщений) и помехоустойчивому кодированию (защите информационных сообщений от помех при передаче по каналам связи).

Аннотация

Дисциплина «Теория информации и кодирования» является одной из основополагающих для инженерных и IT-дисциплин, специализирующихся на построении аппаратно-программных автоматизированных систем и комплексов, решающих задачи хранения, передачи и преобразования информации.

Представленный курс представляет собой сжатое, но достаточно полное изложение наиболее важных аспектов теории информации и теории кодирования; состоит из четырех основных тем-разделов, посвященных основам теории информации, теоретическим основам построения каналов связи; эффективному кодированию (сжатию информационных сообщений) и помехоустойчивому кодированию (защите информационных сообщений от помех при передаче по каналам связи). При подготовке курса лекций особенное внимание уделялось понятности и последовательности изложения, впрочем, лишь читателю дозволено судить о полученном результате.

Эта работа, как и вся моя жизнь, посвящена моей жене. За знания, полученные мной, я благодарю моих Учителей. Спасибо, без вас этой книги никогда бы не было.

Чикрин Д.Е.

Abstract

Theory of coding and information is one of the major branches at the wide tree of IT courses. This course is intended to students, IT-specialists and engineers in the fields of different information systems and hardware complex construction.

In this book we are consider four main topics of the information theory: information theory basics, theory of communication links and channels, effective coding, antinoise coding.

Chickrin D.E.

Содержание

Тема I
Основы теории информации

Тема 1 – список литературы

1 *Дмитриев В.И.* Прикладная теория информации. М.: Букинист, 1989. - 332 с.

2 *Думачев В.Н.* Теория информации и кодирования. Воронеж: Воронежский институт МВД России, 2012. - 200 с.

3 *Прохоров В.С.* Теория информации - курс лекций. - 124 с.

4 *Лидовский В.В.* Теория информации - учебное пособие. М.: Спутник+, 2004. - 111 с.

5 *Финк Л.М.* Сигналы, помехи, ошибки. М.: Радио и связь, 1984. - 256 с.

6 *Липкин И.А.* Статистическая радиотехника. Теория информации и кодирования. М.: Вузовская книга, 2002. - 216 с.

7 *Скляр Б.* Цифровая связь. Теоретические основы и практическое применение. М.: изд. дом Вильямс, 2003. - 1104 с.

8 *Реньи А.* Трилогия о математике. М.: Мир, 1980. - 378 с.

9 *Юсупова Н.И, Сметанина О.Н. и др.* Методические указания «Вычислительные аспекты в задачах теории информации». Уфа: изд. Уфимского ГАТУ, 2003. - 29 с.

10 *Гмурман В.Е.* Теория вероятностей и математическая статистика. М.: Высшая Школа, 2003. - 480 с.

11 *Рогова Н.В., Федосеева О.А.* Комбинаторика и теория вероятностей. Пермь: изд. Пермского ГТУ, 2007. - 39 с.

12 *Гмурман В.Е.* Руководство к решению задач по теории вероятностей и математической статистике. М.: Высшая Школа, 2004. - 407 с.

13 *Орлов В.А., Филиппов Л.И.* Теория информации в упражнениях и задачах. М.: Высшая Школа, 1976. - 136 с.

14 *Кавчук С.В.* Сборник примеров и задач по теории информации. Таганрог: изд. Таганрогского ГРУ, 2002. - 64 с.

Лекция 1

Информация. Базовые понятия теории информации

1.1 Введение

Теория информации является существенной, неотъемлемой частью кибернетики - науки, изучающей общие законы получения, передачи и хранения информации. Основным предметом кибернетики являются так называемые кибернетические системы.

Кибернетические системы *– множество взаимосвязанных объектов – элементов системы –, а также связей между ними, обеспечивающих в своей совокупности восприятие, запоминание, переработку и обмен информацией.*

Под указанное определение подходит самый широкий класс систем - от систем автоматического управления на предприятиях, до биологических популяций и социумов. Подобное многообразие и широта охвата присуща и «дочери» кибернетики - теории информации, изучению которой и будет посвящен текущий курс лекций.

Итак, теорией информации называется наука, изучающая количественные закономерности, связанные с восприятием (получением), запоминанием (хранением), переработкой (обработкой) и обменом (передачей) информации. Приведенные здесь синонимы являются более употребляемыми в системах связи, хранения и обработки информации, которые и будут далее рассматриваться нами. Действительно, теория информации и теория связи являются чрезвычайно взаимосвязанными науками; в частности, возникновение теории информации как отдельной дисциплины чаще всего связывают с фундаментальным трудом Клода Шеннона «Математическая теория связи».

Одной из основных задач теории информации, отображенной в заглавии нашего курса является определение оптимальных методов кодирования, позволяющих передавать и/или хранить и обрабатывать заданную информацию для заданных граничных условий - параметров канала связи; емкости запоминающих устройств; наличествующих вычислительных мощностей и пр.

Что же такое информация? Одним из наиболее часто употребляемых определений является следующее:

Информация - это совокупность сведений, подлежащих хранению, передаче, обработке и использованию в человеческой деятельности.

К сожалению, указанное определение является как минимум незавершенным (определяемым через неопределенное понятие - совокупность сведений), а также неточным с формальной точки зрения. С точки зрения автора, понятие информации может быть максимально точно понято из контекста принципа управления, сформулированного родоначальником кибернетики Норбертом Винером:

Движение и действие больших масс или передача и преобразование больших количеств энергии направляется и контролируется при помощи небольших количеств энергии, осуществляющих управление — несущих ***информацию****.*

Таким образом, если рассматривать всю природу в целом как кибернетическую систему, то энергию можно определить как основное свойство, определяющее элементы системы, а информацию - как основную характеристику связей между элементами - атрибут материи, отражающий взаимосвязь и взаимозависимость явлений и существующее во Вселенной разнообразие.

Указанный результат является в существенной мере философским, что, тем не менее, никак не умаляет его ценности для представленной далее прикладной дисциплины и полностью соответствует точке зрения автора.

1.2 Основные понятия теории информации

1.2.1 Основные термины и предмет теории информации

В любой системе информация представлена в виде ***сообщений*** — совокупности знаков, либо непрерывных сигналов, являющихся переносчиком информации.

Дискретные сообщения формируются в результате последовательной выдачи источником сообщений отдельных элементов - ***знаков***. При этом все множество возможных различных знаков называют ***алфавитом сообщения***, а размер множества - ***объемом алфавита***.

Непрерывные сообщения в свою очередь не разделены на элементы, а описываются непрерывными сигналами - функциями времени, принимающими значения из непрерывного континуума.

Среда, по которой передаются сообщения между источником и приемником сообщений называется ***каналом связи***, либо ***каналом передачи информации***. При этом преобразование сообщения в сигнал, подходящий для передачи по заданному каналу связи, называется ***кодированием*** (в широком смысле слова); обратную операцию называют ***декодированием***.

Во время прохождения сообщения по каналу связи в данном канале могут действовать мешающие воздействия - помехи (как внешние, так и внутренние).

Итак, исходя из введенных терминов, определим предмет теории информации:

Теорией информации исследуются информационные системы (кибернетические системы с ярко выраженными процессами передачи, хранения и преобразования информации), подчиняющиеся следующим постулатам:

1 Источник сообщения осуществляет выбор сообщения из некоторого множества (с определенными вероятностями выбора каждого из сообщений).

2 Сообщения могут передаваться по каналу связи в закодированном виде с возможностью однозначного декодирования на приемной стороне.

3 Сообщения следуют друг за другом, при этом количество сообщений может быть сколь угодно большим.

4 Сообщение считается принятым при успешно осуществленной (и однозначной!) операции декодирования. При этом не имеет значения, сколько времени прошло с момента передачи сообщения и какова вычислительная сложность операций кодирования и декодирования.

5 Количество информации является математической абстракцией; не зависит от смыслового содержания сообщения, его эмоционального воздействия, полезности и отношения к реальной действительности.

1.2.2 Количественная мера информации

Рассмотрим источник дискретных сообщений (дискретный источник информации). Пусть каждое отдельное i-е сообщение представляет собой информационный символ, выбираемый из ансамбля U размерности m с определенной для каждого элемента ансамбля вероятностью появления:

$$U = \begin{pmatrix} u_1 & u_2 \dots u_m \\ p_1 & p_2 \dots p_m \end{pmatrix} \tag{1.1}$$

Представим информацию как меру неопределенности источника сообщений. Так, детерминированные (представляющие собой сингулярный случай 1.1 при $m = 1$ сигналы не несут в себе полезной –информационной нагрузки (величина количества информации, обозначим его $I = 0$).

Возможно перечислить следующие естественные условия к I, как к количественной характеристике меры неопределенности:

1 Функция $I(m)$ должна быть неотрицательной и монотонно возрастающей (за исключением введения в ансамбль вырожденных элементов с вероятностью появления $p = 0$).

2 Функция I_X для любых сообщений X должна обладать свойством аддитивности:

$$I\left(m_{X1}\right) + I\left(m_{X2}\right) = I\left(m_{X1} + m_{X2}\right) \tag{1.2}$$

3 Количество информации I должно зависеть от вероятностей появления элементов ансамбля.

4 Количество информации I должно зависеть от вероятностей появления элементов ансамбля. Действительно, интуитивно ясно, что более редкое событие несет в себе большее количество информации.

Базовым условиям **1** и **2** удовлетворяет функция $I = \log m$; при этом указанная формула достаточно легко расширяется на случай сообщения из n символов, каждый из которых выбирается из ансамбля размерности m. Действительно, в этом случае разнообразие N сообщений дискретного источника определяется как число перестановок с неограниченными повторениями из m по n: $N = m^n$. Таким образом, результирующая формула, полученная Ральфом Хартли в 1928 позволяет определить количество информации в виде следующей функции:

$$I = \log N = \log m^n = n \log m. \tag{1.3}$$

В формуле 1.3 возможно использовать произвольное основание логарифма; от выбранного основания зависит единица измерения количества информации[1]. Наиболее распространенные основания - $e, 10$ и 2; соответствующие единицы измерения - нат, дит и бит. [2] В современной вычислительной технике, в связи с двоичной природой абсолютного большинства современных ЭВМ, в качестве безусловного стандарта принят ***бит***.

1.2.3 Энтропия

Для дискретного источника информации одной из ключевых характеристик является среднее количество информации, передаваемое в одном символе сообщения. Пусть вероятности p_i всех i-х элементов ансамбля U различны и составляют полную систему случайных событий:

$$\sum_{i=1}^{m} p_i = 1. \tag{1.4}$$

Итак, пусть путем эмпирических измерений определено, что в сообщении длины n каждый символ u_i входит n_i раз. В этом случае число всех возможных сообщений длины n определяется как число перестановок с повторениями из n элементов с количеством отдельных элементов

[1] При равновероятных элементах исходного ансамбля.

[2] Нат (nat) - *na*tural digi*t*;
дит (dit) - *d*ecimal digi*t*;
бит (bit) - *b*inary digi*t*; «кусочек» чего-либо.

в $\{n_1, n_2, \ldots, n_n\}$ и равно:

$$N = \frac{K!}{n_1! n_2! \ldots n_n!} \tag{1.5}$$

Таким образом, согласно формулам 1.5 и 1.3 количество информации может быть определено следующим образом:

$$I = \log N = \log n! - (\log n_1! + \log n_2! + \ldots + \log n_m!) \tag{1.6}$$

. Воспользовавшись формулой Стирлинга $\log n! \approx n(\ln n - 1)$ и соотношением $\sum_{i=1}^{m} n_i = n$, получаем:

$$\begin{aligned} I = \ln N &= n(\ln n - 1) - \sum_{i=1}^{m} n_i(\ln n_i - 1) = n \ln n - \sum_{i=1}^{m} n_i \ln n_i = \\ &= -n \left[-\ln n + \sum_{i=1}^{m} \frac{n_i}{n} (\ln \frac{n_i}{n} + \ln n) \right] = \\ &= -n \left[-\ln n + \sum_{i=1}^{m} \frac{n_i}{n} \ln \frac{n_i}{n} + \ln n \sum_{i=1}^{m} \frac{n_i}{n} \right] = \\ &= -n \sum_{i=1}^{m} \frac{n_i}{n} \ln \frac{n_i}{n}. \end{aligned} \tag{1.7}$$

Переходя к вероятностям, получим учитывающую базовое условие **3** ***формулу Шеннона*** для количества информации:

$$I = -n \sum_{i=1}^{m} P_i \log P_i. \tag{1.8}$$

Из указанной формулы возможно получить ***энтропию*** – *среднее количество*
информации на 1 бит информационного сообщения от указанного дискретного источника:

$$H = -\sum_{i=1}^{m} P_i \log P_i. \tag{1.9}$$

В дальнейшем в выражениях количества информации I и энтропии H по умолчанию будут пониматься логарифмы с основанием 2, если не оговорено обратного.

1.2.4 Информационная и физическая энтропия

Предложенная мера среднего количества информации была названа Шенноном энтропией отнюдь не случайно. По легенде, родоначальник компьютерных вычислительных систем фон Нейман, изучая рукопись "Математическая теория связи"сделал замечание, что указанная величина в точности повторяет выражение для определения энтропии физической системы, определенной ранее Больцманом.

Действительно, согласно второму закону термодинамики энтропия H (мера неупорядоченности) замкнутого пространства определяется выражением:

$$H = -\frac{1}{M_n} \sum_{i=1}^{N} m_i \ln \frac{m_i}{M_n}, \tag{1.10}$$

где M_n — число молекул в данном пространстве; m_i - число молекул, обладающих скоростью $\upsilon + \Delta\upsilon$.

Выражение 1.10 может быть также приведено к вероятностной нотации, т.к. m_i/M_n есть вероятность того, что молекула имеет скорость $\upsilon + \Delta\upsilon$. Таким образом, выражение для определения физической энтропии записывается в аналогичной 1.9 форме:

$$H = -\sum_{i=1}^{m} P_i \log P_i. \tag{1.11}$$

Безусловно, указанное совпадение имеет глубокий физический смысл, так как в обоих случаях величина энтропии характеризует степень разнообразия состояний системы.

Парадокс Демона Максвелла: Требуется отметить, что не только традиционная физика послужила примером для подражания для кибернетики. Так, именно с помощью теории информации был разрешен так называемый парадокс Демона Максвелла[3] *Рассматривается два сосуда с разными температурами, соединённые узкой трубкой с затворками, которыми может управлять воображаемый «демон». «Демон» может измерять скорость отдельных летящих молекул, и таким образом избирательно пропускать более быстрые в сосуд с высокой температурой,*

[3]Сформулирован Джеймсом Клерком Максвеллом в 1867 году для демонстрации кажущейся противоречивости второго начала термодинамики

а более медленные — в сосуд с низкой. Из этого мысленного эксперимента вытекает кажущееся противоречие со вторым началом термодинамики - тем, что ***энтропия изолированной системы не может уменьшаться****.*

Парадокс был разрешен при помощи теории информации. Для измерения скорости молекулы «демон» должен получить информацию о её скорости. Но всякое получение информации — материальный процесс, сопровождающийся возрастанием энтропии. Количественный анализ[4] *показал, что приращение энтропии при измерении превосходит по абсолютной величине уменьшение энтропии, вызванное перераспределением молекул «демоном», что полностью разрешает парадокс.*

1.2.5 Семантическая информация

В заключении упомянем о приложении теории информации в лингвистике - определении так называемой семантической информации. Несмотря на то, что сам Шеннон однажды заметил, что смысл сообщений не имеет никакого отношения к теории информации, способ измерения количества информации был применен и для оценки содержательности предложений естественного языка.

Одной из наиболее распространенных мер семантической информации является функция $\inf(s) = -\log_2 p(s)$, где s – это предложение, смысловое содержание которого измеряется, $p(s)$ – вероятность истинности s. Приведем несколько свойств этой функции-меры:

1 если $s_1 \Rightarrow s_2$ (из s_1 следует s_2) – истинно, то $\inf(s_1) \geq \inf(s_2)$;

2 $\inf(s) \geq 0$;

3 если s – истинно, то $inf(s) = 0$;

4 $\inf(s_1 s_2) = \inf(s_1) + \inf(s_2) \Leftrightarrow p(s_1 \cdot s_2) = p(s_1)p(s_2)$, т.е. независимость выражений s_1 и s_2.

Значение этой функция-меры больше для предложений, исключающих большее количество возможностей: например, из s_1 — «$a > 3$» и s_2 — «$a = 7$» следует, что $s_1 \Rightarrow s_2$ или $\inf(s_1) \geq \inf(s_2)$; ясно, что s_2 исключает больше возможностей, чем s_1.

[4]Проведенный в основополагающем труде Энрико Ферми "Термодинамика"

Еще одной достаточно используемой функцией-мерой семантической информации является функция $cont(s) = 1 - p(s)$. Ясно, что $cont(s) = 1 - 2^{-\inf(s)}$ или $\inf(s) = -\log_2(1 - cont(s))$.

Лекция 2

Элементы комбинаторики и теории вероятностей

2.1 Комбинаторика. Разделы комбинаторики

Комбинаторика *– раздел математики, в котором изучаются задачи выбора элементов из заданного множества и расположения их в группы по заданным правилам.*

Наиболее часто используется аппарат *перечислительной* комбинаторики - для решения задач о подсчете числа комбинаций (выборок), получаемых из элементов заданного конечного множества. В каждой задаче такого рода необходимо ответить на вопрос «сколькими способами» возможно осуществить что-либо; подсчитать варианты осуществления некоторого действия.

2.1.1 Базовые правила комбинаторики

Все задачи комбинаторики[1] могут быть решены с помощью трех основных правил, рассмотренных ниже, при этом из указанных правил выводятся все основные формулы комбинаторики.

Правило суммы: *Если некоторый объект A может быть выбран из совокупности объектов m способами, а другой объект B – n способами, то выбрать **либо** A, либо B возможно $m + n$ способами.*

[1]Здесь и далее имеются ввиду задачи перечислительной комбинаторики, если не указано обратное.

Правило произведения: *Если некоторый объект A может быть выбран из совокупности объектов m способами и после каждого такого выбора объект B можно выбрать n способами, то пара объектов (A, B) в указанном порядке может быть выбрана $m \cdot n$ способами.*

Пример 1. Повседневный наряд девушки состоит из блузки, юбки и туфель, а вечерний - платья, шали и туфли. В гардеробе лежит 4 блузки, 5 юбок и 3 туфель; 2 платья, 1 шаль. Внимание вопрос: сколько видов повседневных нарядов может одеть девушка? Сколько всего нарядов у девушки, если вечернее платье можно одеть без шали?

Решение. По принципу умножения получаем, что видов повседневных нарядов у девушки $4 \cdot 5 \cdot 6 = 60$. Аналогично, видов вечерних нарядов - $2 \cdot 2 \cdot 3 = 12$ (вечернее платье можно одеть без шали). Таким образом, по правилу сложения получаем, что общее число нарядов у девушки равно $60 + 12 = 72$.

Формула включения и исключения:
Пусть имеется K частично пересекающихся множеств объектов; k-е множество объединяет объекты, обладающие определенным свойством a_k и имеет размерность N_k. Для решения задачи определения общего количества объектов N_{any}, обладающих хотя бы одним из свойств $\overline{a_1 \dots a_n}$ используется формула включения и исключения, определяющая де-факто мощность объединения указанных множеств.

$$N_{any} = N_1 + N_2 + \ldots + N_K - N_{a_1+a_2} - \ldots - N_{a_1+a_n} - \ldots - N_{a_{n-1}+a_n} + \\ + N_{a_1+a_2+a_3} + \ldots + N_{a_{n-2}+a_{n-1}+a_n} + \ldots + (-1)^{n-1} N_{a_1+a_2+\ldots+a_n}. \tag{2.1}$$

Пример 2. Студенты факультета ВМК К(П)ФУ им. Ульянова-Ленина делятся на знающих языки программирования C++ (30 человек из выпуска), Pascal (20 человек), одновременно два языка (также 20 человек). Какое количество выпускников ВМК знает в достаточной мере хотя бы один язык программирования из вышеперечисленных?

Решение. Согласно формуле включения и исключения указанное количество выпускников $N = 30 + 20 - 20 = 30$.

2.1.2 Основные формулы комбинаторики

Из рассмотренных правил выводятся основные формулы перечислительной комбинаторики.

Размещения

Итак, рассмотрим некоторое множество X, состоящее из n элементов $X = x_1, x_2, \ldots, x_n$. Будем выбирать из этого множества различные упорядоченные подмножества Y из k элементов. В этом случае **размещением** из n элементов множества X по k элементам называется любой упорядоченный набор $(x_{i_1}, x_{i_2}, \ldots, x_{i_k})$ элементов множества X. При этом, если выбор элементов множества Y из X происходит с возвращением (т.е. каждый элемент множества X может быть выбран несколько раз), то число размещений обозначается как $\overline{A_n^k}$ (количество *размещений с повторениями*) и определяется по следующей формуле:

$$\overline{A_n^k} = n^k. \tag{2.2}$$

Если же выбор делается без возвращения, т.е. каждый элемент множества X можно выбрать только один раз, то количество размещений из n по k обозначается A_n^k (количество *размещений без повторений*) и определяется равенством:

$$A_n^k = \frac{n!}{(n-k)!}. \tag{2.3}$$

Пример 3. Даны шесть цифр: 1,2,3,4,5,6. Определить: сколько трехзначных чисел возможно из них составить?

Решение. Если цифры могут повторяться, то количество трехзначных чисел будет равно $\overline{A_6^3} = 6^3 = 216$. Если не могут - то $A_6^3 = \frac{6!}{3!} = 6 \cdot 5 \cdot 4 = 120$.

Перестановки

Частный случай размещения при $n = k$ называется **перестановкой** из n Элементов. Число всех перестановок из n элементов равно:

$$A_n^n = P_n = n! \tag{2.4}$$

Пример 4. 30 книг стоят на книжной полке, из них 27 различных авторов и три книги от одного автора. Сколькими способами возможно расставить эти книги на полке так, чтобы книги одного автора стояли рядом?

Решение. Будем считать три книги одного автора за одну книгу; в этом случае число перестановок будет определяться как P_{28}. При этом три книги возможно переставить между собой P_3 способами. Таким образом, по правилу произведения получаем искомое число способов, равное $P_{28} \cdot P_3 = 28! \cdot 3!$.

Существует также понятие так называемых **перестановок с повторениями**. Такими перестановками из n_1 элементов первого типа, n_2 элементов второго типа, ..., n_k элементов k-го типа называются всевозможные комбинации из этих элементов, каждая из которых содержит n_i элементов i-го вида. Комбинации отличаются друг от друга лишь порядком элементов.

Число перестановок с повторениями обозначается как $P(n_1, n_2, \ldots, n_k)$ и определяется по формуле:

$$P(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_k!}. \tag{2.5}$$

Пример 5. Сколькими способами можно поставить в ряд 3 красных, 4 синих и 5 зеленых кубиков?

Решение. По формуле перестановок с повторениями получаем:
$P(3, 4, 5) = \frac{12!}{3! \cdot 4! \cdot 5!} = 27720$.

Сочетания

Пусть теперь из множества X выбирается неупорядоченное подмножество Y (порядок элементов в котором не важен). **Сочетаниями** из n элементов по k называются подмножества из k элементов, отличающиеся друг от друга хотя бы одним элементом. Общее число всех сочетаний обозначается C_n^k и равно:

$$C_n^k = \frac{A_n^k}{k!} = \frac{n!}{(n-k)!k!} = \frac{n(n-1)\ldots(n-k+1)}{k!}. \tag{2.6}$$

Для сочетаний справедливы следующие равенства: $C_n^0 = 1, C_n^n = 1, C_n^k = C_n^{n-k}$.

Пример 6. Сколькими способами можно выбрать три краски из имеющихся пяти?

Решение. В поставленной задаче порядок выбора красок не важен. Таким образом, количество способов определяется, как количество сочетаний $C_5^3 = \frac{5!}{2!3!} = 10$.

В некоторых случаях необходимо вычислить так называемые **сочетания с повторениями**. Сочетания с повторениями из n элементов по k - это всевозможные комбинации, составленные из элементов n видов по k элементов в каждой. Комбинации считаются различными, если они отличаются составом, но не порядком входящих в них элементов. В комбинацию могут входить элементы одного вида.

Количество сочетаний с повторениями обозначается $\overline{C_n^k}$ и равно:

$$\overline{C_n^k} = C_{n+k-1}^k = \frac{(n+k-1)!}{(n-1)!k!}. \tag{2.7}$$

Пример 7. В кондитерском магазине продаются 4 сорта пирожных: наполеоны, эклеры, песочные и слоеные. Сколькими способами можно купить 7 пирожных?

Решение. Согласно формуле сочетаний с повторениями указанное количество равно $\overline{C_4^7} = \frac{10!}{3!7!} = \frac{8 \cdot 9 \cdot 10}{6} = 120$.

2.1.3 Теоремы Рамсея и Ван-дер-Вардена

В заключение раздела, посвященного основам комбинаторики, невозможно не упомянуть о замечательной комбинаторной теореме Рамсея, изучающей наличие регулярных структур в случайных конфигурациях элементов. Так, примером утверждения, следующего из теоремы Рамсея может служить следующее: *в группе из 6 человек всегда можно найти*

трех человек, которые либо попарно знакомы друг с другом, либо попарно незнакомы. Теорема Рамсея сформулирована в терминах теории графов, но имеет и аналогичную формулировку для натуральных чисел. Указанная формулировка и называется теоремой Ван-дер-Вардена, и определяется следующим образом:

Теорема Ван-дер-Вардена
Для всякого набора чисел $a_1, a_2, ldots, a_n$ существует такое число W, что, как бы мы не покрасили первые W натуральных чисел в n цветов, найдется либо *арифметическая прогрессия* 1-го цвета длины a_1, либо арифметическая прогрессия 2-го цвета длины a_2,..., либо арифметическая прогрессия n-го цвета длины a_n.

Указанные теоремы имеют два характерных свойства:

- Во-первых, они не конструктивны. Доказывается, что некоторая структура существует, но не предлагается никакого способа ее построения, кроме прямого перебора.

- Во-вторых, чтобы искомые структуры существовали, требуется, чтобы объекты их содержащие, состояли из очень большого количества элементов - зависимость числа элементов объекта от размеры структуры как минимум экспоненциальна.

Указанные свойства, с точки зрения автора лекций, носят достаточно философский характер и во многом отражает несколько романтичный дух теории информации в целом.

2.2 Элементы теории вероятности

2.2.1 Базовые понятия теории вероятности

Теория вероятности - область математического знания, изучающая закономерности, возникающие при рассмотрении массовых однотипных случайных событий. Центральным понятием теории вероятности является понятие *случайного события.*

Случайным событием называется событие, которое при осуществлении некоторых условий может произойти или не произойти. Пример случайного события - попадание в некоторый объект или промах при стрельбе по объекту.

Достоверным называется событие, если в результате испытания оно обязательно происходит.

Невозможное - такое событие, которое не может произойти в результате данного испытания.

Несовместными называются такие случайные события, для которых одновременное появление никаких двух из них в рамках данного испытания невозможно.

Независимыми являются такие события, появление одного которых не меняют вероятности появления других.

Зависимыми называются такие события, вероятность которых меняется в зависимости от появления других событий, входящих в эту группу. ***Полная группа*** - множество событий, из которых в результате данного испытания обязательно появится произвольное, но при этом только одно событие.

Исходом называются события, входящие в полную группу равновозможных несовместных случайных событий. Исход называется ***благоприятствующим*** появлению события A, если появление этого исхода влечет за собой появление события A.

Пример 8. В урне находится 8 пронумерованных (от 1 до 8) шаров. Шары с цифрами 1,2 и 3 – красные; остальные – черные. Каким является событие появление шара с номером 4?

Решение. Появление шара с цифрой 4 есть событие (исход), благоприятствующее появлению черного шара.

По результату рассмотрения элементарных терминов теории вероятности, дадим далее классическое определение вероятности:

Вероятностью события A называют отношение числа m благоприятствующих этому событию исходов к общему числу n всех равновозможных несовместных элементарных исходов, образующих полную группу:

$$P(A) = \frac{m}{n}. \tag{2.8}$$

Вероятность в классическом определении обладает следующими элементарными свойствами:

- textbfСвойство 1: Вероятность достоверного события равна 1.

- textbfСвойство 2: Вероятность невозможного события равна 0.
- textbfСвойство 3: Вероятность случайного события A удовлетворяет двойному неравенству $0 \leq P(A) \leq 1$.

2.2.2 Сложение и умножение вероятностей

Событие A называется ***частным случаем*** события B, если при наступлении A наступает и B. То, что A является частным случаем B, записывается $A \subset B$.

События A и B называются ***равными***, если каждое из них является частным случаем другого. Равенство событий A и B записывается как $A = B$.

Суммой событий A и B называется событие $A + B$, наступающее тогда и только тогда, когда наступает хотя бы одно из событий A или B.

Теорема о сложении несовместных событий: *Вероятность появления одного из двух (одного из группы) несовместных событий равна сумме вероятностей этих событий:*

$$\begin{aligned} P(A+B) &= P(A) + P(B) \\ P\left(\sum_{i=1}^{n} A_i\right) &= \sum_{i=1}^{n} P(A_i). \end{aligned} \tag{2.9}$$

Если случайные события $A_1, A_2, \ldots, A_n$ образуют полную группу несовместных событий, то имеет место равенство:

$$P(A_1) + P(A_2) + \ldots + P(A_n) = 1. \tag{2.10}$$

Теорема о сложении совместных событий: *Вероятность суммы совместных событий вычисляется по формуле:*

$$P(A+B) = P(A) + P(B) - P(AB). \tag{2.11}$$

Теорема о умножении вероятностей независимых событий: *Вероятность произведения независимых событий A и B равна:*

$$P(AB) = P(A) \cdot P(B). \tag{2.12}$$

Из данных теорем вытекает следующая:

Вероятность появления хотя бы одного из событий $A_1, A_2, \ldots A_n$, независимых в совокупности, равна разности между единицей и произведением вероятностей противоположных событий ($P(\overline{A}) = 1 - P(A)$ – событие, противоположной событию A):

$$P(A) = 1 - P(\overline{A_1}) \cdot P(\overline{A_2}) \cdot \ldots P(\overline{A_n}). \tag{2.13}$$

Если события $A_1, A_2, \ldots A_n$ имеют одинаковую вероятность реализации p (соответственно, противоположные события - вероятность реализации $1 - p = q$, то формула 2.13 приобретает следующий вид:

$$P(A) = 1 - (1 - p)^n = 1 - q^n. \tag{2.14}$$

2.2.3 Условная вероятность, полная вероятность события

Рассмотрим далее понятия условной вероятности и полной вероятности определенного события.

Итак, случайное событие - это такое событие, которое при осуществлении совокупности условий эксперимента может произойти или не произойти. Если при вычислении вероятности события никаких других условий, кроме условий эксперимента, не налагается, то такая вероятность называется ***безусловной***; если же налагаются и другие дополнительные условия, то вероятность события называют ***условной***. Так, что вычисляют вероятность возникновения события B при дополнительном условии, что произошло событие A.

Условной вероятностью $P_A(B) = P(B|A)$ называют вероятность события B, вычисленную в предположении, что событие A уже наступило. Вероятность ***совместного появления двух зависимых событий*** равна произведению вероятности одного из них на условную вероятность второго, вычисленную при условии, что первое событие произошло. Таким образом:

$$P(AB) = P(B)P(A|B) = P(A)P(B|A); P(A|B) = \frac{P(AB)}{P(B)}. \tag{2.15}$$

Определим далее так называемую полную вероятность события. Итак, если событие A может произойти только при выполнении одного из событий $B_1, B_2, \ldots, B_n$, образующих *полную группу несовместных событий*,

то вероятность события A вычисляется по так называемой ***формуле полной вероятности***:

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \ldots + P(B_n)P(A|B_n). \quad (2.16)$$

2.2.4 Формула Байеса

Закончим лекцию на рассмотрении так называемой формулы Байеса[2], являющейся основополагающей в теории вероятностей, математической статистике и множества областей обработки информации.

Вновь рассмотрим полную группу несовместных событий $B_1, B_2, \ldots, B_n$, вероятности появления которых $P(B_1), P(B_2), \ldots, P(B_n)$. Событие A может произойти только вместе с каким-либо из событий $B_1, B_2, \ldots, B_n$, которые будем называть ***гипотезами***. Тогда по формуле полной вероятности получаем:

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \ldots + P(B_n)P(A|B_n).$$

Если событие A произошло, то это может изменить вероятности гипотез $P(B_1), P(B_2), \ldots, P(B_n)$. Далее, по теореме умножения вероятностей:

$$P(AB_1) = P(B_1)P(A|B_1) = P(A)P(B|A),$$

откуда

$$P(B_1|A) = \frac{P(B_1)P(A|B_1)}{P(A)}.$$

Формулой Байеса называется аналогичное выражение, обобщенное для всех гипотез:

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{P(A)}, i = 1, \ldots, n. \quad (2.17)$$

В формуле Байеса вероятности гипотез $P(B_i|A)$ называются ***апостериорными вероятностями***[3], тогда как $P(B)$ - ***априорными вероятностями***[4].

В общем случае Байесовская теория представляет собой метод адаптации существующих вероятностей к вновь полученным экспериментальным данным.

---

[2]В честь Томаса Байеса (1702–1761), доказавшего частный случай приведенной ниже теоремы. В реальности, более общий случай теоремы Байеса был доказан Лапласом, не считавшим, однако, ее важной для развития теории вероятности.

[3]a-*posteriori* (лат. - после опыта).

[4]a-*priori* (лат. - до опыта).

На сегодняшний день Байесовы методы и Байесов анализ широко применяется в теории информации, при построении интеллектуальных семантических фильтров (например спам–фильтров), в алгоритмах анализа массивов данных, в предсказании состояний различных динамических процессов, а также во множестве алгоритмов цифрового приема и обработки сигналов.

Практика 1-2 – комбинаторика, количество информации

2.3 Элементарная комбинаторика

2.3.1 Комбинаторные формулы-задачи с решениями

Пример 9. Сколькими способами возможно расставить на полке 5 книг?

Решение. Искомое количество способов равно числу перестановок из 5 элементов (книг):

$$P_5 = 5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120.$$

Пример 10. Сколько "слов"по две буквы возможно составить из букв a,b,c,d,e таким образом, чтобы буквы в этих "словах"не повторялись?

Решение. Так как каждое "слово"должно содержать две буквы, то итоговое количество способов равно числу размещений из 5 элементов (букв) по две, т.е.:

$$A_5^2 = \frac{5!}{(5-2)!} = \frac{5!}{3!} = 20.$$

Пример 11. Сколькими способами возможно выбрать 1 красную гвоздику и 2 розовых из вазы, в которой стоят 10 красных и 4 розовых гвоздики?

Решение. Так как порядок выбора цветов не имеет значения, то красную гвоздику можно выбрать

$$C_1 0^1 = \frac{10!}{1! \cdot (10-1)!} = \frac{10!}{9!} = 10$$

способами. Выбрать две розовые гвоздики из имеющихся четырех возможно

$$C_4^2 = \frac{4!}{2! \cdot 2!} = \frac{4 \cdot 3}{1 \cdot 2} = 6$$

способами. Поэтому букет из одной красной и двух розовых гвоздик можно составить по правилу умножения $C_1 0^1 \cdot C_4^2 = 10 \cdot 6 = 60$ способами.

Пример 12. Набирая номер телефона, абонент забыл последние 3 цифры и помня лишь, что эти цифры различны, набрал их наугад. Найти вероятность того, что номер телефона набран правильно.

Решение. Благоприятствующий исход здесь один - правильный выбор последних цифр ($m = 1$). Всех возможных исходов n здесь столько же, сколько существует комбинаций из 3-х цифр, порядок которых имеет значение. Таким образом,

$$n = A_1 0^3 = 720$$

и вероятность целевого события A (того, что номер набран правильно):

$$P(A) = \frac{m}{n} = \frac{1}{720}.$$

Пример 13. Среди 100 колес 5 нестандартных. Для контроля выбирается 7 колес. Найти вероятность того, что среди них будет ровно 3 нестандартных.

Решение. Число всех возможных исходов равно количеству комбинаций из 100 колес по 7 штук, т.к. порядок значения не имеет, то $n = C_1 00^7$. Благоприятствующий исход состоит в выборе ровно 3 нестандартных колес из 5 и совместном выборе (7-3) стандартных колес из (100-5), при этом порядок значения не имеет. По правилу произведения

$$m = C_5^3 \cdot C_9 5^4$$

. Следовательно, вероятность того, что среди взятых для контроля колес будет ровно 3 нестандартных (целевое событие A):

$$P(A) = \frac{m}{n} = \frac{C_5^3 \cdot C_9 5^4}{C_1 00^7} = \frac{17967600}{90345024} = 0.199.$$

2.3.2 Комбинаторные формулы-задачи

1 Сколько четных положительных чисел можно составить из цифр числа 13754, если каждую цифру можно использовать в записи не более 1 раза?

2 Сколькими способами можно составить трехцветный флаг, если имеется материал пяти различных цветов?

3 Необходимо доставить рекламные проспекты в 6 различных фирм. Сколькими способами это могут сделать трое курьеров?

4 В коробке 48 шариковых ручек и 3 гелевых. Наудачу извлекают одну ручку и, не возвращая ее обратно, извлекают еще одну. Какова вероятность того, что последняя рука шариковая, если первая извлеченная ручка - гелевая?

5 В группе 25 студентов, среди них 5 отличников. Выбирают по списку 10 студентов. Найти вероятность того, что среди них окажется 3 отличника.

2.4 Количество информации дискретного источника

2.4.1 Количество информации-задачи с решениями

Пример 14. Какая информация содержится в сообщении о том, что монетка упала гербом?

Решение. Вероятность того, что монетка упала гербом, равна $p = \frac{1}{2}$. Поэтому проведение данного испытания дало нам:

$$I = -\log p = -\log \frac{1}{2} = \log 2 = 1bit,$$

т.е. один бит информации.

Пример 15. В ящике 10 гранат, из которых 8 без взрывателя. Из ящика наудачу выбираются 3 гранаты. Какое количество информации содержится в сообщении о том, что все выбранные гранаты оказались без взрывателя?

Решение. Определим вероятность выбора из ящика 3-х гранат без взрывателя:

$$p = \frac{8}{10} \cdot \frac{7}{9} \cdot \frac{6}{8} = \frac{7}{15} = 0.467.$$

Количество информации в данном сообщении определяется по формуле:

$$I = -\log p = -\log \frac{7}{15} = 1.1bit$$

.

Пример 16. В группе 20 курсантов, среди которых 4 отличника, 6 хорошистов, 7 троечников, остальные двоечники. По списку наудачу отбираются 5 курсантов. Какое количество информации содержится в сообщении о том, что среди отобранных курсантов 3 отличника, 1 хорошист и 1 троечник?

Решение. Используем понятие т.н. обобщенной гипергеометрической вероятности, используемое, когда интересующие объекты не возвращаются в выборку. Достаточно очевидно, что общее количество исходов определяется, как $C_N^M = C_2 0^5$, в то же время, число благоприятных исходов по комбинаторному правилу произведения определяется, как:

$$C_{n1}^{m1} \cdot C_{n2}^{m2} \cdot C_{n3}^{m3} \cdot C_{n4}^{m4} = C_4^3 \cdot C_6^1 \cdot C_7^1 \cdot C_3^0.$$

Итоговое значение вероятности совпадает с формулой гипергеометрического распределения и равно:

$$P = \frac{C_4^3 \cdot C_6^1 \cdot C_7^1 \cdot C_3^0}{C_2 0^5} = \frac{7}{17 \cdot 19 \cdot 2 = 0.011.}$$

Исходя из этого, и получаем искомое количество информации:

$$I = -\log P = -\log 0.011 = 6.528 bit.$$

2.4.2 Количество информации-задачи

1 В отделе работают 6 мужчин и 4 женщины. По табельным номерам наудачу отобраны 7 человек. Какое количество информации содержится в сообщении о том, что среди отобранных лиц окажутся 3 женщины?

2 В лифт 7-этажного дома сели 3 пассажира. Каждый независимо от других с одинаковой вероятностью может выйти на любом (начиная со второго) этаже. Какое количество информации содержится в сообщении о том, что все вышли на разных этажах?

3 Имеются изделия четырех сортов, причем число изделий каждого сорта равно $n_1 = 2, n_2 = 3, n_3 = 1, n_4 = 3$. Для контроля наудачу берутся 5 изделий. Какое количество информации содержится в сообщении о том, что среди них $m_1 = 2$ - первосортных; $m_2 = 1$ - второго, $m_3 = 0$ - третьего и $m_4 = 2$ - четвертого сорта?

4 Таможенный контроль проходят 3 таджика, 4 грузина и 6 азербайджанцев. На досмотр наудачу вызывают трех человек. Какое количество информации содержится в сообщении о том, что все три вызванных являются грузинами?

Лекция 3

Свойства энтропии. Взаимная информация. Непрерывные случайные величины

3.1 Энтропия

3.1.1 Свойства дискретной энтропии

Энтропия - это важнейшее понятие в теории информации. Рассмотрим далее свойства энтропии и доказательства их корректности:

Свойство 1. Энтропия – вещественная и неотрицательная величина: $H(U) \geq 0$.

Доказательство.
По определению, энтропия $H(U)$ сообщения U определяется по формуле

$$H = -\sum_{i=1}^{m} P_i \log P_i.$$

Каждое слагаемое в данной сумме является неотрицательным. Действительно:

1. $\forall i \quad 1 \geq i \geq Np_i \in [0; 1]$.

2. Для указанных значений $p \log p \geq 0 \Rightarrow -p_i \log p_i \geqslant 0$, что и требовалось доказать.

3. $H(U) = 0$ тогда и только тогда, когда какой-то k-й из исходов является достоверным ($p_k = 1$, соответственно $\forall p_i, \overline{i \neq k}$). Данное утверждение доказывается от противного – $\forall p_i, 0 < p_i < 1$ все $-p_i \log p_i > 0$, что и требовалось доказать.

Свойство 2. Для произвольного источника информации значение энтропии удовлетворяет неравенству $H(U) \geq \log N$, где N - объем алфавита источника, при этом $\max H(U) = \log N$ для случая n равновероятных исходов: $p_i = \frac{1}{n}, i = \overline{1 \dots n}$.

Доказательство.

1. Рассмотрим разность $H(U) - \log N$:

$$H(U) - \log N = -\sum_{i=1}^{n} (P(U_i) \cdot \log P(U_i)) - \log N \cdot \underbrace{\sum_{i=1}^{n} P(U_i)}_{=1} = \tag{3.1}$$
$$-\sum_{i=1}^{n} P(U_i) \left[\log P(U_i) + \log N\right] = -\sum_{i=1}^{n} P(U_i) \cdot \log \left(N P(U_i)\right).$$

2. Воспользуемся далее неравенством:

$$\frac{\log x}{\log e} = \ln x \leq x - 1 (\ln x = x + 1 \text{при} x = 1). \tag{3.2}$$

Используя его в выражении 3.1, получаем:

$$H(U) - \log N = -\log e \cdot \sum_{i=1}^{n} P(U_i) \cdot \ln N \cdot P(U_i) \leq$$
$$\leq -\log e \cdot \sum_{i=1}^{n} P(U_i) \left[N \cdot P(U_i) + 1\right] = -\log e \cdot \sum_{i=1}^{n} \left[N + P(U_i)\right] = \tag{3.3}$$
$$= \log e \cdot (1 - 1) = 0 \Longrightarrow H(U) \leq \log N,$$

что и требовалось доказать. Из этого же равенства, а также из 3.2 следует, что $H(U) - \log N = 0$ тогда, когда $\frac{1}{N \cdot P(U)} = 1$, т.е. когда $P(U) = \frac{1}{N}$, что соответствует случаю равновероятных исходов.

Для рассмотрения следующего свойства энтропии введем понятие объединения источников сообщений (без ограничения общности - для случая двух источников сообщений):

Объединением источников сообщений *U и Z с объемами алфавита N и M соответственно понимают обобщенный источник сообщений U, Z, характеризующийся совместными вероятностями $P(U, Z_j)$ всех возможных комбинаций, выбираемых из алфавита размерностью $N \cdot M$*

Свойство 3. Свойство аддитивности энтропии:
Энтропия объединения нескольких независимых источников сообщений равна сумме их исходных значений энтропии:

$$H(U_1, U_2, \ldots, U_K) = \sum_{k=1}^{K} H(U_k).$$

Доказательство.

1. Не теряя общности, ограничимся рассмотрением объединения двух источников U и Z с объемами алфавита, соответственно, N и M. Энтропия такого объединенного источника определяется по формуле:

$$H(U, Z) = -\sum_{i=1}^{N}\sum_{j=1}^{M} P(U_i, Z_j) \cdot \log P(U_i, Z_j). \tag{3.4}$$

2. Для случая статистической независимости U и Z $P(U, Z_j)$, тогда:

$$\begin{aligned} H(U, Z) &= -\sum_{i=1}^{N}\sum_{j=1}^{M} P(U) \cdot P(Z_j) \cdot \log P(U_i)P(Z_j) = \\ &= -\underbrace{\sum_{j=1}^{M} P(Z_j)}_{1} \cdot \underbrace{\sum_{i=1}^{N} P(U) \cdot \log P(U)}_{H(U)} - \underbrace{\sum_{i=1}^{N} P(U)}_{1} \cdot \underbrace{\sum_{j=1}^{M} P(Z_j) \cdot \log P(Z_j)}_{H(Z)} = \\ &= H(U) + H(Z), \text{ что и требовалось доказать.} \end{aligned} \tag{3.5}$$

3.1.2 Условная энтропия и взаимная информация

Часто необходимо определить энтропию сложного опыта U, Z для случая, когда U и Z не являются независимыми - т.е. если на исход источника U влияет результат опыта Z. Элементарным примером является случай с нахождением в некотором ящике всего двух разноцветных шаров, когда результат U состоит в извлечении первого шара, а Z - второго. Для этого случая $H(U, Z) = H(U)$, а не $H(U) + H(Z)$ согласно озвученному выше правилу аддитивности.

Итак, пусть источники U и Z являются зависимыми. Тогда, согласно 2.15:

$$P(U,Z) = P(U) \cdot P(Z|U); \quad \log P(U,Z) = \log P(U) + \log P(Z|U). \quad (3.6)$$

Используя 3.4, получаем следующий вывод для энтропии описанного сложного опыта:

$$\begin{aligned}
H(U,Z) &= -\sum_{i=1}^{N}\sum_{j=1}^{M} P(U_i,Z_j)\cdot \log P(U_i,Z_j) = \\
&= -\sum_{i=1}^{N}\sum_{j=1}^{M} P(U_i)P(Z_j|U_i)\cdot (\log P(U_i) + \log P(Z_j|U_i)) = \\
&= -\sum_{i=1}^{N}\left(P(U_i)\cdot \log P(U_i) \underbrace{\sum_{j=1}^{M} P(Z_j|U_i)}_{1}\right) - \\
&- \sum_{i=1}^{N}\left(P(U_i)\left(\sum_{j=1}^{M} P(Z_j|U_i)\cdot \log P(Z_j|U_i)\right)\right) = \\
&= H(U) + \sum_{i=1}^{N} P(U_i)\cdot H(Z|U_i) = H(U) + H_U(Z).
\end{aligned} \quad (3.7)$$

В указанном выражении

$$H(Z|U_i) = -\sum_{j=1}^{M} P(Z_j|U_i)\cdot \log P(Z_j|U_i) - \quad (3.8)$$

так называемая ***частная условная энтропия*** источника Z с учетом реализации исхода U_i;

$$\begin{aligned}
H_U(Z) &= \sum_{i=1}^{N} P(U_i)\cdot H(Z|U_i) = \\
&- \sum_{i=1}^{N}\left(P(U_i)\sum_{j=1}^{M} P(Z_j|U_i)\cdot \log P(Z_j|U_i)\right) -
\end{aligned} \quad (3.9)$$

полная условная энтропия или просто условная энтропия источника Z *по отношению к* источника U.

Исходя из определения, условная энтропия источника Z по отношению к источнику U представляет собой среднее количество информации, даваемое сообщением источника U при условии, что сообщение источника Z уже известно.

Для условной энтропии справедливо следующее неравенство:

$$0 \leq H_U(Z) \leq H(Z), \tag{3.10}$$

при этом $H_U(Z) = 0$, когда по сообщению источника U возможно точно определить сообщение источника Z[1] и $H_U(Z) = H(Z)$, когда источники U и Z независимы и знание реализации U ничего не говорит о реализации Z.

Итак, в общем случае $H_U(Z) < H(Z)$ и знание реализации U снижает первоначальную неопределенность Z. На основании сделанного вывода возможно ввести специальную информационную характеристику двух зависимых источников U и Z - так называемую *взаимную информацию.*

Взаимная информация *- количество информации, содержащееся в U относительно Z* и равна:

$$I(Z,U) = H(Z) - H_U(Z). \tag{3.11}$$

Взаимная информация определяется в тех же единицах, что и энтропия[2]. Величина $I(Z,U)$ показыает, сколько в среднем бит информации о реализации сообщения источника Z дает наблюдение о реализации сообщения источника U.

Определим $I(Z,U)$ с использованием выражений 3.7, 3.9 и 3.11, а также учитывая, что $P(Z|U) = \frac{P(Z,U)}{P(U)}$ и $P(Z_j) = \sum_{i=1}^{N} P(Z_j, U_i)$:

$$I(Z,U) = -\sum_{j=1}^{M} P(Z_j) \cdot \log(P(Z_j)) + \sum_{j=1}^{M}\sum_{i=1}^{N} \left(P(Z_j,U_i) \cdot \log \frac{P(Z_j,U_i)}{P(U_i)} \right) =$$

$$-\sum_{j=1}^{M}\sum_{i=1}^{N} (P(Z_j,U_i) \cdot \log P(Z_j)) + \sum_{j=1}^{M}\sum_{i=1}^{N} \left(P(Z_j,U_i) \cdot \log \frac{P(Z_j,U_i}{P(Z_j)} \right) =$$

$$\sum_{j=1}^{M}\sum_{i=1}^{N} \left(P(Z_j,U_i) \cdot \log \frac{P(Z_j,U_i)}{P(Z_j) \cdot P(U_i)} \right). \tag{3.12}$$

[1]Такой случай идентичен каналу связи без помех, когда Z - сообщение, переданное по каналу связи, а U - сообщение, принятое на приемной стороне

[2]В битах, детах и натах, см. лекцию 1.

3.1.3 Свойства взаимной информации

Взаимная информация обладает рядом замечательных свойств, перечисленных ниже:

1 Взаимная информация удовлетворяет неравенству:

$$I(Z,U) \geq 0, \tag{3.13}$$

причем равенство имеет место только в том случае, когда Z и U независимы между собой. Это следует из определения 3.11 и неравенства 3.10.

2 Свойство симметрии взаимной информации:

$$I(Z,U) = I(U,Z) \Longrightarrow I(U,Z) = H(U) - H(U|Z), \tag{3.14}$$

т.е. U содержит в себе столько же информации относительно Z, сколько и Z относительно U, это свойство вытекает из симметрии заключительного выражения 3.12 для количества информации.

3 Количество взаимной информации для двух источников всегда не больше энтропии любого из этих источников:

$$I(Z,U) \leq H(U); I(Z,U)leqH(Z), \tag{3.15}$$

при этом равенство имеет место, когда по реализации U возможно точно восстановить реализацию Z и наоборот. Это свойство следует из самого определения количества информации и из свойства симметрии.

4 Полагая в 3.11 $U = Z$ и учитывая, что $H(Z|Z) = 0$, получаем:

$$I(Z,Z) = H(Z) \tag{3.16}$$

Из этого выражения следует, что *энтропия источника является собственной информацей о самом себе.*

3.1.4 Преобразования информации

В заключении данного раздела лекции упомянем о обратимых и необратимых преобразованиях информации и влиянию характера преобразований на свойства информационных характеристик.

Итак, пусть Z – ансамбль дискретных сообщения, а U - ансамбль дискретных сигналов, в которые преобразуется сообщение Z, тогда $I(Z,U) = H(Z) \Longleftrightarrow$, если преобразование Z в U обратимо, т.е. однозначно. При необратимом преобразовании $I(Z,U) < H(Z)$ и разность $H(Z) - I(Z,U) = H(Z|U)$ называют ***потерей информации*** или ***ненадежностью преобразования*** Z в U.

Таким образом, информация не теряется только при обратимых преобразованиях; величина $H(U|Z) = H(U) - I(Z,U)$ называется ***энтропией шума преобразования*** или ***ложной информацией, создаваемой при образовании***.

3.2 Непрерывные случайные величины

3.2.1 Функция и плотность распределения вероятностей

Если случайная величина X может принимать конечное число дискретных значений x_i, исчерпывающей вероятностной характеристикой данной величины служит распределение вероятностей этих значений P_i. В том же случае, если случайная величина X непрерывна и может принимать любое значение на интервале $[x_min, x_max]$, ее статистической характеристикой служит так называемый **интегральный закон распределения** или **функция распределения вероятностей**:

$$F(x) = P(X < x) \Longrightarrow P(x_1 \leq X < x_2) = F(x_2) - F(x_1). \qquad (3.17)$$

Функция распределения $F(x)$ обладает следующими свойствами, вытекающими из ее определения:

1 $F(x)$ - монотонная неубывающая функция;

2 $F(-\infty) = 0; F(+\infty) = 1.$

Если функция $F(x)$ является дифференцируемой, то используют т.н. **дифференциальный закон распределения** или **закон распределения плотности вероятности**[3]:

$$\omega(x) = \frac{dF(x)}{dx} \qquad (3.18)$$

[3]Сокращенно практически всегда говорится просто «плотность распределения вероятностей»

Для плотности распределения $\omega(x)$ справедливы следующие соотношения:

1 $F(x) = \int_{-\infty}^{x} \omega x dx$;

2 $\omega(x) = \int_{-infty}^{+infty} dx = 1$;

3 $P(x_1 \leq X < x_2) = \int_{x_1}^{x_2} \omega x dx$.

3.2.2 Моменты распределения

Помимо законов распределения часто используются числовые характеристики случайных величин, называемые **моментами распределения**. Моменты, характеризующие распределение случайных величин относительно нуля, называются **начальными**. Для непрерывных случайных величин начальный момент *k-го порядка* определяется по следующей формуле:

$$m_k(X) = \int_{-\infty}^{+\infty} x^k \omega(x) dx, . \tag{3.19}$$

при этом предполагается, что интеграл 3.19 абсолютно сходится, т.е. выражение $\int_{-\infty}^{+\infty} |x|^k \omega(x) dx$ конечно.

Самым важным начальным моментом является момент первого порядка, называемый **математическим ожиданием** и представляющим собой среднее значение случайной величины[4]:

$$m(X) = mean(X) = \int_{-\infty}^{+\infty} x\omega(x) dx. \tag{3.20}$$

Разность $\triangle X = X - m(x)$ называется **отклонением** случайной величины. Моменты распределения отклонений случайной величины называются **центральными**, отображают разброс случайной величины относительно среднего значения и обозначаются $M_k(X)$.

Наиболее важным центральным моментом, определяющим амплитуду разброса относительно среднего значения является так называемая **дисперсия** - второй центральный момент:

$$D(X) = M_2(X) = \int_{-\infty}^{+\infty} (x - m(x))^2 \omega(x) dx. \tag{3.21}$$

[4]В английской литературе математическое ожидание случайной величины X записывается как $\mathbf{E}\{X\}$.

Корень из дисперсии имеет ту же размерность, что и случайная величина, называется **средним квадратическим отклонением** случайной величины и обозначается греческой буквой «сигма»:

$$\sigma(X) = \sqrt{D(X)}. \tag{3.22}$$

3.2.3 Нормальный закон распределения

Среди реальных физических процессов, представляющих собой дефакто случайные величины, большой класс подчиняется так называемому **нормальному (гауссовому) закону распределения**[5]:

$$\omega(x) = \frac{1}{\sigma(x)\sqrt{2\pi}} e^{\left(-\frac{(x-m(x))^2}{2\sigma^2}\right)}. \tag{3.23}$$

Условия принадлежности случайной величины к классу нормального распределения определяются одной из важнейших теорем теории вероятности:

Центральная предельная теорема[6]
Если независимые случайные величины $X_1, X_2, \dots X_n$ имеют одинаковые распределения с конечной, отличной от нуля дисперсией $sigma^2$, то при $n \to \infty$ сумма этих величин стремится к нормальному распределению со средним значением и дисперсией:

$$m_\Sigma = \sum_{i=1}^{n} m(X_i); \sigma_\Sigma^2 = n\sigma^2. \tag{3.24}$$

В завершение упомянем важное частное правило для случайных величин, подчиняющихся нормальному закону распределения:

Правило трех сигм Вероятность попадания случайной величины с нормальным законом распределения $\mathcal{N}(m, \sigma)$ в окрестность $\{m-3\sigma; m+3\sigma\}$ составляет 99,75%, т.е. является практически обеспеченной. Аналогичное правило для двух и одной сигм обеспечивает точность в 96% и 68% соответственно.

[5]Обозначается в англоязычной литературе как $\mathcal{N}(m, \sigma)$.
[6]Представлена в упрощенном виде.

Практика 3 – энтропия

3.3 Вероятностные и информационные характеристики

3.3.1 Качественные задачи и повторение пройденного

1 Эксперимент состоит в прочтении первой буквы на каждой странице собрания сочинений на русском языке. Указать примеры реализаций, алфавит источника, определить несколько событий различных типов, пояснить, как найти относительные частоты различных сообщений алфавита и введенных вами событий.

2 Привести несколько собственных примеров эксперимента и определить для них указанные выше основные параметры описывающей математической модели.

3 В эксперименте определены следующие вероятности алфавита источника:

$$P(\text{ж})=0{,}009;\ P(\text{э})=0{,}002;\ P(\text{о})=0{,}11;$$
$$P(\text{ч})=0{,}015;\ P(\text{ю})=0{,}007;\ P(\text{е})=0{,}087;$$
$$P(\text{ш})=0{,}007;\ P(\text{я})=0{,}022;\ P(\text{а})=0{,}075;$$
$$P(\text{щ})=0{,}004;\ P(\text{ы})=0{,}019;\ P(\text{и})=0{,}075;$$
$$P(\text{ъ, ь})=0{,}017.$$

Рис. 3.1: Вероятности алфавита источника

Определить вероятности следующих событий:
1). Получение гласной буквы.
2). Получение шипящей буквы.
3). Получение буквы, стоящей в упорядоченном алфавите после буквы "ц". Какие из указанных событий совместимы?

4 Эксперимент состоит в подбрасывании четырех монет. Определить: 1). Все возможные исходы эксперимента. 2). Вероятности результата, состоящего в выпадении одного герба и трех решеток, результатов "два герба, две решетки "три герба, одна решетка".

5 Составной эксперимент состоит в чтении текста из букв A_1, A_2, A_3, A_4 в условиях плохой освещенности. Заданы вероятности выборочных точек первого эксперимента (появление букв в тексте):

$$P(A_1) = 0.5; P(A_2) = 0.25; P(A_3) = P(A_4) = 0.125.$$

Заданы условные вероятности выборочных точек второго эксперимента:

$$\left|\begin{array}{cccc} 0.5 & 0.25 & 0.5 & 0 \\ 0.4 & 0.1 & 0.5 & 0 \\ 0.3 & 0.5 & 0.1 & 0.1 \\ 0 & 0.25 & 0.25 & 0.5 \end{array}\right| A_i$$
$$B_j$$

Перечислить исходы (выборочные точки) составного эксперимента и определить их вероятности.

3.3.2 Энтропия как мера неопределенности

1 Имеются два дискретных источника информации, заданные матрицами:

$$\left|\begin{array}{c} X \\ P \end{array}\right| = \left|\begin{array}{cc} x_1 & x_2 \\ p_1 & p_2 \end{array}\right|, \left|\begin{array}{c} Y \\ Q \end{array}\right| = \left|\begin{array}{ccc} y_1 & y_2 & y_3 \\ q_1 & q_2 & q_3 \end{array}\right|$$

Определить, какой источник обладает большей неопределенностью, если:
1). $p_1 = p_2, q_1 = q_2 = q_3$. 2). $p_1 = q_1, p_2 = q_2 + q_3$.

2 На выходе двоичного источника информации элементы «0» и «1» появляются с вероятностями P и $1-P$, соответственно. При каком значении P энтропия источника максимальна? Построить график $H(P)$ для двоичного источника.

3 Имеются два дискретных троичных источника с независимыми элементами. На выходе каждого источника появляются сообщения одинаковой длины - по 15 элементов. Количество различных элементов

в сообщении каждого источника постоянно; сообщения отличаются только порядком элементов.
Зафиксированы два типичных сообщения: 021202120212021 - первого источника и 012101201101201 - второго. Элемент какого источника несет в себе в среднем большее количество информации?

3.3.3 Условная энтропия. Взаимная информация

1 Пусть X и Y - два алфавита, при этом $Z = X + Y$. Чему равна условная энтропия $H(z|x)$, если:
1). X и Y - независимы.
2). X и Y - зависимы.
3). $X \equiv Y$?

2 Элементы алфавитов X и Y статистически связаны. Известно, что $H(x) = 8$ бит; $H(y) = 12$ бит. В каких пределах меняется условная энтропия $H(y|x)$ при изменении $H(x|y)$ в максимально возможных пределах?

3 Дана матрица

$$P(X,Y) = \begin{vmatrix} 0.125 & 0.125 & 0.125 \\ 0.125 & 0 & 0.125 \\ 0.125 & 0.125 & 0.125 \end{vmatrix}$$

Определить энтропии $H(x), H(y), H(x|y), H(y|x), H(x|y_1), H(y|x_2), H(x|y)$.

4 Значения д.с.в. X_1 и X_2 определяются подбрасыванием двух идеальных монет, а д.с.в. Y равна сумме количества «гербов», выпавших при подбрасывании этих монет. Сколько информации о X_1 содержится в Y?

5 Сколько информации о X_1 содержится в д.с.в. $Z = (X_1 + 1)^2 - X_2$, где независимые д.с.в. X_1 и X_2 могут с равной вероятностью принимать значение либо 0, либо 1? Найти $H(X_1)$ и $H(Z)$. Каков характер зависимости между X_1 и Z?

Лекция 4

Дифференциальная энтропия. Эпсилон-энтропия

4.1 Дифференциальная энтропия

4.1.1 Определение дифференциальной энтропии

В предыдущих лекциях рассматривалась энтропия как мера неопределенности выбора для дискретных источников информации. Большинство же физически существующих источников информации являются ***непрерывными***, т.е. такими, множество возможных состояний которых составляет континуум. Для обобщения на случай непрерывного источника формулы Шеннона для определения энтропии разобъем интервал возможных состояний случайной непрерывной величины X на равные непересекающиеся отрезки $\triangle x$ и рассмотрим множество дискретных состояний $x_1, x_2, \ldots, x_m$ с вероятностями $P_i = p(x_i)\triangle x (i = 1 \ldots m)$. В этом случае энтропия вычисляется по формуле:

$$\begin{aligned} H &= -\sum_{i=1}^{m} p(x_i)\triangle x \log p(x_i)\triangle x = \\ &= -\sum_{i=1}^{m} p(x_i)\triangle x \log p(x_i) - \sum_{i=1}^{m} p(x_i)\triangle x \log \triangle x. \end{aligned} \tag{4.1}$$

В пределе при $\triangle x \to 0$ с учетом соотношения $\int_{-\infty}^{\infty} p(x)dx = 1$ получаем:

$$h(x) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx - \log \triangle x. \tag{4.2}$$

Первое слагаемое в правой части соотношения имеет конечное значение, которое зависит только от закона распределения непрерывной случайной

величины X и не зависит от шага квантования $\triangle x$.[1] Оно имеет такую же структуру, как энтропия дискретного источника, но для его определения используется плотность распределения вероятности состояний непрерывного источника, т.е. *дифференциальный закон распределения* данного источника. В соответствии с этим указанная энтропия называется ***дифференциальной энтропией*** непрерывного источника информации.

Величина $\log \triangle x$ зависит только от выбранного интервала квантования $\triangle x$ и при $\triangle x = const$ она также постоянна. В связи с этим, в качестве дифференциальной энтропии в общем случае принимают значение лишь первого слагаемого, называемого также ***приведенной энтропией***; данное слагаемое целиком определяет информативность сообщений, обусловленных статистикой состояний их элементов. При этом в общем случае значение дифференциальной энтропии является *относительной величиной*, т.е. зависит от характеристик масштаба и размерности описываемой величины (информационного источника) и может быть использована для сравнения информационных источников лишь в равных условиях.

4.1.2 Свойства дифференциальной энтропии

Свойство 4. При наличии для случайной величины X единственного ограничения - области ее возможных значений $[\alpha, \beta]$, максимальной дифференциальной энтропией обладает равномерное распределение вероятностей в этой области:

$$p(x) = \frac{1}{\beta - \alpha}, \quad \alpha \leq x \leq \beta, \tag{4.3}$$

при этом

$$h_{uniform}(x) = \log \beta - \alpha. \tag{4.4}$$

Доказательство.

При доказательстве данного свойства решается задача определения плотности распределения $p(u)$, обеспечивающей максимальное значение функционала:

$$h(x) = -\int_{\alpha}^{\beta} p(x) \log p(x) dx \tag{4.5}$$

при ограничении $\int_{\alpha}^{\beta} p(x)dx = 1$.

[1] **Квантование** - разбиение диапазона значений некоторой величины на конечное число интервалов. Величина такого интервала и называется **шагом квантования**.

Приведем леммы, требуемые для доказательств указанных утверждений:

Лемма 1 (Лемма Гиббса). $\forall p, q$, таких что $\sum p = 1, \sum q = 1$, имеет место неравенство

$$\sum p \cdot \ln p \geq \sum p \cdot \ln q. \tag{4.6}$$

Доказательство леммы. Воспользовавшись известным неравенством $\ln p \leq p - 1$, для выражения

$$\begin{aligned} &\sum p \cdot \ln \frac{q}{p} \leq \sum p \cdot \left(\frac{q}{p} - 1\right) = \sum p \cdot \left(\frac{q-p}{p}\right) = \\ &= \sum (q - p) = \sum q - \sum p = 1 - 1 = 0 \rightarrow \\ &\sum p \cdot \ln \frac{q}{p} = \sum p \cdot (\ln q - \ln p) \leq 0, \end{aligned} \tag{4.7}$$

отсюда и следует утверждение леммы,

Лемма 2 (О энтропии источника с равномерным распределением). Энтропия произвольной дискретной случайной величины $X = (x_1, x_2, \ldots x_n)$ меньше энтропии равномерной величины U, распределенной на том же интервале, при этом:

$$H(X) \leq H(U) = \ln n. \tag{4.8}$$

Доказательство леммы. Обозначим распределение случайной величины X через $\mathbf{p} = (p_1, p_2, \ldots p_n)$, а распределение U - $\mathbf{q} = (q_1, q_2, \ldots, q_n) = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$. По определению, энтропия:

$$\begin{aligned} &H(X) = -\sum p \cdot \ln p \leq -\sum p \cdot \ln q = -\sum p \cdot \ln \frac{1}{n} = \\ &= \sum p \cdot \ln n = \ln n \cdot \sum p = \ln n \cdot 1 = \ln n. \end{aligned} \tag{4.9}$$

Переходя в лемме 2 к интервалу между возможными реализациями $\triangle \to 0$, получаем $n = \frac{\beta - \alpha}{\triangle}$ и значение дифференциальной энтропии (с точностью до константы $\triangle$) $H(X) = ln(\beta - \alpha)$, что и требовалось доказать.

Свойство 5. Если ограничения на область значений непрерывной случайной величины X отсутствуют, но известно, что дисперсия ее ограничена, то максимальной дифференциальной энтропией обладает нормальное

распределение величины X:

$$p(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}, \tag{4.10}$$

при этом

$$h_{normal}(X) = \log \sigma\sqrt{2\pi e}. \tag{4.11}$$

Доказательство.При доказательстве данного свойства решается задача определения плотности распределения $p(u)$, обеспечивающей максимальное значение функционала:

$$h(x) = -\int_{\infty}^{\infty} p(x) \log p(x) dx \tag{4.12}$$

при следующих ограничениях:

$$\int_{-\infty}^{\infty} p(x)dx = 1 \quad \text{(условие нормировки);} \tag{4.13}$$

$$\int_{-\infty}^{\infty} (x - \overline{x})^2 p(x)dx = \sigma^2 \quad \text{(условие ограничения дисперсии).} \tag{4.14}$$

Для этого запишем задачу Лагранжа в следующем виде:

$$F(p, \lambda, \mu) = p \ln p + \lambda \cdot p + \mu \cdot (x - \overline{x})^2 p. \tag{4.15}$$

Экстремум функции, соответствующий максимальному значению исходного функционала:

$$\frac{\delta F}{\delta p} = 0, \quad \ln p + 1 + \lambda + \mu \cdot (x - \overline{x})^2 = 0, \tag{4.16}$$

откуда:

$$p = e^{1-\lambda} e^{-\mu \cdot (x-\overline{x})^2}. \tag{4.17}$$

Из условия нормировки далее находим:

$$\int p(x)dx = e^{\lambda - 1} \int e^{\mu \cdot (x-\overline{x})^2} dx = e^{\lambda - 1} \sqrt{\frac{\pi}{\mu}} = 1 \quad \to e^{1-\lambda} = \sqrt{\frac{\mu}{\pi}}. \tag{4.18}$$

Таким образом,

$$f = \sqrt{\frac{\mu}{\pi}} e^{-\mu \cdot (x-\overline{x})^2}. \tag{4.19}$$

В свою очередь из ограничения на дисперсию следует:

$$\int (x-\overline{x})^2 p(x)dx = \sigma^2 \to \sqrt{\frac{\mu}{\pi}} \int (x-\overline{x})^2 \cdot e^{-\mu\cdot(x-\overline{x})^2} dx = \frac{1}{2\mu} = \sigma^2; \tag{4.20}$$

итого в результате получаем нормальный закон распределения, что и требовалось доказать:

$$f = \sqrt{\frac{1}{2\pi\sigma^2}} e^{-\frac{(x-\overline{x})^2}{2\sigma^2}}. \tag{4.21}$$

В случае сигналов в радиосвязи дисперсия σ^2 случайного процесса X пропорциональна средней мощности сигнала. Таким образом, корректным является следующее утверждение:

При заданной мощности наибольшей средней неопределенностью выбора будет обладать источник, генерирующий сигналы, амплитуды которых распределены по нормальному (гауссовому) закону[2].

Свойство 6. Свойства аддитивности энтропии:
Свойства аддитивности зависимых и независимых непрерывных источников совпадают с случаем дискретных источников информации и доказываются аналогично.

4.2 Эпсилон-энтропия случайной величины

Рассмотрим состояние некоторой системы в виде ансамбля реализаций случайной величины U, описываемой плотностью распределения вероятностей $p(U)$. Пусть состояние данной системы измеряется некоторыми датчиками, сведения от которых представляют собой ансамбль реализации случайной величины Z с плотностью распределения $p(Z)$[3]. Для количественной оценки степени сходства вектора состояния и вектора измерения требуется ввод функции ***метрики сходства*** $p(Z,U)$, имеющую областью определения все множество значений Z и U.

[2]К сожалению, аналогичное утверждение справедливо и для помех: наиболее пагубное воздействие на сигнал оказывают помехи, чье распределение вероятности наиболее близко к гауссовому.

[3]В этом случае говорят, что Z воспроизводит U.

Говорят, что вектор наблюдения Z верно воспроизводит вектор состояния U, если т.н. ***критерий верности*** $V(ZU)$[4] не превышает заданной верности определения - определенного порога ϵ:

$$V(ZU) = \iint p(u)p_u(z)(z-u)^2 du dz \leq \epsilon^2, \tag{4.22}$$

где $p_u(z)$ - условная плотность распределения - *функция правдоподобия* того, что конкретный сигнал u будет воспринят датчиками как сигнал z с заданным значением верности не хуже[5], чем ϵ. При этом, так как плотность $p(u)$ определена, то для выполнения условия 4.22 варьировать возможно только рассмотренной условной плотностью распределения $p_u(z)$.

Итак, если случайная величина Z воспроизводит случайную величину U с некоторой верностью ϵ, то количество информации, содержащееся в Z относительно U конечно и может быть записано в следующей форме:

$$I(ZU) = \iint p(u)p_u(z) \log \frac{p_u(z)}{p(z)}, \quad p(z) = \int p(u)p_u(z)du. \tag{4.23}$$

Очевидным является тот факт, что желательно обеспечить заданную верность воспроизведения при минимальном количестве получаемой информации - т.е. из множестве функций $p_u(z)$, удовлетворяющих условию 4.22 целесообразно выбрать минимизирующую значение $I(ZU)$. Исходя из вышесказанного, дадим определение эпсилон-энтропии наблюдаемой случайной величины U:

Эпсилон-энтропией $H_\epsilon(U)$ величины U называется минимальное количество информации в наблюдаемой случайной величине Z относительно U, при котором удовлетворяется заданное требование к верности воспроизведения величины U:

$$H_\epsilon(U) = \min_{|p_u(z)|} I(ZU) = h(U) - \max_{|p_u(z)|} h_z(U) \quad \text{при} \quad V(ZU) \leq \epsilon^2, \tag{4.24}$$

где $h(U)$ и $h_z(U)$ - безусловная и условная дифференциальные энтропии величины U, соответственно.

Пример 17. Найти $H_\epsilon(U)$ источника информации, ансамбль состояний которого описывается нормально распределенной случайной величиной U с дисперсией σ^2 при верности воспроизведения $V(ZU) \leq \epsilon^2$.

[4]Представляющий собой математическое ожидание метрики $p(Z,U)$, взятое по всей области определения.

[5]Не больше.

Решение. Будем считать, что заданная вероятность воспроизведения обусловлена действием аддитивной статистически независимой от сигнала помехи Ξ, при этом $E[\Xi] = 0; \quad E[\Xi^2] = \epsilon^2$. Вектор состояния системы[6] u в данном случае возможно рассмотреть как сумму вектора наблюдения[7] z и помехи ζ:

$$u = z + \zeta. \tag{4.25}$$

Т.к. в данном случае $h_z(U)$ в выражении 4.24 полностью определяется помехой:

$$h_z(U) = h_z(Z + \Xi) = h(\Xi), \tag{4.26}$$

то

$$H_\epsilon(U) = h(U) - \max_{p(\zeta)} h(\Xi), \tag{4.27}$$

где $h(\Xi)$ - дифференциальная энтропия помехи; $p(\zeta)$ - плотность распределения помехи Ξ. /par Ранее (см. 4.11) нами было установлено, что при ограничении на дисперсию случайной величины максимальной дифференциальной энтропией обладает нормальное распределение. В соответствии с этим получаем:

$$\max_{p(\zeta)} h(\Xi) = \log \epsilon\sqrt{2\pi e}; \quad H(U) = \log \sigma\sqrt{2\pi e}, \tag{4.28}$$

откуда получаем результирующее выражение для эпсилон-энтропии:

$$H_\epsilon(U) = \log \sigma\sqrt{2\pi e} - \log \epsilon\sqrt{2\pi e} = \frac{1}{2} \log \frac{\sigma^2}{\epsilon^2}. \tag{4.29}$$

В описанном выражении 4.29 σ^2 определяет собой среднюю мощность P_U сигнала, а ϵ^2 - среднюю мощность P_ζ помехи. Данный результат является очень важным для теории связи и характеризует зависимость эпсилон-энтропии от величины P_U/P_ζ, представляющую собой так называемое ***отношение сигнал-помеха***.

Таким образом:

При заданном отношении сигнал-помеха значение $H_\epsilon(U)$ для нормально распределенной случайной величины является максимально возможным.

Аналогично, для произвольно распределенной случайной величины U при том же критерии верности и малых ϵ[8], справедливо следующее

[6] По сути-передаваемый сигнал.

[7] Воспроизводящего сигнала

[8] Когда значение дифференциальной энтропии $H_\epsilon(U)$ велико.

приближенное равенство:

$$H_\epsilon(U) \approx h(U) - \log \epsilon\sqrt{2\pi e}. \tag{4.30}$$

Практика 4 – дифференциальная энтропия

4.3 Энтропия непрерывного источника

4.3.1 Дифференциальная энтропия

1 Определить дифференциальную энтропию равномерного на интервале $[-W_1 + W_2]$ распределения.

2 Определить дифференциальную энтропию $h(x)$ нормального распределения с плотностью вероятности

$$\rho(x) = \frac{1}{\sigma(x)\sqrt{2\pi}} e^{-\frac{(x-\overline{x})^2}{2\sigma^2(X)}}.$$

Как влияет на величину $h(x)$ увеличение в два раза
а). Среднего $\overline{x}$; б). Дисперсии $\sigma^2(x)$.

3 Определить энтропию двумерного равномерного распределения, заданного плотностью

$$\rho(x,y) = \rho(x)\rho(y) = \begin{cases} 1, 0 \leq x \leq 1, \\ 0, 0 \leq y \leq 1. \end{cases}$$

4 Найти энтропию системы m случайных величин, распределенных по нормальному закону.

5 Информация передается с помощью частотно-модулированных сигналов, рабочая частота F которых изменяется с равной вероятностью в пределах от $F_1 = 10$ МГц до $F_2 = 50$ МГц. Определить энтропию частоты, если точность измерения частоты $\triangle F = 2$ кГц.

6 В результате повреждения аппаратуры центра управления полетами n самолетов летят произвольными курсами. Через некоторое время

управление было восстановлено и все самолеты взяли общий курс со среднеквадратической ошибкой отклонения от курса $\sigma = 3^\circ$. Найти изменение энтропии, считая, что в первом случае имело место равномерное распределение вероятностей углов, а во втором случае-нормальное.

7 Точность выдерживания курса БПЛА под действием управляющих команд изменилась с 3° до $10'$ при равномерном распределении ошибки курса. ДО какой величины пришлось бы изменить СКО, если бы ошибка была распределена по нормальному закону, чтобы обеспечить такое же изменение энтропии, как в первом случае?

Тема II

Теоретические основы каналов связи

Тема 2 – список литературы

1 *Дмитриев В.И.* Прикладная теория информации. М.: Букинист, 1989. - 332 с.

2 *Думачев В.Н.* Теория информации и кодирования. Воронеж: Воронежский институт МВД России, 2012. - 200 с.

3 *Липкин И.А.* Статистическая радиотехника. Теория информации и кодирования. М.: Вузовская книга, 2002. - 216 с.

4 *Скляр Б.* Цифровая связь. Теоретические основы и практическое применение. М.: изд. дом Вильямс, 2003. - 1104 с.

5 *Орлов В.А., Филиппов Л.И.* Теория информации в упражнениях и задачах. М.: Высшая Школа, 1976. - 136 с.

6 *Кавчук С.В.* Сборник примеров и задач по теории информации. Таганрог: изд. Таганрогского ГРУ, 2002. - 64 с.

Лекция 5

О каналах связи и источниках сообщений

5.1 Источники информации и каналы связи

Вторая тема представленного курса лекций посвящена вопросам формализованного описания источников сообщений и каналов связи; установления пути повышения эффективности систем передачи информации; определению условий достижения максимальной скорости передачи данных по каналам с помехами и без помех.

5.1.1 Основные определения

Введем далее основные формулировки, необходимые для рассмотрения различных моделей источника сообщений и канала связи.

Сигнал *– материальный носитель информации, используемый для передачи сообщений в системе связи.*

Сигнал может генерироваться, но его прием не обязателен, в отличие от сообщения, которое должно быть принято принимающей стороной. Сигналом может быть **любой физический процесс**, параметры которого изменяются в соответствии с передаваемым сообщением.

Источник дискретных сообщений *– объект, формирующий дискретные последовательности из ограниченного числа элементарных сообщений.*

Источник непрерывных сообщений *– объект, формирующий непрерывную последовательность, некоторые параметры которой определяются исходя из статистических свойств указанного объекта.*

Источником без памяти *– такой источник сообщений, выбор каждого $i-$го сообщения в котором не зависит от предыдущего ($i-1$-го).*

Источником с памятью *– источник, в котором выбор каждого сообщения зависит от того, выбор каких сообщений совершался до этого.*

В теории информации произвольный источник сообщений полностью определяется статистическими данными о формируемых им сообщениях.

Канал связи *– совокупность устройств и физических сред, обеспечивающих на определенном временном интервале передачу сообщений от одного объекта к другому.*

Дискретным *называется канал связи, использующийся для передачи дискретных сообщений.*

Непрерывным *называется канал, предназначенный для передачи непрерывных сообщений (сигналов*[1]*).*

[1]Для упрощения рассуждений здесь и далее будем использовать как синонимичные понятия *сообщения* и *сигнал*

Каналом без помех *называется идеализированная модель канала связи, в которой возможно пренебречь вредным действием помех.*

В канале без помех каждому сообщению на входе однозначно соответствует определенное сообщение на выходе и наоборот.

Когда требования к достоверности достаточно велики и пренебрежение неоднозначностью связи между сообщениями недопустимо, используется более сложная модель - ***канал с помехами***.

Канал считается заданным, если известны статистические данные о сообщениях на его входе и выходе, а также ограничения, накладываемые на входящие сообщения физическими характеристиками канала.

Каналом без памяти *называется канал связи, для которого статистические свойства сигнала на выходе в некоторый момент времени t определяются только сигналом на входе, переданном в этот момент времени и, следовательно, не зависят от сигналов, переданных до и после этого момента времени t.*

Каналом с памятью[2] *называется канал связи, для которого статистические свойства сигнала на выходе в некоторый момент времени t определяются сигналами на входе, переданными в интервале времен t',*

$$t - \tau \leq t' \leq t.$$

Величина τ называется ***длиной*** канала с памятью.

Каналом с обратной связью *называют* ***прямой канал передачи***[3]*, дополненный* ***обратным каналом***.[4]

5.1.2 Стационарность и эргодичность источников информации

Сигнал в теории информации представляется в виде случайного процесса. Введем базовые категории интересующих нас классов случайных процессов:

Стационарным *называется такой случайный процесс, у которого вероятностные закономерности неизменны во времени*[5].

[2]Точнее – каналом с *конечной* памятью.

[3]От источника сообщений к получателю.

[4]От получателя сообщений к источнику, зачастую является служебным каналом – например, для запроса повторной передачи в случае обнаружения ошибки.

[5]В противном случае случайный процесс называется *нестационарным*.

Эргодическим *называется такой случайный процесс, для которого операцию усреднения по статистическому ансамблю*[6] *возможно заменить усреднением по времени.*

Понятия *стационарности* и *эргодичности* применимо к источникам сообщений понимаются следующим образом:

Стационарным источником сообщений *называется такой источник, для которого вероятности появления отдельных знаков* [7] *и их сочетаний не зависят от расположения последних по длине сообщения.*

Эргодическим источником сообщений *называется такой источник, для которого статистические закономерности, полученные при исследовании одного достаточно длинного сообщения справедливы (эквивалентны с вероятностью, близкой к единице) для всех сообщений, создаваемых источником.*

Примем без доказательства также два интересных взаимоотношения стационарных и эргодических источников сообщений:

- Стационарный источник сообщений без памяти **всегда** является эргодическим.
- Любой стационарный источник сообщений может быть представлен совокупностью нескольких эргодических источников, различающихся режимами работы.

Источники без памяти являются частным случаем (при длине памяти $\tau = 0$) источников с конечной памятью. В свою очередь, источники с конечной памятью в теории информации классически описываются т.н. ***цепями Маркова***:

Цепь Маркова *порядка n характеризует последовательность событий, вероятности которых зависят от того, какие n событий предшествовали данному.*

Простой цепью Маркова называется цепь Маркова порядка 1. В общем случае порядок n цепи Маркова показывает количество предшествующих знаков, которые определяют состояние источника, в котором он находится при выдаче текущего знака.

[6] Возможным вариантам реализации в фиксированный момент времени.

[7] Знаки в данной трактовке - отдельные элементы сообщения, выбираемые из заданного алфавита. Синонимы понятия «знак» в данной трактовке - понятия «символ», «сигнальный символ» и «канальный символ».

Согласно элементарным правилам комбинаторики при объеме алфавита знаков, равном L число R различные состояний источника, описываемого цепью Маркова порядка n не превышает L^n:

$$R \leq L^n. \tag{5.1}$$

Обозначим множество R возможных состояний за $S = \{S_1 \dots S_q \dots S_R\}$, а вероятность выбора в состоянии S_q знака z_l через $p_q(z_l)$. При определении вероятности $p_q(z_l)$ естественно предположить, что к моменту выдачи источником очередного знака известны все знаки, сгенерированные им ранее – история состояния источника на текущий момент.

Итак, если источник находится в состоянии S_q, его частная энтропия $H(S_q)$ определяется соотношением:

$$H(S_q) = -\sum_{l=1}^{L} p_q(z_l) \log p_q(z_l). \tag{5.2}$$

Усредняя случайную величину $H(S_q)$ по всем возможным состояниям $q = \overline{1, R}$, получаем искомую энтропию источника сообщений:

$$H(Z) = -\sum_{q=1}^{R} p(S_q) \sum_{l=1}^{L} p_q z_l \log p_q(z_l), \tag{5.3}$$

где $p(S_q)$ - вероятность того, что источник сообщений находится в состоянии S_q. Итоговая величина $H(Z)$ характеризует неопределенность, приходящуюся в среднем на один знак, выдаваемый источником сообщений.

Согласно выражению 5.3 для источника без памяти при выборе источником знака z_l на любом из предшествующих шагов его состояние не меняется ($R = 1$). Следовательно, $p(S_1) = 1$ и для энтропии источника сообщений справедливо следующее выражение:

$$H(Z) = -\sum_{l=1}^{L} p(z_l) \log p(z_l). \tag{5.4}$$

В завершение раздела рассмотрим случай простой цепи Маркова. В этом случае максимальное число различных состояний источника равно объему алфавита. Следовательно, $R = L$ и $p_q(z_l) = P(z_l|z_q)$, где $q = \overline{1, L}$. В

этом случае выражение для определения энтропии приобретает следующий вид:

$$H(Z) = -\sum_{q=1}^{L} p(z_q) \sum_{l=1}^{L} p(z_l|z_q) \log p(z_l|z_q). \tag{5.5}$$

Аналогично можно получить выражения для энтропии источника сообщений с памятью произвольной длины.

Пример 18. Определить, является ли эргодическим стационарный дискретный источник сообщений, алфавит которого состоит из четырех знаков z_1, z_2, z_3, z_4, при этом безусловные вероятности выбора знаков одинаковы: $p(z_1) = p(z_2) = p(z_3) = p(z_4) = 0.25$, а условные вероятности заданы таблицей 5.1:

Таблица 5.1: значения условных вероятностей

z_q \ z_l	z_1	z_2	z_3	z_4
z_1	$1/3$	$1/3$	$1/3$	0
z_2	$1/3$	$1/3$	$1/3$	0
z_3	$1/3$	$1/3$	$1/3$	0
z_4	0	0	0	1

Решение. Анализ таблицы 5.1 показывает, что источник имеет два режима работы: с вероятностью, равной $3/4$, первым будет выбран один из знаков z_1, z_2 или z_3 и источник начнет формировать последовательность с равновероятным появлением знаков. Если же первым будет выбран знак z_4 (вероятность этого равна, как видно из таблицы, $1/4$), то будет сгенерирована последовательность, содержащая только знаки z_4.

Усреднение по ансамблю предполагает наличие множества однотипных источников, примерно три четверти из которых будут работать в первом режиме, а одна четверть - во втором. При этом, в соответствии с 5.5 энтропия источника равна:

$$H(Z) = -\frac{3}{4} \log \frac{1}{3} - \frac{1}{4} \log 1 = 1,19 \text{ bit}.$$

В свою очередь, среднее по последовательности (по времени) вычисляется с использованием конкретной последовательности и поэтому зависит от режима функционирования источника. В первом режиме неопределенность, приходящаяся на один символ достаточно длинной последовательности[8], равна 1,586 бит, а во втором случае - нулю. Таким образом, поскольку энтропии формируемых последовательностей не совпадают с энтропией источника, то **данный источник не является эргодическим**.

5.2 Характеристики источников сообщений

5.2.1 Свойство асимптотической равномерности

Одним из основных свойств для эргодических сообщений является т.н. свойство асимптотической равномерности. Определим данное свойство в виде теоремы, а также приведем доказательство указанной теоремы для простейшего случая эргодического источника без памяти:

Свойство 7 (**Асимптотическая равномерность длинных последовательностей эргодического источника сообщений**). $\forall \delta > 0, \mu > 0$ для достаточно большой длины последовательности N все последовательности, генерируемые эргодическим источником могут быть разбиты на две группы:

1 Т.н. *типичные* последовательности, характеризующиеся тем, что вероятности их появления при достаточно большом N практически одинаковы, при этом вероятность ρ любой такой последовательности удовлетворяет условию:

$$\left| \frac{\log(1/\rho)}{N} - H(Z) \right| < \mu, \tag{5.6}$$

где $H(Z)$ - энтропия рассматриваемого источника сообщений.

2 Последовательности с ничтожными вероятностями осуществления - так называемые *нетипичные* последовательности: сумма вероятностей всех таких последовательностей $\sum p_i < \delta$. При этом допу-

[8]Так называемая энтропия последовательности

щении[9] их количество (при достаточно большом N является существенно большим по отношению к количеству типичных последовательностей.

Раскроем подробнее неравенство, которому подчиняются типичные последовательности:

Итак, поскольку при $N \longrightarrow \infty$ источник сообщений с вероятностью, сколь угодно близкой к единице, выдает только типичные последовательности, что и принимаемое во внимание число последовательностей равно $1/\rho$. Неопределенность создания каждой такой последовательности с учетом их равновероятности составляет $\log 1/p$. Тогда величина $\log 1/p/N$ представляет собой неопределенность, приходящуюся в среднем на один знак. Интуитивно ясно, что данная величина практически не должна отличаться от энтропии источника, что и констатируется соотношением 5.6.

Доказательство.

Докажем вначале первую часть теоремы, касающихся типичных последовательностей:

Из закона больших чисел[10] следует следующее - для достаточно длинной последовательности из N элементов алфавита $(z_1, z_2, \ldots, z_L)$, имеющих вероятности появления $p_1, p_2, \ldots, p_L$, содержится $N \cdot p_1$ элементов z_1, $N \cdot p_2$ элементов z_2 и т.д. В этом случае вероятность ρ реализации любой типичной последовательности (последовательности, состоящей именно из такого количества элементов) близка к величине:

$$p = p_1^{p_1 \cdot N} p_2^{p_2 \cdot N} \ldots p_L p_L \cdot N. \tag{5.7}$$

Логарифмируя обе части данного выражения, получаем:

$$\log p = N \sum_{i=1}^{L} p_i \log p_i,$$

[9]За исключением случая равновероятного и независимого выбора знаков источником (источник является источником без избыточности; само понятие избыточности будет рассмотрено ниже), т.к. в этом случае нетипичные последовательности отсутствуют.

[10]В данном случае под законом больших чисел понимается следующее:
Всегда найдётся такое количество испытаний, при котором с любой заданной наперёд вероятностью относительная частота появления некоторого события будет сколь угодно мало отличаться от его вероятности.
При этом **относительная частота** - отношение количества исходов, соответствующих данному событию к общему количеству исходов.

откуда и получаем искомое выражение:

$$\frac{\log(1/p)}{N} = H(Z).$$

Для общего случая теорема доказывается с привлечением мат. аппарата свойств цепей Маркова.

Докажем теперь вторую часть теоремы:[11]

При объеме алфавита источника L и количестве знаков в последовательности N число всех возможных последовательностей определяется по формуле размещений с повторениями:

$$n_L = L^N = 2^{N \log_2 L}. \tag{5.8}$$

Принимая во внимание соотношение 5.6, число типичных последовательностей n_{typ} возможно записать в следующем виде:

$$n_{typ} = 2^{NH(Z)}. \tag{5.9}$$

В этом случае соотношение между нетипичными и типичными последовательностями определяется следующим образом:

$$\frac{n_{untyp}}{n_{typ}} = 2^{N[\log_2 L - H(Z)]}.$$

При этом, т.к. $H(Z) < \log_2 L$ для случая неравномерных последовательностей, то

$$n_{typ} \ll n_{untyp},$$

при этом неравенство усиливается с увеличением N.

К. Шеннон показал, что рассмотренные свойства длинных последовательностей являются основанием для осуществления эффективного кодирования информации, виды которого рассматриваются в теме 3 представленного курса лекций.

Пример 19. Оценить, какую долю общего числа возможных последовательностей следует учитывать в практических расчетах, если эргодический источник характеризуется следующими параметрами: $L = 16, H(Z) = 3,5\ bit, N = 50$.

[11] О существенном превалировании количества нетипичных последовательностей перед типичными.

Решение.

$$n_{untyp} = 16^{50} = 2^{200};\ n_{typ} = 2^{50 \cdot 3.5} = 2^{175},$$

откуда

$$\frac{n_{typ}}{n_{untyp}} = \frac{2^{175}}{2^{200}} = \frac{1}{2^{25}} \approx \frac{1}{30 \cdot 10^6}.$$

Таким образом, к типичным последовательностям в данном случае относится только одна тридцатимиллионная доля всех возможных реализаций.

5.2.2 Избыточность источника сообщений

Большинство физических источников сообщений с точки зрения генерации информационных последовательностей работают неоптимально: в большинстве случаев количество информации, переносимое сообщениями существенно отличаются от оптимального случая[12]. Таким образом, в большинстве случаев ту же информационную нагрузку на знак возможно обеспеить, используя алфавит меньшего объема. В связи с этим говорят о ***избыточности алфавита*** источника сообщений или просто о ***избыточности источника***.

Мерой избыточности служит величина D, показывающая, насколько хорошо используются знаки данного источника:

$$D = \frac{H_{max}(Z) - H(Z)}{H_{max}(Z)}, \tag{5.10}$$

где $H_{max}(Z)$ - максимально возможная энтропия источника, равная $\log L$; $H(Z)$ - реальная энтропия источника.

Если избыточность источника равна нулю, то формируемые им сообщения оптимальны в смысле наибольшего количества переносимой информации. Для передачи определенного количества информации I при отсутствии помех в этом случае необходимо $k_I = \frac{I}{H_{max}(Z)}$ знаков.

Поскольку энтропия сообщений, формируемых реальным источником, обладающим некоторой избыточностью, меньше максимальной, то для передачи того же количества информации I знаков требуется больше, а именно:

$$k_{I-real} = /fracIH(Z) > k_I.$$

Поэтому говорят также о ***избыточности знаков в сообщении*** или просто о ***избыточности сообщения***, характеризуя ее тем же параметром D:

$$D = \frac{k_{I-real} - k_I}{k_{I-real}} = \frac{H_{max}(Z) - H(Z)}{H_{max}(Z)}. \tag{5.11}$$

Безусловно, избыточность нельзя рассматривать как признак несовершенства источника сообщений. Обычно она является следствием его физических свойств - так, ограничения, существующие в любом человеческом языке связаны, например, с особенностями артикуляции, не позволяющими формировать слова, состоящие из произвольных сочетаний букв.

[12]При независимом и равновероятном выборе знаков.

Последствия от наличия избыточности неоднозначны. С одной стороны, избыточные сообщения требуют дополнительных затрат при передаче - например, увеличения времени передачи или расширения полосы частот. С другой стороны, при адекватном учете избыточности (или введении искусственной избыточности) становится возможно применять алгоритмы обнаружения и исправления ошибок, что обеспечивает повышение помехоустойчивости передачи. Так, высокая избыточность большинства естественных языков обеспечивает достаточно надежное общение людей даже при наличии дефектов речи и неидеальной слышимости.

В большинстве реальных систем связи естественная избыточность сообщений перед передачей устраняется - при помощи алгоритмов эффективного (оптимального) кодирования[13], а искусственная - добавляется для повышения помехоустойчивости - при помощи помехоустойчивого кодирования информации[14].

Пример 20. Определить возможный эффект от устранения избыточности при передаче текста на русском языке.

Решение. Максимальная энтропия текста на русском языке (с учетом пренебрежения при передаче различий в буквах е и ё, ъ и ь) равна 5 бит (см. Дмитриев, пример 3.3.). Там же определена энтропия с учетом неравномерного распределения вероятностей появления отдельных букв (4.42 бит). Имея сведения о переходных вероятностях и упрощенно представляя текст в виде простой цепи Маркова, можно определить, что энтропия уменьшается до 3.52 бит. Учет всех ограничений в языке, включая связи между словами, позволяет по различным экспертным оценкам ограничить минимальное значение энтропии значением 1.5 бит. Таким образом, избыточность русского языка составляет:

$$D = \frac{H_{max}(Z) - H(Z)}{H_{max}(Z)} = \frac{5 - 1.5}{5} = 0.7 = 70\%.$$

Таким образом, полное устранение избыточности при передаче текстов на русском языке позволит повысить эффективность использования каналов связи более чем в три раза.

---

[13]Рассматриваются в теме 3 приведенного курса лекций.

[14]Рассматривается в теме 4 приведенного курса лекций.

5.2.3 Производительность источника сообщений

Рассмотрим последнюю интересующую нас в рамках данной лекции характеристику источников сообщений. Итак:

Производительностью источника сообщений *называется количество информации, вырабатываемое источником в единицу времени*[15].

Поскольку возможное воздействие помех на источник сообщений принято учитывать эквивалентным изменением характеристик модели канала связи, то производительность источника сообщений равна энтропии источника, приходящейся на единицу времени и измеряется в ***бит-с***.

Длительность выдачи знаков источником в каждом из состояний в общем случае может быть различной. Обозначим длительность выдачи знака z_l, формируемого источником в состоянии S_q через τ_{qz_l}. В этом случае средняя длительность выдачи источником одного знака равна:

$$\tau_{mean} = \sum_{q=1}^{R} p(S_q) \sum_{l=1}^{L} p_q(z_l)\tau_{qz_l}. \tag{5.12}$$

Производительность источника $\overline{I}(Z)$ таким образом возможно выразить формулой:

$$\overline{I}(Z) = \frac{H(Z)}{\tau_{mean}}. \tag{5.13}$$

Как следует из 5.12, повышение производительности источника возможно не только за счет увеличения энтропии, но и за счет снижения средней длительности формирования знака. Длительность знаков желательно выбирать обратно пропорциональными вероятностям их появления.

Если длительность выдачи знака не зависит от состояния источника, для всех знаков одинакова и равна τ, то $\tau_{mean} = \tau$ и соответствует передаче $v = \frac{1}{\tau}$ символов в единицу времени. В этом случае выражение 5.13 для производительности источника имеет смысл средней скорости поступления информации от источника сообщений и принимает следующий вид:

$$\overline{I}(Z) = \frac{H(Z)}{\tau} = H(Z) \cdot v. \tag{5.14}$$

[15] Данную характеристику называют также ***скоростью сообщений*** или ***интенсивностью потока*** входной информации.

Лекция 6

Дискретные каналы связи

6.1 Дискретные каналы связи

6.1.1 Модели дискретных каналов связи

В общем виде информационная модель канала с помехами задается множеством символов на его входе и выходе и описанием вероятностных свойств передачи отдельных символов. В общем случае канал может иметь множество состояний и переходить из одного состояния в другое как с течением времени, так и в зависимости от последовательности передаваемых символов.

В каждом своем состоянии канал характеризуется матрицей условных вероятностей $P(U_i|Z_j)$ того, что переданный символ U_i будет воспринят на выходе кака символ Z_j. Таким образом, ***нестационарный*** канал связи имеет зависимость матрицы P от времени, но при этом может быть представлен рядом стационарных каналов, соответствующим различным временным интервалам.

Информационная модель канала без помех является частным случаем с определением однозначных соответствий между множествами символов на входе и выходе канала (матрица условных вероятностей заполняется единицами). В общем случае **M-ичным каналом** (с помехами, либо без помех) называется такой канал связи, в котором размеры ансамблей U и Z совпадают.

Важной частной моделью, широко использующейся при анализе и построении систем и сетей телекоммуникаций является *стационарный дискретный двоичный канал без памяти*. Данная модель однозначно определяется четырьмя условными вероятностями: $p(0|0),\ p(1|0),\ p(0|1),\ p(1|1)$.

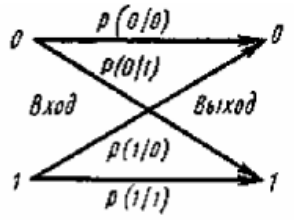

Рис. 6.1: Дискретный двоичный канал без памяти

На указанном рис. 6.1 $p(0|0)$ и $p(1|1)$ - вероятности неискаженной передачи символов, а $p(0|1)$ и $p(1|0)$ - вероятности искажения (трансформации) символов 0 и 1 соответственно.

Если вероятности искажения символов приблизительно равны:

$$p(0|1) \approx p(1|0) = q,$$

то такой канал называют двоичным ***симметричным*** каналом, иначе - ***несимметричным***.

Достаточно интересной также является мат. модель т.н. двоичного дискретного канала со *стиранием*. На выходе такого канала возможен символ неопределенности состояния (когда символ с равным основанием может быть отнесен к нулю, либо единице) - *символ стирания*, обозначающийся S:

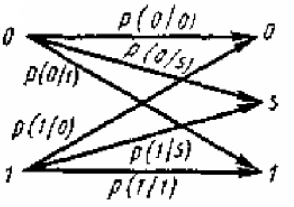

Рис. 6.2: Дискретный двоичный канал со стиранием

При декодировании символы стирания существенно проще исправить, чем ошибочно определенные.

6.2 Теоремы Шеннона для дискретных каналов связи

6.2.1 Теорема Шеннона для дискретного канала без помех

Определим понятие пропускной способности канала связи:

Пропускной способностью канала связи *называется величина C, характеризующая максимально возможную скорость передачи информации по каналу с данными техническими характеристиками.*

Исходя из указанного определения, возможно записать выражение для пропускной способности канала связи:

$$= \max \overline{I}(U, Z) = \max v \cdot I(U, Z). \tag{6.1}$$

Для случая канала без помех имеет место взаимно-однозначного соответствия между множествами символов U и Z на входе и выходе канала, соответственно. Таким образом, для канала без помех:

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

При этом максимум возможного количества информации на символ равен $\log M$, где M - объем алфавита символов; таким образом, пропускная способность дискретного канала без помех определяется следующим образом:

$$C = v_{channel} \log M. \tag{6.2}$$

Исходя из свойства асимптотической сходимости становится возможно доказать **теорему Шеннона** для дискретного канала без помех:

Теорема 1 (Теорема Шеннона для дискретного канала без помех)**.** Если пропускная способность дискретного канала без помех превышает производительность источника сообщений, т.е. удовлетворяется условие:

$$v_{channel} \log M > v_{source} H(U), \tag{6.3}$$

то существует способ кодирования и декодирования сообщений источника с энтропией $H(A)$, обеспечивающий сколь угодно высокую однозначного и верного декодирования. В противном случае такого способа не существует.

Доказательство.Для доказательства теоремы пронумеруем все типичные последовательности достаточно большой временной длины $T = \frac{N}{v_{source}}$ цифровыми комбинациями B с основанием M, равным объему алфавиту канала. Таким образом, число различных кодовых комбинаций равно:

$$N(B) = M^N = 2^{T \cdot v_{source} \log M}. \tag{6.4}$$

Число типичных последовательностей длительностью N в соответствии с (5.9)

$$N_{typ}(U) = 2^{T \cdot v_{source} \cdot H(U)}$$

. Условие теоремы Шеннона 6.3 возможно записать в виде $v_{channel} \log M = v_{source} H(A) + \epsilon$, где ϵ - сколь угодно малая величина и, следовательно:

$$\frac{N(B)}{N_{typ}(U)} = 2^{T \cdot \epsilon}. \tag{6.5}$$

Примем

$$\epsilon = \log \frac{e}{T N_{typ}(U)}. \tag{6.6}$$

В этом случае:

$$\frac{N(B)}{N_{typ}(U)} = e^{1/N_{typ}(U)} = 1 + \frac{1}{N_{typ}(U)} + \frac{1}{2![N_{typ}(U)]^2} + \dots$$

или

$$N(B) > N_{typ}(U) + 1.$$

Таким образом, при выполнении базового условия теоремы Шеннона число различных комбинаций по крайней мере на единицу больше числа типичных последовательностей источника. Эту избыточную кодовую комбинацию поставим в соответствие всем нетипичным последовательностям, предопределив их недостоверную передачу. Поскольку при $T \longrightarrow \infty$ и $N \longrightarrow \infty$, то вероятность появления нетипичной последовательности стремится к нулю, а величина ϵ бесконечно мала, то первую часть теоремы возможно считать доказанной.

Докажем теперь вторую часть теоремы - в случае нарушения условия 6.3, когда $v_{channel} \log M > v_{source} H(U)$, используя аналогичный подход при доказательстве, получаем неравенство

$$N_{typ}(U) > N(B) + 1.$$

Из полученного неравенства следует, что даже при равновероятном способе кодирования (соответствующем предельной скорости передачи информации по каналу связи) невозможно закодировать и передать все типичные последовательности N_{typ}, что и требовалось доказать. Оптимальное кодирование, использованное при доказательстве теоремы Шеннона для дискретного канала без помех, сводится к предельному укрупнению

алфавита канала, когда каждый укрупненный символ (кодовая комбинация) отвечает бесконечно длинной последовательности символов источника сообщений. При этом одновременно устраняется корреляция между символами укрупненного алфавита канала и благодаря сохранению лишь типичных последовательностей обеспечивается равная вероятность их появления. В результате полностью устраняется избыточноть сообщения, передаваемого по каналу.

Кодирование данным способом связано с задержкой передачи сообщения (*латентностью передачи*) на время:

$$\tau_{latency} = 2T + T_0, \tag{6.7}$$

где время $2T$ определяется тем, что кодирование может начаться, лишь когда известна уже вся последовательность символов источника длительностью T, а декодирование - когда уже принята кодовая комбинация той же длительности; T_0 - время, затрачиваемое на технические операции кодирования, декодирования и на прохождение сигнала по каналу[1].

6.2.2 Теорема Шеннона для дискретного канала с помехами

При наличии помех соответствие между множествами символов на входе и выходе канала связи перестает быть однозначным. Среднее количество информации $I(U, Z)$, передаваемое по каналу одним символом, определяется в этом случае соотношением:

$$I(VU) = H(U) - H_Z(U) = H(Z) - H_U(Z). \tag{6.8}$$

Если статистические связи между символами отсутствуют, энтропия на выходе канала связи равна:

$$H(U) = -\sum_{m=1}^{M} p(u_m) \log p(u_m). \tag{6.9}$$

При наличии статистической связи энтропию определяют с использованием цепей Маркова. Для простоты рассуждений ограничимся здесь случаем отсутствия корреляции между отдельными символами.

[1]Необходимо отметить, что в канале без помех источником ненадежности отождествления передаваемых сообщений может быть только операция кодирования, т.к. нарушение соответствия при передаче сообщений по каналу исключено.

Итак, если объем алфавита входных символов U равен M, а выходных символов Z - L, то апостериорная энтропия[2] входного сообщения U после приема конкретного символа может быть вычислена по следующей формуле:

$$H_Z(U) = -\sum_{m=1}^{M}\sum_{l=1}^{L} p(u_m, z_l)\log p(u_m|z_l). \tag{6.10}$$

Величина $H_Z(U)$ по терминологии, введенной Клодом Шенноном, называется ***ненадежностью канала***. Таким образом, получаем скорость передачи информации по каналу с помехами:

$$\overline{I}(U,Z) = v_{symbol}\sum_{m=1}^{M}\sum_{l=1}^{L} p(u_m, z_l)\log\frac{p(u_m, z_l)}{p(u_m)p(z_l)}, \tag{6.11}$$

где v_{symbol} - символьная скорость передачи по каналу связи. Согласно своему определению, пропускная способность канала с помехами является пределом функции $\overline{I}(U,Z)$:

$$C = \max_{p(U)} v_{symbol}\overline{I}(U,Z).$$

Для двоичного симметричного канала с помехами пропускная способность определяется по формуле

$$C = F[1 + (1-\beta)\cdot\ln(1-\beta) + \beta\cdot\ln\beta], \tag{6.12}$$

где β - вероятность ошибочного приема; $F = 1/\tau$ - символьная частота передачи данных (τ - длительность канального символа).

С учетом вышеизложенного, реальная степень загрузки канала характеризуется ***коэффициентом использования канала***:

$$\lambda = \underline{I}(U)/C, \tag{6.13}$$

где $\underline{I}(U)$ - производительность источника сообщений, а C - пропускная способность канала связи.

6.2.3 Теорема Шеннона для дискретного канала с помехами

Шенноном было доказано, что и в случае канала с помехами его пропускная способность определяет верхнюю достижимую границу скорости

[2]В случае канала без помех значение апостериорной энтропии равно нулю.

достоверной передачи информации по каналу. Таким образом, для дискретного канала с помехами теорему Шеннона возможно сформулировать следующим образом:

Теорема 2 (Теорема Шеннона для дискретного канала с помехами)**.** Для дискретного канала с помехами существует такой способ кодирования, при котором может быть обеспечена безошибочная передача всей информации, поступающей от источника сообщений, если только пропускная способность канала превышает производительность источника сообщений, т.е. выполняется условие:

$$v_{channel}[\log M - H(U|Z)] > v_{source}H(U). \tag{6.14}$$

Доказательство данной теоремы приводится, например, [Липкин] и может быть детально рассмотрено читателем. Отметим лишь основные следствия данной теоремы:

1 Теорема Шеннона для дискретного канала с помехами под оптимальным понимает кодирование, связанное с увеличием задержки передачи сообщения[3] на время $\tau_{latency} = 2T + T_0$.

2 Теорема Шеннона для канала с шумами не указыает конкретного способа кодирования, обеспечивающего достоверную передачу информации, а лишь доказывает принципиальное существование такого кода.

3 Чем выше длительность кодированной последовательности (и, соответственно, латентность) и чем меньше коэффициент использования канала, тем больше помехоустойчивость сообщений.

Теорема Шеннона для канала с помехами имела огромное значение для становления правильных воззрений на принципиальные возможности техники связи. До Шеннона считалось, что сколь угодно малую вероятность ошибки можно получить лишь при стремлении скорости передачи информации к нулю. Введя в рассмотрение кодирование последовательностей бесконечной длительности, Шеннон впервые показал, что принципиально существуют коды, которые обеспечивают сколь угодно малую вероятность ошибки при конечной скорости передачи информации, причем эти коды обладают сравнительно небольшой избыточностью, необходимой для компенсации вредного действия помех в канале связи.

[3]Латентности сообщения.

Практика 5 – каналы связи

6.3 Каналы связи

6.3.1 Взаимная информация, производительность канала связи

1 Энтропия дискретного источника сообщений всегда положительна. Дифференциальная энтропия, в свою очередь, может быть отрицательной. Может ли быть отрицательной полная взаимная информация двух непрерывных систем?

2 Лектор произносит в среднем около сорока шестибуквенных слов в минуту. Рассматривая его как источник дискретных сообщений, определить его производительность. Для простоты принять, что все буквы алфавита равновероятны и статистически независимы.

3 Двоичный источник с равновероятными элементами имеет производительность 1000 бит-с. При передаче по каналу в среднем один из переданных 100 символов искажается. Определить скорость передачи информации по данному каналу.

4 По двоичному симметричному канала связи с помехами передаются сигналы (x_1, x_2) с априорными вероятностями $p(x_1) = 3/4$; $p(x_2) = 1/4$. Из-за наличия помех вероятность правильного приема каждого из сигналов (x_1, x_2) уменьшается до $\alpha = 7/8$. Найти:
1. Среднее количество информации $I(x; y)$;
2. Пропускную способность канала связи $C = \max\limits_{p(x)} I(x; y)$.

5 Имеется источник информации с энтропией в единицу времени $H = 100$ бит и два канала связи: каждый из них может передавать в единицу времени 70 бит информации, при этом в результате

помехи, действующей на каждый из этих каналов, значение бита может быть заменено на противоположное с вероятностью 0.1. Вопрос - достаточна ли пропускная способность этих каналов для передачи информации, поставляемой источником?

6 Определить пропускную способность канала связи, если средняя мощность полезного сигнала равна S, полоса частот канала - F, а помехами являются тепловые шумы приемного устройства, имеющего температуру T°. Построить (качественно) график зависимости пропускной способности от полосы частот F.

7 Найти пропускную способность канала с амплитудно-импульсной модуляцией (АИМ), если число уровней сигнала равно 16, полоса частот исходного сигнала - F, сигнал $u(t)$ равномерно распределен в диапазоне $(-U_M, +U_M)$; при этом вероятность искажения, выражающая возможность перехода в соседний уровень, равна 5%.

Лекция 7

Непрерывные каналы связи

7.1 Непрерывные каналы связи и источники сообщений

7.1.1 Гауссова модель канала связи

Реальные непрерывные каналы связи являются сложными инерционными объектами с нелинейной передаточной характеристикой[1]. Для представления непрерывных каналов связи разработано множество моделей представления; в рамках данного курса лекций мы рассмотрим одну из наиболее примитивных, но и наиболее используемых для целей аналитического вывода моделей - *гауссов канал связи*. Итак, под ***гауссовым каналом связи*** понимают математическую модель канала связи, построенную при следующих допущениях:

1 Основные физические параметры канала являются известными детерминированными величинами;

2 Полоса пропускания ограничена верхней частотой в $F_{max} = W$ герц;

3 В канале действует аддитивный белый гауссовый шум[2]. Данный тип шума является случайным сигналом с равномерной спектральной плотностью[3], нормально распределенным значением амплитуды и аддитивным способом воздействия. АБГШ - тип случайного процесса с отсутствием корреляции между соседними временными от-

[1]Упрощенно - с нелинейной зависимостью между входом и выходом рассматриваемой динамической системы - канала связи.

[2]Здесь и далее - АБГШ. В английской нотации - AWGN (additive white gaussian noise).

[3]Спектральная плотность - форма представления распределения энергии сигнала на частотную плоскость.

счетами; в рамках данной лекции рассматриваемый канал также является каналом без памяти.

Аналогично случаю дискретного канала, ***пропускной способностью непрерывного канала связи*** называется максимально возможная скорость C передачи информации по данному каналу:

$$C = \max \overline{I}(Z, U), \tag{7.1}$$

где $\overline{I}(Z, U)$ - среднее количество информации, передаваемое в канале связи с учетом принимаемого ансамбля сообщений Z и передающего ансамбля U.

7.1.2 Дискретизация, квантование и отношение сигнал-шум

Для определения пропускной способности непрерывного гауссового канала связи необходимо вначале определить некоторые дополнительные понятия и теоремы из теории связи:

Дельта-функция[4] *определяется как функция, удовлетворяющая следующим условиям:*

- $\delta(x) = \begin{cases} +\infty, x = 0, \\ 0, x \neq 0. \end{cases}$
- $\int\limits_{-\infty}^{+\infty} \delta(x)dx = 1;$
- $\int\limits_{-\infty}^{+\infty} \delta(x)f(x)dx = f(0)$ - т.н. ***фильтрующее свойство*** дельта-функции для любой функции f.

Схематичное изображение дельта-функции приведено на рис. ниже:

[4]Синонимы - δ-функция; δ-функция Дирака; единичная импульсная функция. Дельта-функция позволяет записать пространственную плотность физической величины (массы, заряда, силы и т.п.), сосредоточенной или приложенной в одной точке.

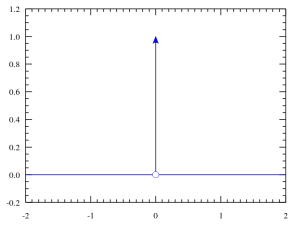

Рис. 7.1: Дельта-функция Дирака

Дискретизацией *называется процесс отображения непрерывного*[5] *сигнала в соответствующий ему дискретный, т.е. представленный в виде конечного множества отдельных отсчетов*[6] *для некоторых моментов времени.*

Равномерной дискретизацией *называется дискретизация с фиксированным временным шагом* τ_{sample} *между отсчетами.* Процесс представления непрерывного сигнала $U(t)$ на всей области определения в виде определенной на том же интервале функции $U_{discrete}$ из I отсчетов при помощи равномерной дискретизации исходного может быть определен с помощью фильтрующего свойства дельта-функции:

$$\underset{i=0,I-1}{U_{discrete}} = \int_{-\infty}^{+\infty} \delta(i \cdot \tau_{sample} t) f(t) dt \tag{7.2}$$

Процесс дискретизации отображен на рисунке ниже:

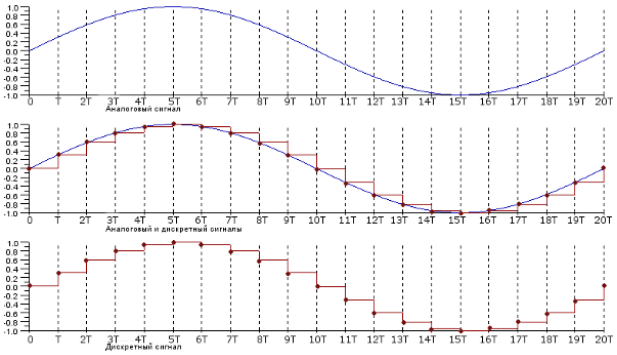

Рис. 7.2: Процесс дискретизации непрерывного сигнала

Кроме операции дискретизации для представления исходного непрерывного сигнала в виде дискретного сообщения производится также операция квантования:

[5] В радиотехнике более принято понятие *аналогового* сигнала.

[6] Мгновенных значений исходного сигнала; англ. эквивалент - *sample*.

Квантование *- операция разбиения диапазона значений непрерывного или дискретного сигнала на конечное число интервалов.*

Простейшим видом квантования является равномерное[7] квантование, когда область значения исходной функции разбивается на конечное множество равных интервалов, величина которых называется ***шагом квантования*** и для каждого входного значения выбирается ближайшее квантованное. Также достаточно часто применяется термины ***глубина дискретизации***, называемая также ***битностью*** или ***разрядностью***. Разрядность определяет количество бит, выражающих амплитуду сигнала. Пусть сигнал $U(t)$ имеет область значений $U_1; U_2$. Зададимся допустимым значением шага квантования U_{step}. В этом случае количество интервалов квантования определяется как округление в большую сторону до целого:

$$N_{step} = \frac{U_2 - U_1}{U_{step}} \tag{7.3}$$

Количество чисел, представляемых битностью сигнала, равно количеству интервалов квантования + 1. Таким образом, получаем значение требуемой битности сигнала:

$$N_{bitness} = round(\log_2 N_{step} + 1) \tag{7.4}$$

Результат совместного использования операций дискретизации и квантования к исходному сигналу называется его представлением в виде ***цифрового сигнала*** или ***цифровым представлением*** и продемонстрирован на рис. ниже.

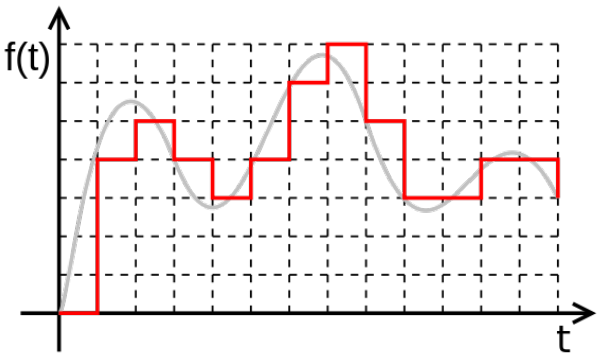

Рис. 7.3: Представление сигнала в цифровой форме

Перед рассмотрением одной из основных теорем теории связи - *теоремы отсчетов Котельникова* рассмотрим еще одно базовое понятие -

[7]Синонимы - однородное, линейное

отношение сигнал-шум и рассмотрим его базовый физический смысл.

Отношение сигнал-шум[8] *- безразмерная величина, равная отношению мощности полезного сигнала к мощности шума:*

$$SNR = \frac{P_{signal}}{P_{noise}} = \left(\frac{A_{signal}}{A_{noise}}^2\right), \tag{7.5}$$

где P - средняя мощность, а A - среднеквадратическое [9] значение амплитуды. Здесь и далее для сокращения выкладок будем обозначать $P_{signal} = S$, а $P_{noise} = N$.

7.2 Теорема Котельникова и пропускная способность непрерывных каналов связи

7.2.1 Теорема Котельникова

Теорема 3 (Теорема Котельникова[10])**.** Если аналоговый сигнал u(t) имеет ограниченный спектр (т.е. ограничен верхней частотой F_{max}, что при присутствии сигнала на всех частотах соответствует ширине полосы $W = F_{max}$), то данный сигнал может быть восстановлен однозначно и без потерь по дискретному сигналу, сформированному из исходного; при этом частота дискретизации должна быть более удвоенной F_{max}:

$$f_s \geq 2 \cdot F_{max}; \triangle t < \frac{1}{2 \cdot F_{max}} \tag{7.6}$$

[8]Англ. обозначение - ***SNR*** - signal-to-noise ratio.

[9]Среднеквадратическое значение (англ. эквивалент - RMS (root-mean-square) имеет синонимичные термины - *действующее* или *эффективное* значение. Для любой периодической функции (сигнала) среднеквадратическое значение функции $F = \sqrt{(\frac{1}{T}\int_0 T f^2(t)dt}$.

Если же функция задана в виде суммы гармоник: $f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} F_n cos(nx - \psi_n)$, то $F = \sqrt{\frac{a_0}{2}^2 + \sum_{n=1} \infty \frac{F_n^2}{2}}$.

[10]В англоязычной литературе - теорема Найквиста-Шеннона. Изначально теорема была сформулирована Гарри Найквистом в 1928 г. в работе «Certain topics in telegraph transmission theory» и является одной из основополагающих теорем в теории и технике цифровой связи. Приблизительно такие же результаты были опубликованы в том же году в Германии Карлом Купфмюллером. В СССР и России данная теорема традиционно связывается с именем Котельникова, независимо опубликовавшего аналогичные результаты в 1933 г. в своей работе «О пропускной способности эфира и проволоки в электросвязи».

, при этом непрерывный сигнал возможно восстановить по отсчетам имеющегося дискретного сигнала в виде ряда следующей формы:

$$\sum u(k \cdot \triangle t) \frac{sin(\pi f_s(t - k\triangle t))}{\pi f_s(t - k\triangle t)}. \tag{7.7}$$

Указанная теорема справедлива для идеального случая бесконечного сигнала, не имеющего во временной характеристике точек разрыва. Для реальных сигналов из теоремы Котельникова следует два следствия:

1 Любой реальный непрерывный сигнал может быть восстановлен *с какой угодно точностью* по своим дискретным отсчетам, взятым с частотой $f_s > 2 \cdot F_{max}$.

2 Если максимальная частота в сигнале превышает половину частоты прерывания, то способа восстановить сигнал из дискретного в аналоговый без искажений не существует.

Половина частоты дискретизации непрерывного сигнала называется ***частотой Найквиста*** и широко используется при решении задач цифровой обработки сигналов.

7.2.2 Пропускная способность и формула Шеннона

Случай белого гауссового шума в канале соответствует максимальной энтропии, вносимой шумом в полученное непрерывное сообщение, т.е. соответствует максимальной ненадежности канала по Шеннону при заданном отношении сигнал-шум. Таким образом, модель гауссового канала возможно записать следующим образом:

$$Z(t) = U(t) + n(t), \tag{7.8}$$

где $Z(t)$ - принимаемый сигнал; $U(t)$ - передаваемый; $n(t)$ - шум в канале связи. Ширина спектра сигнала $U(t)$, передаваемого по каналу, как уже было указано выше, равна W, Гц. Таким образом, согласно теореме Котельникова, сигнал полностью определяется дискретными отсчетами U_i, идущими с частотой $F_s = 2 \cdot W$.

Так как отсчеты белого шума являются некоррелированными; рассматриваемый канал является гауссовым без памяти и исходный источник сообщений также являются источником без памяти и с нормальным законом распределения сообщений[11], следовательно и дискретные отсчеты

[11] Что соответствует максимуму энтропии канала и источника, см. предыдущие лекции

принимаемого сигнала Z будут некоррелированными и количество информации, содержащейся в принимаемом сигнале, будет равно сумме количеств информации, содержащихся в его независимых отсчетах, следующих с частотой $2 \cdot F_{max}$. Количество информации о текущем значении передаваемого сигнала U_i, вносимое дискретным отсчетом принимаемого сигнала Z_i может быть представлено в виде разности априорной энтропии этого отсчета и его апостериорной энтропии при известном отсчете передаваемого сигнала. Заменяя разность энтропий разностью соответствующих дифференциальных энтропий, получаем:

$$h(Z_i, U_i) = H(Z_i) - H(Z_i|U_i). \tag{7.9}$$

Отсчеты принимаемого сигнала, представляющие собой сумму двух нормальных случайных сигналов $U(t)$ и $Z(t)$, распределены по нормальному закону с дисперсией $\sigma_Z^2 = \sigma_U^2 + \sigma_{AWGN}^2$. Апостериорная неопределенность принимаемого сигнала при известных значениях отсчетов передаваемого сигнала определяется значением шума, имеющего нормальное распределение с дисперсией σ_{AWGN}^2. Таким образом, разность дифференциальных энтропий в 7.9 в соответствии с формулой для дифференциальной энтропии нормальной величины [12], равняется

$$\log\left(\sigma_Z/\sigma_{AWGN}\right)$$

и информация, вносимая одним дискретным отсчетом сигнала, равна:

$$\begin{aligned} h(Z_i, U_i) &= \log\left(\sigma_Z/\sigma_{AWGN}\right) = \\ &= \log\sqrt{\frac{\sigma_U^2 + \sigma_{AWGN}^2}{\sigma_{AWGN}^2}} = \log\sqrt{1 + \frac{\sigma_U^2}{\sigma_{AWGN}^2}} = \\ &= \log\sqrt{1 + \frac{S}{N}} = \log\sqrt{1 + SNR}. \end{aligned} \tag{7.10}$$

Таким образом, пропускная способность канала, определяемая количеством информации, передаваемая по каналу в единицу времени при рассмотренных условиях, равна информации, содержащейся в $2 \cdot F_{max}$ независимых дискретных отсчетах сигнала:

$$C = 2 \cdot F_{max} \log\sqrt{1 + SNR} = F_{max} \cdot \log(1 + SNR). \tag{7.11}$$

[12] $h_{gauss}(U) = \log\sqrt{2\pi e}\sigma$

Формула 7.11 называется ***формулой Шеннона***.

Если же аддитивная помеха в канале связи не представляет собой АБГШ, то формула 7.11 занижает его пропускную способность.

7.2.3 Ограничения пропускной способности канала

К сожалению, из формулы 7.11 не следует возможность неограниченного роста пропускной способности канала при неограниченном расширении его полосы. Определим мощность шума N [Вт] через спектральную плотность мощности $N_0 = \frac{N}{W}$ [Вт/Гц] и полосу частот, на которой рассматривается шум. В этом случае максимально возможная полоса представляет собой предел:

$$C_{max} = \lim_{F_{max} \to \infty} \frac{\log_2 1 + \frac{S}{N_0 \cdot W}}{1/W}. \tag{7.12}$$

По правилу Лопиталя, при $\frac{1}{W} \longrightarrow 0$) получаем предельное значение пропускной способности канала:

$$C_{max} = 1.443 \cdot \frac{S}{N_0}. \tag{7.13}$$

При этом о характере зависимости $C_{max} = f(W)$ возможно судить по следующему графику, показывающему неперспективность увеличения полосы канала связи после некоторого предела.

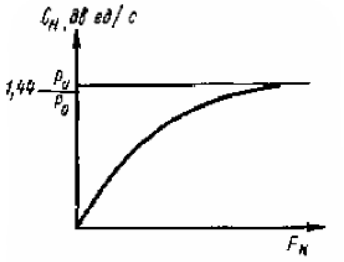

Рис. 7.4: Зависимость пропускной способности от предельной частоты

Лекция 8

О практическом определении помехоустойчивости и пропускной способности

8.1 Дополнения к формуле Шеннона

Непосредственно формулу Шеннона следует применять с достаточно большой осторожностью (по крайней мере, в части, касающейся построения реальных систем связи и передачи данных). Так, из данной формулы следует бесконечная производительность канала связи при отсутствии шума, что не является корректным.

Для уточнения применимости рассмотренного в предыдущих лекциях мат. аппарата рассмотрим далее некоторые понятия и теоремы, логически дополняющие формулу Шеннона и использующиеся в процессе проектирования и оценки параметров реальных систем и каналов связи.

8.1.1 Нормированное отношение сигнал-шум

Рассматриваемая в предыдущей лекции величина SNR представляет собой отношение средней мощности сигнала к средней мощности шума. К сожалению, данный показатель часто является неинформативным[1] На практике в цифровой связи используется нормированный показатель сигнал-шум - отношение энергии бита к спектральной плотности мощности шума:

$$SNR_{norm} = \frac{E_b}{N_0}.$$

[1]Так, данный показатель абсолютно неинформативен в т.н. широкополосных системах связи и передачи данных, которые будут рассматриваться в курсе «Вычислительные сети и системы телекоммуникаций».

E_b - энергия бита и ее возможно определить, как мощность сигнала S, умноженную на время передачи бита T_b. Т.к. время передачи бита и скорость передачи битов V_b взаимно обратны, $T_b = 1/V_b$. Таким образом, получаем следующее выражение для определения нормированного отношения сигнал-шум:

$$\frac{E_b}{N_0} = \frac{ST_b}{N/W} = \frac{S}{N}\left(\frac{W}{V_b}\right). \tag{8.1}$$

Одной из важнейших метрик качества в реальных системах цифровой связи является так называемая функция битовой ошибки BER[2] от нормированного отношения сигнал шум – $BER = f(E_b/N_0)$. Графики функции BER чаще всего имеют логарифмический масштаб по оси ординат (меняющийся от 0 до 0,5) и линейный по оси абсцисс (нормированное отношение сигнал-шум, измеряющееся в дБ). Пример графика такого рода показан на рис. ниже.

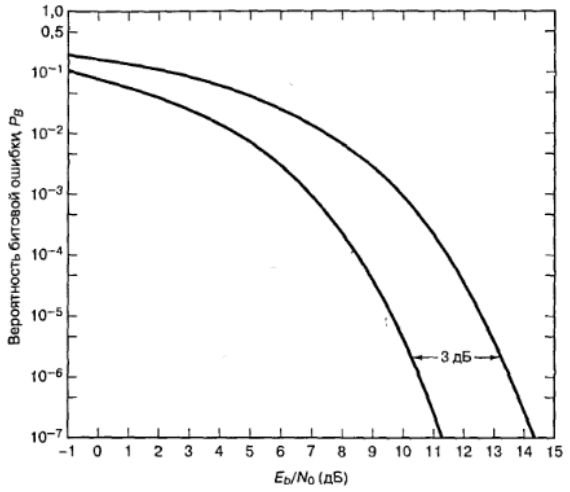

Рис. 8.1: Пример графика функции битовой ошибки

Итак, отношение E_b/N_0 представляет собой метрику, оценивающую качество различных систем - чем меньше требуемое отношение E_b/N_0 , тем эффективнее происходит процесс детектирования исходного сигнала при заданной вероятности ошибки.

8.1.2 Теорема Найквиста

Предельное значение пропускной способности, ограничивающее формулу Шеннона, дает теорема Найквиста для канала без шумов, но с фиксированным количеством уровней квантования:

[2]BER-bit error rate - вероятность битовой ошибки.

Теорема 4 (Теорема Найквиста)**.** Максимальная пропускная способность канала связи без шумов C при прохождении квантованного (с количеством уровней квантования N) сигнала определяется по следующей формуле[3]:

$$C = 2W \cdot \log N. \tag{8.2}$$

Пример 21. Определить максимальную пропускную способность телефонного канала с полосой пропускания в 3 кГц и отношением сигнал-шум в 30 дБ. Определить теоретическую пропускную способность данного канала при предположении отсутствия шумов, но условии, что канал адекватно описывается гауссовой моделью.

Решение.
1. Переведем отношение сигнал-шум из децибел в разы:

$$S/N_{dB} = 10 \lg S/N,$$

$$S/N = 10^{\frac{S/N_{dB}}{10}} = 10^3 = 1000.$$

Пропускная способность непрерывного канала с известным шумом и ограниченной полосой определяется по формуле Шеннона и равна:

$$C = W \cdot \log 1 + S/N = 3000 \cdot \log_2 1 + 1000 = 29901 \text{ bit}.$$

2. Пропускная способность данного канала в предположении, что он является каналом без шумов, может быть определена при фиксированном количестве уровней квантования N_{step}. Возьмем шаг квантования равный σ_N, в этом случае требуемое количество шагов квантования для покрытия диапазона от 0 до σ_S равно

$$N_{1\sigma} S/N + 1 = 1001$$

. В предположении, что исходный сигнал также представляет собой гауссовую случайную величину, обеспечим покрытие его динамического диапазона согласно правилу трех сигма:

$$N_{step} = 3 \cdot N_{1\sigma} = 3003.$$

[3]Известной также как формула Найквиста.

Таким образом, согласно формуле Найквиста, потенциальная пропускная способность такого канала равна:

$$C = 2W \cdot \log N_{step} = 6000 \cdot \log_2 3003 = 69313 \text{ bit}.$$

8.1.3 Предел Шеннона.

Кроме упомянутых выше ограничений на пропускную способность, существует еще одно фундаментальное ограничение возможностей канала связи. Это ограничение представляет собой нижнее предельное значение отношения сигнал-шум, при котором ни при какой скорости передачи нельзя осуществить безошибочную передачу информацию. Данное пороговое значение и называется пределом Шеннона.

Итак, нам известно, что выражение для пропускной способности канала связи с АБГШ определяется по формуле Шеннона:

$$C = W \cdot \log \left(1 + \frac{S}{N_0 W}\right).$$

Если битовая скорость передачи равна пропускной способности канала, то с помощью выражения 8.1 возможно записать следующее:

$$\frac{S}{N_0 C} = \frac{E_b}{N_0}. \tag{8.3}$$

Таким образом, формулу Шеннона возможно модифицировать следующим образом:

$$\begin{aligned} \frac{C}{W} &= \log \left(1 + \frac{E_b}{N_0} \cdot \frac{C}{W}\right), \\ 2^{C/W} &= 1 + \frac{E_b}{N_0}\frac{C}{W}, \\ \frac{E_b}{N_0} &= \frac{W}{C}(2^{C/W} - 1). \end{aligned} \tag{8.4}$$

Функция $\frac{W}{C} = f(SNR)$ называется нормированной пропускной способностью канала и измеряется в Гц/бит/с. График данной функции представлен на рис. ниже и отражает допустимые возможности реальных систем связи и передачи данных.

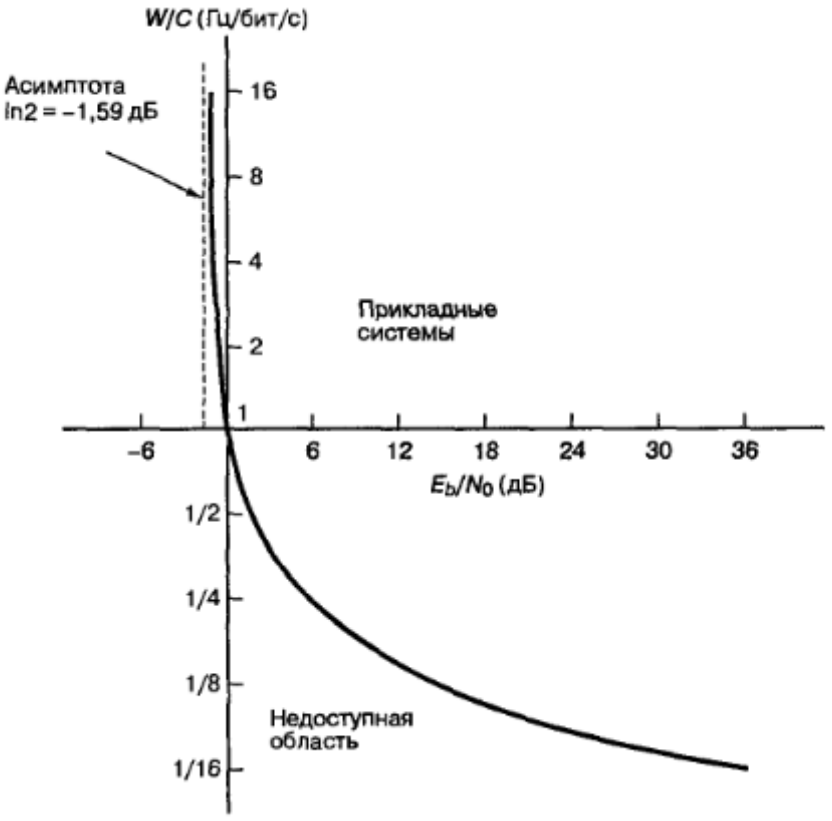

Рис. 8.2: Нормированная пропускная способность от нормированного SNR

С помощью формулы 8.4 и соотношения

$$\lim_{x \longrightarrow 0} (1+x)^1/x = e$$

рассчитаем значение предела Шеннона в заданной постановке – граничное значение E_b/N_0.

Итак, пусть

$$x = \frac{E_b}{N_0}\frac{C}{W}.$$

Тогда из уравнения 8.4 следует:

$$\frac{C}{W} = x \log_2 (1+x)^1/x$$

и

$$1 = \frac{E_b}{N_0} \log_2 1 + x^1/x.$$

. В пределе, при $C/W \longrightarrow 0$, получаем значение предела Шеннона в линейной и децибельной шкале:

$$\frac{E_b}{N_0} = \frac{1}{\log_2 e} = 0.693 = -1.6dB. \tag{8.5}$$

На настоящий момент в системах связи с использованием т.н. турбокодов достигнут эмпирический предел порядка -0.5 dB.

Пример 22 (Кажущееся противоречие с пределом Шеннона)**.** График зависимости битовой ошибки от E_b/N_0 обычно показывает плавный рост BER при увеличении E_b/N_0. Например, на рис. 8.1 видно, что в пределе при $E_b/N_0 \longrightarrow 0$, $BER \longrightarrow 0$. Таким образом, кажется, что всегда (при сколь угодно малом значении E_b/N_0) имеется ненулевая скорость передачи информации. На первый взгляд это *не согласуется* с величиной предела Шеннона $E_b/N_0 = -1.6dB$, ниже которого невозможна безошибочная передача информации. Показать способ разрешения кажущегося противоречия.

Решение. Величина энергии бита E_b, традиционно используемая для расчета каналов в прикладных системах - это энергия принятого сигнала, приходящаяся на *переданный символ*. Однако E_b в уравнении 8.4 - это энергия сигнала, приходящаяся на один бит *принятой информации*. Таким образом, для разрешения описанного выше противоречия следует учитывать потери информации, вызываемые помехами канала.

Практика 6 – дополнительные вопросы передачи информации

8.2 Дополнительные вопросы передачи информации

8.2.1 Квантование, дискретизация, сигнал-шум

1 По радиоканалу с полосой частот 100 кГц, в котором действует АБГШ со спектральной плотностью мощности $N_0 = 1.6 \cdot 10^{-3}$ Вт-кГц, передается сигнал $u(t)$, имеющий граничную частоту 10 кГц и среднюю мощность $S = 14$ Вт. Сколько времени займет передача сигнала по данному каналу?

2 Аналоговый сигнал считывается с частотой Найквиста $1/T$ и квантуется с использованием L уровней квантования, после чего полученный цифровой сигнал передается по каналу связи.
а). Покажите, что длительность T одного бита передаваемого двоично-кодированного сигнала должна удовлетворять условию $T \leq T_S/(\log_2 L)$.
б). Когда имеет место равенство?

3 Определите максимальную частоту дискретизации, необходимую для выборки и точного восстановления сигнала $x(t) = sin(6280t)/(6280t)$.

4 Рассмотрим аудиосигнал, спектральные компоненты которого ограничены полосой частот от 300 до 3300 Гц. Предположим, что для создания цифрового сигнала используется частота дискретизации в 8000 выборок/с. Предположим также, что отношении пиковой мощности сигнала к средней мощности сигнала должно быть равно 30 дБ.

Чему равно минимальное число уровней квантования с равномерным шагом и минимальное число битов на выборку?

5 В 1962 году компания AT&T первой предложила цифровую телефонную линию связи, названную службой T1. Каждый кадр T1 разбивается на 24 канала (интервала времени). Каждый интервал при этом содержит 8 бит (одна речевая выборка) и один бит для выравнивания. Каждый кадр считывается с частотой Найквиста в 8000 выборок/с, а ширина полосы, используемая для передачи составного сигнала, равна 386 кГц.
Определите для этой схемы эффективность использования полосы (в бит/с/Гц).

6 Опишите два точных способа сравнения различных кривых, описывающих зависимость вероятности появления ошибочного бита от отношения E_b/N_0.

7 Предложите график, характеризующий помехоустойчивость канала связи для случая трех равновероятных АБГШ-помех (различной мощности) в канале связи, для следующих случаев: а). Вероятности которых образуют полную группу (события возникновения помех являются несовместными). б). Помехи могут возникать в любой комбинации, в том числе возможно и отсутствие всех типов помех.

Тема III
Эффективное кодирование

Тема 3 – список литературы

1 *Дмитриев В.И.* Прикладная теория информации. М.: Букинист, 1989. - 332 с.

2 *Аршинов М.Н., Садовский Л.Е.* Коды и математика (рассказы о кодировании). М.: Наука, 1983 г. - 144 с.

3 *Прохоров В.С.* Теория информации - курс лекций. - 124 с.

4 *Лидовский В.В.* Теория информации - учебное пособие. М.: Спутник+, 2004. - 111 с.

5 *Липкин И.А.* Статистическая радиотехника. Теория информации и кодирования. М.: Вузовская книга, 2002. - 216 с.

6 *Скляр Б.* Цифровая связь. Теоретические основы и практическое применение. М.: изд. дом Вильямс, 2003. - 1104 с.

7 *Шульгин В.И.* Основы теории передачи информации. Ч.I. Экономное кодирование - учебное пособие. Харьков: изд. ХАИ, 2003. - 102 с.

8 *Орлов В.А., Филиппов Л.И.* Теория информации в упражнениях и задачах. М.: Высшая Школа, 1976. - 136 с.

9 *Хэмминг Р.В.* Теория кодирования и теория информации. М.: Радио и связь, 1985. 176

10 *Шеннон К.* Работы по теории информации и кибернетике. Теория связи в секретных системах. М.: ИЛ, 1963. с. 333-369.

Лекция 9

О кодировании. Статистическое кодирование

9.1 Понятие кодирования. Типы кодирования

Предыдущие темы представленного курса лекций были посвящены основам теории информации, необходимым для проведения базовых расчетов и понимания принципов функционирования информационных систем различного рода. Вторая половина данного курса дает понятие о принципах, типах и методах кодирования в информационных системах: системах передачи, хранения и обработки информации.

9.1.1 Позиционное кодирование. Код Грея

Сообщения, передаваемые произвольным источником сообщений, чаще всего не удовлетворяют тем или иным требованиям, предъявляемым со стороны канала связи - требованиям по скорости передачи данных и требованиям по помехоустойчивости передаваемых сообщений. Для адаптации сообщений к параметрам канала связи применяется методы кодирования - ***преобразующего***[1], ***эффективного кодирования***[2], минимизирующие избыточность сообщений и методы помехоустойчивого кодирования, повышающие их помехоустойчивость.

Итак, что же такое кодирование?

Кодирование *в общем случае - это преобразование алфавита исходного сообщения* $U\{u_i\}, i = 1 \dots M$ *в алфавит кодовых символов*

[1]Осуществляющего преобразование алфавитов с какой-либо целью.

[2]Здесь и далее в качестве синонимичного используется термин «*кодирование источника*»), поскольку методы эффективного кодирования применяются преимущественно для минимизации избыточности сообщений источника информации

$R\{r_j\}, j = 1 \ldots K$.

При этом обычно, но не обязательно, $K < M$ или даже $K << M$.

Примером преобразующего кодирования является запись исходного сообщения, представляемого[3] в виде числа, в выбранной позиционной системе счисления разрядности m:

$$u_i = \sum_{k=0}^{n-1} r_k m^k = r_{n-1}m^{n-1} + r_{n-2}m^{n-2} + \ldots + r_0 m^0. \qquad (9.1)$$

Достаточно очевидно, что чем больше основание системы счисления, тем меньшее число разрядов требуется для представления данного сообщения, следовательно, требуется меньшее время для передачи[4] К сожалению, с ростом разрядности существенно повышаются требования к разрешающим устройствам и снижается помехоустойчивость итогового алфавита. Наиболее простой для реализации и традиционной для современных вычислительных машин является двоичная система: в данном случае задача различения сигналов в общем сводится к так называемой *задаче обнаружения*[5]. (есть или нет импульс), являющейся минимальной по вычислительной сложности и обеспечивающей наибольшую уверенность приема.

Крайне интересным типом преобразующего кодирования, не связанного с преобразованием систем счисления, являются коды с циклическим преобразованием алфавитов. Ярким примером таких кодов является так называемый ***код Грея***, используемый ранее в АЦП с аналоговым выбором вариантов итоговой строки[6], в которых данный код позволяет свести к единице младшего разряда ошибку неоднозначности при считывании[7]. На следующем рис. описаны комбинации кода Грея для десятичных чисел от 0 до 15.

---

[3]Без ограничения общности.

[4]При одинаковой длительности канального символа - т.е. сообщения итогового алфавита - для любого значения размерности m.

[5]Термин статистической радиотехники; в рамках данного курса детально не раскрывается.

[6]АЦП – аналого-цифровой преобразователь. Англ. эквивалент - ADC.

[7]В данном коде два соседних значения различаются только в одном разряде. Де-факто код Грея также представляет собой специфичную систему счисления. Описываемый двоичный код Грея является наиболее часто используемым на практике, хотя в общем случае существует бесконечное множество кодов Грея для систем счисления с любым основанием. Под кодом Грея здесь и далее будем понимать именно указанный двоичный код.

Число в десятичном коде	Код Грея	Число в десятичном коде	Код Грея	Число в десятичном коде	Код Грея
0	0000	6	0101	11	1110
1	0001	7	0100	12	1010
2	0011	8	1100	13	1011
3	0010	9	1101	14	1001
4	0110	10	1111	15	1000
5	0111				

Рис. 9.1: Комбинации кода Грея для десятичных чисел от 0 до 15

Пример использования кода Грея объясним на примере патента самого Фрэнка Грея, посвященного созданию нового типа АЦП на базе электронно-лучевой трубки. Итак, схема работы указанного АЦП была следующей:

Измеряемое АЦП напряжение отклоняет луч ЭЛТ по вертикали - задает строчку отображаемых значений. Основная проблема указанного АЦП была в следующей - если луч оказывался около границы двух строк, на выходе получались результаты, сильно отличающиеся друг от друга в каждый момент времени. Идея Грея заключалась в том, что если строки отличаются только одним битом, то и погрешность АЦП будет составлять только один младший разряд.

9.2 Методы эффективного кодирования

9.2.1 Статистическое кодирование

Хотя оптимальное кодирование, используемое для доказательства теоремы Шеннона в дискретном канале без помех является недостижимым[8], близкие к данному пределу результаты возможно получить и для сообщений (последовательностей символов) ограниченной длины, используя т.н. *статистическое кодирование.*

Оптимальным статистическим кодированием *называется кодирование, при котором обеспечивается распределение времени на передачу отдельных символов алфавита в зависимости от априорных вероятностей их появления.*

[8]Условие данного оптимального кодирования - укрупнение кодирующего алфавита до бесконечности и, соответственно, бесконечная длина исходящего сообщения.

Действительно, физическую сущность достижения предела производительности канала связи по определению возможно трактовать как постоянство скорости передачи информации (равной максимально достижимому значению) на элементарных интервалах, отводимых на передачу каждого символа (при этом все символы несут одинаковую информацию и имеют одинаковую длительность). Однако данного условия можно добиться и при использовании алфавита с неравновероятными, но независимыми символами, если время, отводимое на передачу отдельных символов, принимать пропорциональным доставляемой ими информации, т.е. выбирать из условия:

$$t_j = -\frac{\log_2 P_j}{C_k}, \tag{9.2}$$

где C_k - пропускная способность канала связи.

Для случая оптимального статистического кодирования среднее время передачи одного символа, при котором обеспечивается пропускная способность канала C_k равно:

$$t_{av} = \sum_{i=1}^{M} P_i t_i = -\frac{1}{C_k} \sum_{i=1}^{M} P_i \log_2 P_i = \frac{H(U)}{C_k}, \tag{9.3}$$

т.е. соответствует времени на передачу одного символа, отвечающему той же пропускной способности при идеальном кодировании по Шеннону.

Необходимым условием для использования любого типа оптимального статистического кодирования является независимость символов алфавита. Примером мер для ослабления корреляционных связей между символами алфавита является преобразующее кодирование, осуществляющее укрупнение алфавита канала. Указанное укрупнение сводится к тому, что блок из L символов исходного сообщения (от источника сообщений) объединяется в один символ алфавита кодирующего сообщения - при этом преобразовании статистические связи между символами практически устраняются.[9]

Рассмотрим далее основные виды оптимального статистического кодирования, используемые на практике.

[9]К сожалению, даже при достаточно большом укрупнении алфавита уменьшения избыточности не происходит – исходные последовательности обычно все еще являются слишком короткими для выделения существенной группы нетипичных последовательностей, вероятностью появления которых можно было бы пренебречь, исключив их из укрупненного алфавита.

9.2.2 Кодирование Шеннона-Фано

Кодирование методом Шеннона-Фано является одним из простейших методов статистического кодирования. Кодирование по Шеннону-Фано производится в соответствии с следующим алгоритмом:

1 Все символы из исходного алфавита U записывают в порядке убывания вероятностей.

2 Всю совокупность символов разбивают на две примерно равные по сумме вероятностей группы: одной из них (в группе могут быть один или более символов) присваивают значение «1», другой - «0».

3 Каждую из этих групп снова разбивают (если это возможно) на две части и каждой из частей присваивают значения «1» или «0». Данный пункт повторяется итеративно и рекурсивно до разбиения всех групп.

Указанную процедуру кодирования легче всего осуществлять при помощи построения т.н. *кодового дерева*, построение кода с помощью которого показано на следующем рисунке.

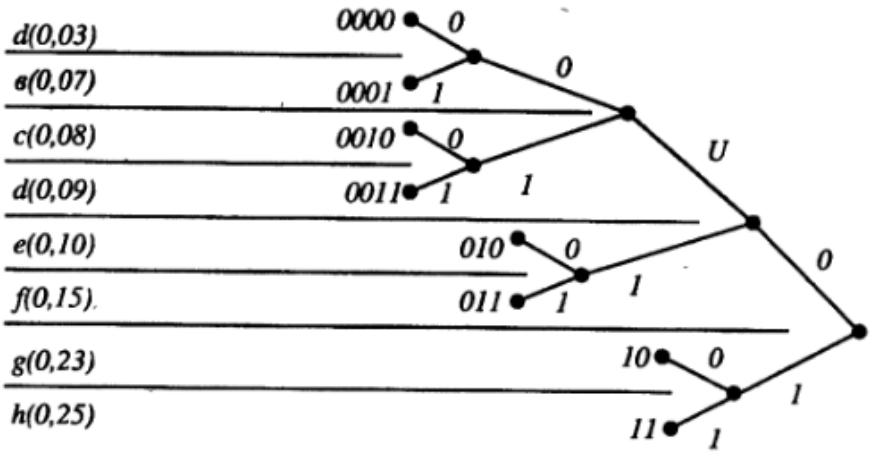

Рис. 9.2: Пример построения кода Шеннона-Фано с помощью кодового дерева

Для полученного кода среднее число двоичных символов, приходящихся на одну букву, равно:

$$\overline{n} = \frac{1}{8}\sum_{i=0}^{8} n_i = \frac{26}{8} = 3.25. \tag{9.4}$$

9.2.3 Кодирование по Хаффману

Простейшим подмножеством кодов, на основе которых может выполняться сжатие данных, являются коды без памяти. Для определения понятия «код без памяти» введем определение т.н. *префиксного множества*[10]:

Префиксным множеством *двоичных последовательностей U называется конечное множество двоичных последовательностей, таких, что ни одна последовательность в этом множестве не является префиксом, или началом, никакой другой последовательности из U.*

В коде без памяти *каждый символ в кодируемом векторе данных заменяется кодовым словом из префиксного множества двоичных последовательностей или слов.*

Алгоритм Хаффмена реализует идею оптимального статистического кодирования с использованием аппарата префиксных множеств и функционирует согласно следующему алгоритму:

1 Все символы алфавита выписываются в ряд в порядке возрастания (либо убывания) вероятности их появления в исходном сообщении.

2 Два символа с наименьшими вероятностями появления объединяются в новый составной символ, вероятность которого полагается равной сумме вероятностей составляющих символов. В результате итеративного рекурсивного выполнения данного пункта строится кодовое дерево, каждый узел которого имеет суммарную вероятность всех узлов, находящихся ниже него.

3 От вершины дерева прослеживается путь к каждому листу, помечая направление на каждом разветвлении (например – вниз - 0; вверх - 1). Полученные последовательности для каждого листа и есть кодовые слова, соответствующие исходным символам.

Построение кодового дерева для алгоритма Хаффмана для шестизначного множества показано на рис. ниже.

[10]Более подробно префиксные множества будут рассмотрены в следующей лекции.

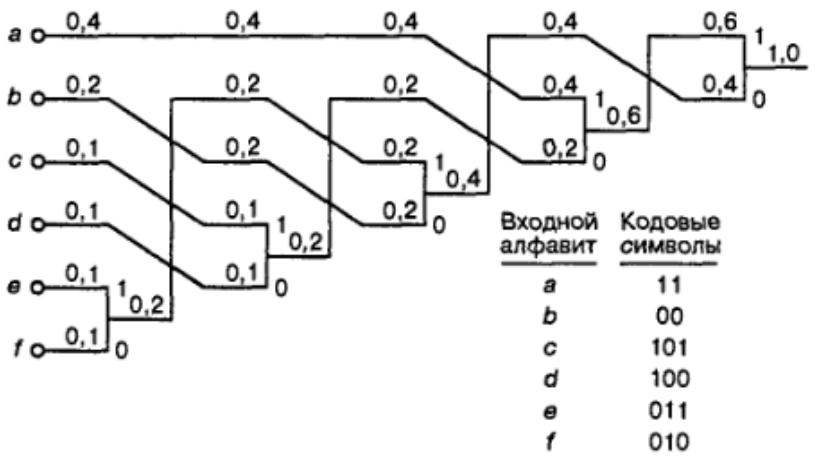

Рис. 9.3: Пример построения кода Хаффмана с помощью кодового дерева

9.2.4 Арифметическое кодирование

Алгоритм кодирования Хаффмана не способен передавать на каждый символ сообщения менее одного бита информации, что является его достаточно серьезным недостатком (пример – в сообщении, состоящем из нулей и единиц, единицы встречаются в 10 раз чаще нулей). Одной из наиболее популярных схем кодирования, позволяющих кодировать некоторые символы менее, чем одним битом, является методология ***арифметического кодирования***, детально проработанная в 70-е года XX века[11].

Основная идея арифметического кодирования заключается в том, чтобы присваивать битовые коды не отдельным символам исходного сообщения, а последовательностям данных символов. Определим алгоритм арифметического кодирования согласно следующей схеме:

1 Введем понятие **рабочего интервала** – полуинтервала $[a; b)$ с расположенными на нем точками; при этом точки расположены таким образом, что длины образованных ими отрезков равны вероятностям появления исходных символов. На первом шаге алгоритма $a = 0;\ b = 1$.

2 Основной шаг алгоритма: для кодируемого символа ищется соответствующий участок на рабочем интервале. Указанный участок рекурсивно становится новым рабочим интервалом (т.е. мы переразби-

[11]Наиболее известная версия разработана И.Уиттеном с соавторами.

ваем с использованием известных значений вероятностей символов). Данная операция выполняется до конца исходного сообщения.

3 Результатом кодирования исходного сообщения является любое число (а также длина его битовой записи), выбранное из финального рабочего интервала[12] При декодировании происходит сначала определение первого символа, затем нормализация рабочего интервала, соответствующего второму символу, затем нормализация рабочего интервала для третьего символа и т.д.

Рассмотрим алгоритм арифметического кодирования на примере, которым и закончим данную лекцию.

Пример 23. Пусть необходимо закодировать входное сообщение SWISS_MISS с заданными вероятностями появления символов, показанными на рис. ниже:

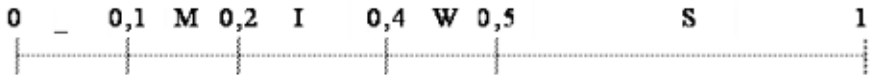

Рис. 9.4: Вероятности появления символов для арифметического кодирования

Решение. Процесс кодирования начинается со считывания первого символа входного потока и выбора нового рабочего интервала, соответствующего данному символу. В данном случае для первого символа S получаем диапазон [0.5, 1). После этого считывается новый символ W, которому соответствует поддиапазон [0.4, 0.5) уже во втором рабочем интервале и т.д. Схема преставления новых границ символа W представлена на рис. ниже.

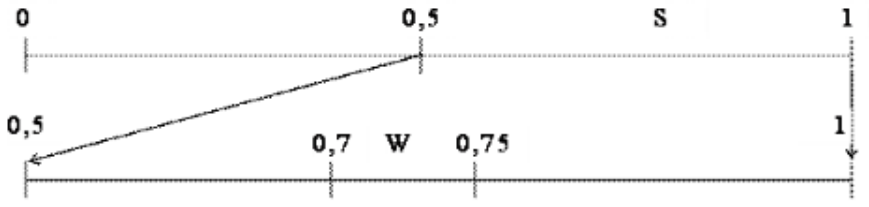

Рис. 9.5: Новый рабочий интервал

Указанная схема может быть записана в виде следующих формул:

[12]Чаще всего берется левая граница интервала.

$$\text{NewHigh} = \text{OldLow} + (\text{OldHigh-OldLow})*\text{HighRange(X)}, \text{NewLow} = \text{OldLow} + (\text{OldHigh-OldLow})*\text{LowRange(X)},$$

где OldLow – нижняя граница интервала, в котором представляется текущий символ; OldHigh – верхняя граница интервала; HighRange(X) – исходная верхняя граница кодируемого символа; LowRange(X) – исходная нижняя граница кодируемого символа.

Повторяя итеративно данный процесс, получаем значение последнего рабочего интервала [0.71753375, 0.717535). В качестве результирующего кода берется значение левой границы отрезка - 0.71753375, из которых достаточно записать в битовой форме число 71753375, кодируемое лишь 27 битами (27 битами кодируется число от 0 до $2^27 = 134217727$).

Теперь рассмотрим возможность восстановления закодированной информации по восьми цифрам 71753375 и известным интервалам символов. Первая из восьми цифр – это 7, т.е. 0,7. Она принадлежит одному из заданных интервалов [0.5, 1), который соответствует символу S. Поэтому первый декодированный символ – это S. Теперь вернемся к рис. 3 и заметим, что второй символ был представлен в интервале символа S, т.е. [0.5, 1). Но для удобства декодирования его лучше представить в исходном интервале [0, 1). Для этого достаточно интервал [0.5, 1) увеличить до начального, т.е. умножить на два и границы сдвинуть на величину $0.5 \cdot 2 = 1$.

Применяя данную схему к числу 0.71753375, получаем нижнюю границу следующего закодированного символа как будто он был начальным при кодировании:

$$0.71753375 * 2 - 1 = 0.4350675.$$

Полученное значение принадлежит диапазону [0.4, 0.5), который соответствует символу W. Затем, также полученное число 0.4350675 следует нормировать, что в общем случае выполняется по формуле:

$$\text{Code} = (\text{Code-LowRange(X)})/(\text{HighRange(X)-LowRange(X)}),$$

где Code – текущее значение кода.

Итоговая схема декодирования для данного примера приведена на рис. ниже.

Символ	Code-Low	Область	
S	0.71753375 – 0.5	= 0.21753375	/ 0.5 = 0.4350675
W	0.4350675 – 0.4	= 0.0350675	/ 0.1 = 0.350675
I	0.350675 – 0.2	= 0.150675	/ 0.2 = 0.753375
S	0.753375 – 0.5	= 0.253375	/ 0.5 = 0.50675
S	0.50675 – 0.5	= 0.00675	/ 0.5 = 0.0135
_	0.0135 – 0	= 0.0135	/ 0.1 = 0.135
M	0.135 – 0.1	= 0.035	/ 0.1 = 0.35
I	0.35 – 0.2	= 0.15	/ 0.2 = 0.75
S	0.75 – 0.5	= 0.25	/ 0.5 = 0.5
S	0.5 – 0.5	= 0	/ 0.5 = 0

Рис. 9.6: Общая схема декодирования для исходного сообщения

Лекция 10

Неравенство Крафта. Словарные методы кодирования

10.1 Эффективные методы кодирования

10.1.1 Неравенство Крафта-Макмиллана

Для получения некоторых общих результатов для эффективных методов кодирования введем более строго понятие «кодирование» и сопутствующую терминологию:

Кодирование - это процесс отображения множества U **кодируемого алфавита** в множество Z **кодирующего алфавита.** При этом элементы указанных множеств называются **символами**, а строки (последовательности символов конечной длины) - **словами**.

Кодом *C для алфавита U называется функция C, которая для каждого символа u_i из U указывает слово $C(u_i)$, кодирующее этот символ.*

Префиксным кодом *называется такой код, в котором ни одно из кодовых слов не является префиксом никакого другого кодового слова.*

Полным *называется такой префиксный код, для которого добавление любого нового кодового слова (в заданном алфавите) нарушает свойство префиксности.«* ***Разделимым***[1] *называется такой код, в котором никаким двум словам кодируемого алфавита не может быть сопоставлено одно и то же кодовое слово.* ***Любой префиксный код является разделимым.***

[1]Или однозначно декодируемым.

Рассмотрим далее т.н. теорему Макмиллана, известную также под названием неравенства Крафта-Макмиллана[2] Указанная теорема в теории кодирования дает необходимое и достаточное условие существования разделимых и префиксных кодов, обладающих заданным набором длин кодовых слов.

Теорема 5 (Теорема Макмиллана)**.** Пусть заданы кодируемый и кодирующий алфавиты, состоящие из n и d символов, соответственно, а также заданы желаемые длины кодовых слов: $l_1, l_2, \ldots, l_n$. В этом случае необходимым и достаточным условием существования разделимого и префиксного кодов, обладающих заданным набором длин кодовых слов, является выполнение неравенства:

$$\sum_{i=1}^{n} d^{-l_i} \leq 1. \tag{10.1}$$

Указанную теорему мы примем без доказательства, но рассмотрим при этом доказательство ее частного случая - неравенства Крафта для префиксных кодов:

[2]Впервые данное неравенство было выведено Леоном Крафтом в своей магистерской дипломной работе в 1949 году, однако он рассматривал только префиксные коды, поэтому при обсуждении префиксных кодов это неравенство часто называют просто *неравенством Крафта*. В 1956 году Броквэй Макмиллан доказал необходимость и достаточность этого неравенства для более общего класса кодов - разделимых кодов.

Теорема 6 (Неравенство Крафта)**.** Для любого префиксного кода C, отображающего произвольный алфавит U размерности M на 2-ичное множество Z, длины кодовых слов должны удовлетворять неравенству

$$\sum_{i=1}^{M} 2^{-l_i} \leq 1. \tag{10.2}$$

Доказательство.
1). Рассмотрим непрерывный отрезок (континуум) $[0;1]$ на числовой прямой.
2). Разделим его пополам, при этом левую половину обозначим M_0, а правую M_1.
3). Поделим M_0 пополам и обозначим его левую половину за M_{00}, а правую за M_{01}. Аналогично, для M_1 получаем M_{10} и M_{11}.
4). Указанные действия выполняем до тех пор, пока длина индекса полученного отрезка M_j не превосходит $\max(l_1, l_2, \ldots, l_M)$.
Далее, заметим следующее:

- Любому кодовому слову C_j сопоставлен свой отрезок M_{C_j} (например, кодовому слову 1011 соответствует отрезок M_{1011});
- Длина отрезка M_{C_i} равна 2^{-l_i} (например, M_0 имеет длину 0.5, а M_{00} - соответственно, 0.25.
- Если кодовое слово x является префиксом кодового слова y, то отрезок M_x содержит M_y (например, кодовое слово 01 является префиксом 0111, а отрезок M_{01} содержит M_{0111} - это его самая правая четверть.

Теперь рассмотрим префиксный код C - так как ни одно из кодовых слов не является префиксом никакого другого кодового слова, то никакие два отрезка не пересекаются.
Если на отрезке $[0;1]$ выбрать некоторое количество непересекающихся отрезков, то очевидно, что сумма их длин не превзойдет 1, то есть

$$\sum_{i=1}^{M} M_{C_i} \leq 1.$$

Таким образом, получаем искомое неравенство:

$$\sum_{i=1}^{M} M_{C_i} = \sum_{i=1}^{M} 2^{-l_i} \leq 1.$$

Указанная теорема может быть обобщена на случай k-ичного кодирующего алфавита следующим образом:
1). Отрезок $[0;1]$ необходимо делить не на 2, а на k равных частей;
2). Непосредственно само неравенство принимает вид

$$\sum_{i=1}^{M} k^{-l_i} \leq 1. \tag{10.3}$$

Наконец, если неравенство 10.2 переходит в строгое равенство, то такой код называется **компактным** и обладает наименьшей среди кодов с данным алфавитом длиной, то есть является оптимальным.

10.1.2 Вектор Крафта и код Хаффмана

Вектор $L = [l_1, l_2, \ldots, l_M]$ длин кодовых слов называется **вектором Крафта**.

Пусть необходимо разработать код без памяти для сжатия вектора данных $X = [x_1, x_2, \ldots, x_n]$ с алфавитом U размерностью в M символов. Введем в рассмотрение т.н. **вектор частот** $F = [F_1, F_2, \ldots, F_M]$, где F_i – количество появления i-го наиболее часто встречающегося символа из U в X. Закодируем X кодом без памяти, для которого вектор Крафта $L = [L_1, L_2, \ldots, L_M]$. В этом случае длина двоичной кодовой последовательности $B(X)$ на выходе кодера составит:

$$L \cdot F = L_1 \cdot F_1 + L_2 \cdot F_2 + \ldots + L_M \cdot F_M. \tag{10.4}$$

Очевидно, что оптимальным является код с $\min B(X)$. Именно рассмотренный выше алгоритм кодирования Хаффмена представляет собой эффективный способ поиска оптимального вектора Крафта.

10.2 Словарные методы кодирования

Рассмотренные ранее методы Шеннона-Фано, Хаффмана и арифметического кодирования называются *статистическими методами*. Рассматриваемые далее методы называются словарными и носят менее математически обоснованный, но более практичный характер.

Практически все словарные методы кодирования основываются на семействе алгоритмов LZ-сжатия, принадлежащих перу израильских ученых Зиву и Лемпелу. Общая суть данных методов состоит в следующем:

фразы в сжимаемом тексте заменяются указателем на то место, где они в данном тексте ранее появлялись.

Все словарные методы кодирования возможно разбить на две группы - основывающиеся на алгоритме LZ77 (1977 г.) и на алгоритме LZ78.

В рамках первой группы повторяющиеся цепочки символов заменяются указателями на предыдущие повторения. Наиболее совершенным представителем данной группы является алгоритм LZSS, опубликованный в 1982 Сторером и Шимански.

В рамках второй группы в дополнение к исходному словарю источника создается словарь фраз, представляющих собой повторяющиеся комбинации символов исходного словаря, встречающиеся во входных данных. Рассмотрим подробнее указанные типы кодирования.

10.2.1 Группа методов LZ77

LZ77 заменяет ссылками второе и последующее вхождение некоторой строки символов на ее первое вхождение. При этом алгоритмом используется скользящее по сообщению окно, разделенное на две неравные части. Первая, большая по размеру, включает уже просмотренную часть сообщения. Вторая, существенно меньше, представляет собой буфер, содержащий еще незакодированные символы входного потока. Обычно размер окна составляет несколько килобайт, а размер буфера - не более 100 айт. Алгоритм пытается найти в большей части окна (называемой словарем) фрагмент, совпадающий с содержимым буфера.

В результате работы алгоритм LZ77 выдает коды, состоящие из трех элементов:

- Смещение в словаре относительно начала подстроки, совпадающей с началом содержимого буфера.
- Длина данной подстроки.
- Первый символ буфера, следующий за подстрокой.

Пример 24. Закодировать по алгоритму LZ77 строку «КРАСНАЯ КРАСКА».

СЛОВАРЬ(8)	БУФЕР(5)	КОД
"........"	"КРАСН"	<0,0,'К'>
".......К"	"РАСНА"	<0,0,'Р'>
"......КР"	"АСНАЯ"	<0,0,'А'>
".....КРА"	"СНАЯ "	<0,0,'С'>
"....КРАС"	"НАЯ К"	<0,0,'Н'>
"...КРАСН"	"АЯ КР"	<5,1,'Я'>
".КРАСНАЯ"	" КРАС"	<0,0,' '>
"КРАСНАЯ "	"КРАСК"	<0,4,'К'>
"АЯ КРАСК"	"А...."	<0,0,'А'>

Рис. 10.1: Пример кодирования по алгоритму LZ77

Декодирование кодов LZ77 проще их получения, т.к. при этом не требуется осуществлять поиск в словаре. LZ77 также обладает следующими недостатками:

- С ростом размеров словаря скорость работы кодера пропорционально замедляется.

- Кодирование одиночных символов крайне неэффективно.

10.2.2 Группа методов LZ78

LZ78 не использует метод скользящего окна, а хранит словарь из уже просмотренных фраз. При старте алгоритма данный словарь содержит только одну пустую строку. Алгоритм считывает символы сообщения до тех пор, пока накапливаемая подстрока входит целиком в одну из фраз словаря. Как только эта строка перестает соответствовать хотя бы одной фразе словаря, алгоритм генерирует код, состоящий из индекса строки в словаре, которая до последнего введенного символа содержала входную строку, и символа, нарушившего совпадение. Затем в словарь добавляется введенная подстрока. Если словарь уже заполнен, то из него предварительно удаляют менее всех используемую в сравнениях фразу.

Пример 25. Закодировать по алгоритму LZ78 строку «КРАСНАЯ КРАСКА», используя словарь длиной 16 фраз.

ВХОДНАЯ ФРАЗА (В СЛОВАРЬ)	КОД	ПОЗИЦИЯ СЛОВАРЯ
""		0
"К"	<0,'К'>	1
"Р"	<0,'Р'>	2
"А"	<0,'А'>	3
"С"	<0,'С'>	4
"Н"	<0,'Н'>	5
"АЯ"	<3,'Я'>	6
" "	<0,' '>	7
"КР"	<1,'Р'>	8
"АС"	<3,'С'>	9
"КА"	<1,'А'>	10

Рис. 10.2: Пример кодирования по алгоритму LZ78

10.2.3 RLE и дифференциальное кодирование

В завершение лекции рассмотрим два простейших метода сжатия данных.

Сжатие RLE (Run Length Encoding) - простейший метод сжатия данных, основывающийся на замене последовательности символов специальным кодом-маркером, указывающим, сколько раз повторить следующий байт. Данный метод сжатия крайне неэффективен в общем случае, но дает неплохие результаты для изображений.

Дифференциальное кодирование основано на том факте, что для многих типов данных разница между соседними отсчетами несущественна, даже если сами отсчеты достаточно велики по значению (пример - разница в цвете соседних пикселов изображения). Таким образом, данные при этом методе кодирования кодируются блоками, различие в которых не превышает определенного порога (определяемого требуемым количеством бит для кодирования); первый символ блока кодируется полноценным количеством бит, последующие - сокращенным, достаточным лишь для представления разницы между соседними данными.