Уральский федеральный университет

ИНФОРМАТИКА

Старший преподаватель департамента информационных технологий и автоматики
Шеклеин Алексей Александрович

Уральский федеральный университет

ТЕМА 6. ЭФФЕКТИВНОЕ КОДИРОВАНИЕ ИНФОРМАЦИИ

- Сущность проблемы кодирования:
  - Коды
  - Длина кода
  - Свойства кода
  - Моментальные коды
  - Кодовые деревья
- Неравенство Крафта
- Средняя длина кода и энтропия
- Первая теорема Шеннона

Уральский федеральный университет

КОДЫ

- Что такое код?
- Длина кода
- Свойства кода
- Моментальные коды
- Кодовые деревья
- Неравенство Крафта
- Средняя длина кода
- Первая терема Шеннона

Уральский федеральный университет

ПРОБЛЕМА КОДИРОВАНИЯ

Какие коды являются эффективными?

Шеннон в своей теории кодирования развил две главные идеи:

1. Использовать короткие коды для событий, являющихся весьма вероятными.
2. Кодировать нескольких событий одновременно, рассматривая группу этих событий как пакет или метасобытие.

Результатом теории стала **первая теорема Шеннона**, которая показывает связь между средней длиной кода и энтропией.

Уральский федеральный университет

ЧТО ТАКОЕ КОД?

События источника информации – это символы, подлежащие передаче, допустим s_1, s_2, …,s_m.

Символы источника:

- буквы алфавита a, b,..., z;
- цифры от 0 до 9;
- абстрактные символы.

Уральский федеральный университет

Код состоит из **кодовых слов**, включающих знаки из кодового алфавита.

Кодовый алфавит может быть двоичным алфавитом, состоящим из нулей и единиц. Количество знаков в кодовом алфавите обозначается *r*.

Например,

011 – это возможное двоичное кодовое слово, состоящее из трех знаков.

Слово читается слева направо.

Код – это присвоение кодовых слов символам источника.

Например,

источник имеет символы A, B, C,

а кодовый алфавит состоит из 0 и 1.

Закрепление

$A \rightarrow 0$

$B \rightarrow 01$

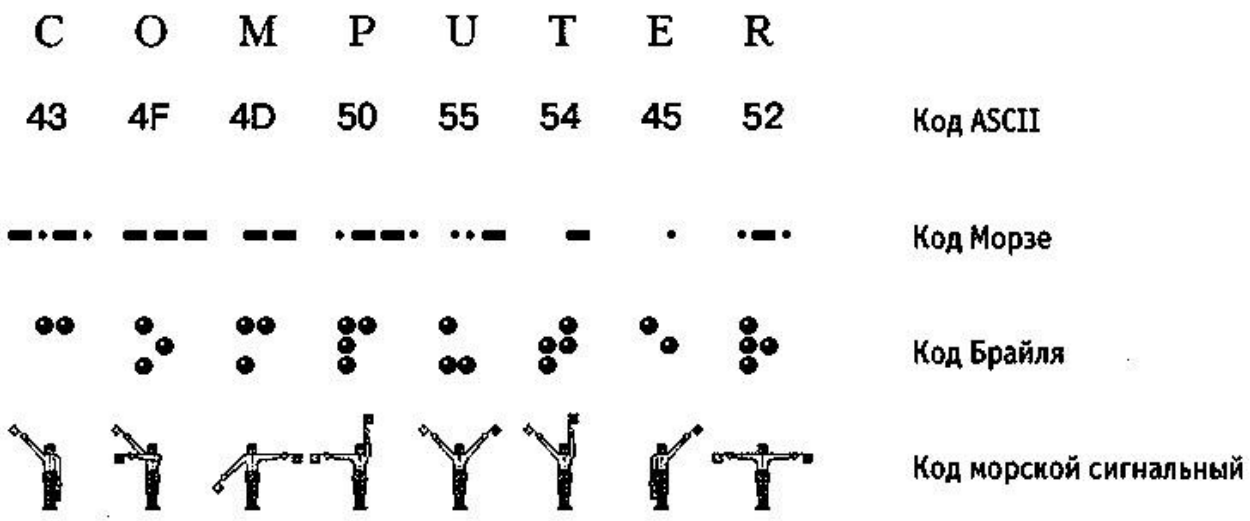
$C \rightarrow 010$

– это код, в котором символы источника переведены в кодовые слова.

Уральский федеральный университет

Примеры кодов:

- Азбука Морзе (точки, тире и пробелы)
- Код ASCII (двоичные разряды)
- Товарный идентификационный код (толстые и тонкие вертикальные линии)



Уральский федеральный университет

ДЛИНА КОДА

- Важной характеристикой кода является **длина его кодовых слов**.
- Чем короче кодовое слово, тем лучше.
- Код, в котором все слова имеют одну и ту же длину, называется **блок-кодом.**
- В некоторых случаях выгоднее пользоваться словами различной длины, тогда с мерой кода связывают **среднюю длину кода**.

Средняя длина кода:

$$L = \sum_{i=1}^{m} p_i l_i$$

где

m – количество символов источника с вероятностями $p_1, p_2, \ldots, p_m$ соответственно,

$l_1, l_2, \ldots, l_m$ – длины соответствующих кодовых слов.

Говорят, что **код является эффективным,** если он имеет наименьшую возможную среднюю длину слова.

Уральский федеральный университет

СВОЙСТВА КОДА

- Коды бывают сингулярными и несингулярными.
- Код является **несингулярным,** если каждое кодовое слово соответствует уникальному символу источника; в противном случае он является **сингулярным.**

Пример. Сингулярный и несингулярный коды. Код является сингулярным, если отсутствует уникальное соответствие между кодовыми словами и символами.

Символ источника	Сингулярный код	Несингулярный код
s_1	00	0
s_2	10	10
s_3	01	00
s_4	10	01

Пример. (Продолжение).

Воспользуемся несингулярным кодом.

Допустим мы получили последовательность:

0010.

Тогда декодируемое сообщение будет:

или $s_1 s_4 s_1$ или $s_3 s_2$ или $s_1 s_1 s_2$.

Т.е. однозначность отсутствует!

Таким образом, **несингулярность кода – это еще не гарантия его эффективности!**

Коды, которые могут декодироваться однозначно, даже когда произвольные номера символов источника кодируются в последовательности, называются **однозначно декодируемыми**.

Именно такими кодами мы должны пользоваться.

Уральский федеральный университет

МОМЕНТАЛЬНЫЕ КОДЫ

- Если каждое слово может быть однозначно декодировано сразу же, как только оно будет получено, то такой код называется **моментальным кодом.**

Например,

Несингулярный блок-код (со всеми кодовыми словами равной длины) – это моментальный код.

Пример. Моментальный блок-код. Как только принимаются два знака, мгновенно можно определить соответствующий символ источника.

Символ источника	Кодовое слово
s_1	00
s_2	01
s_3	10
s_4	11

Тогда, например:

Полученная последовательность: 01101100.

Декодированное сообщение: $s_2 s_3 s_4 s_1$.

Проблема эффективности кодов

- Блок-коды просты для декодирования, но они не всегда бывают эффективными, т.к. мы хотим присвоить короткие кодовые слова высоковероятным символам источника.
- Например, в азбуке Морзе за самой распространенной буквой *E* закрепляется одна точка, в то время как менее распространенной букве Q придается относительно длинное кодовое слово «тире тире точка тире».

Пример. Код запятой и заглавный код. Оба кода являются несингулярными и однозначно декодируемыми, однако заглавный код моментальным не является.

Символ источника	Код запятой	Заглавный код
s_1	0	0
s_2	10	01
s_3	110	011
s_4	1110	0111

Пример. Декодируйте последовательность:

Код запятой: последовательность 01011100 → сообщение $s_1 s_2 s_4 s_1$

Заглавный код: последовательность 00101110 → сообщение $s_1 s_2 s_4 s_1$

Пояснения к примеру:

Оба кода могут однозначно декодироваться.

Но код запятой является моментальным, а заглавный код таким не является.

Нуль в коде запятой указывает на конец слова, и отсюда мы можем декодировать это слово сразу же (моментально).

Нуль в заглавном коде указывает на начало нового слова, и отсюда мы можем вернуться на один знак и декодировать предыдущее слово. Но в этом случае декодирование отстает от приема кодового слова на один знак. Поэтому этот код не является моментальным.

Итак, **моментальный код** – это код, в котором ни одно кодовое слово не является приставкой следующего кодового слова.

Уральский федеральный университет

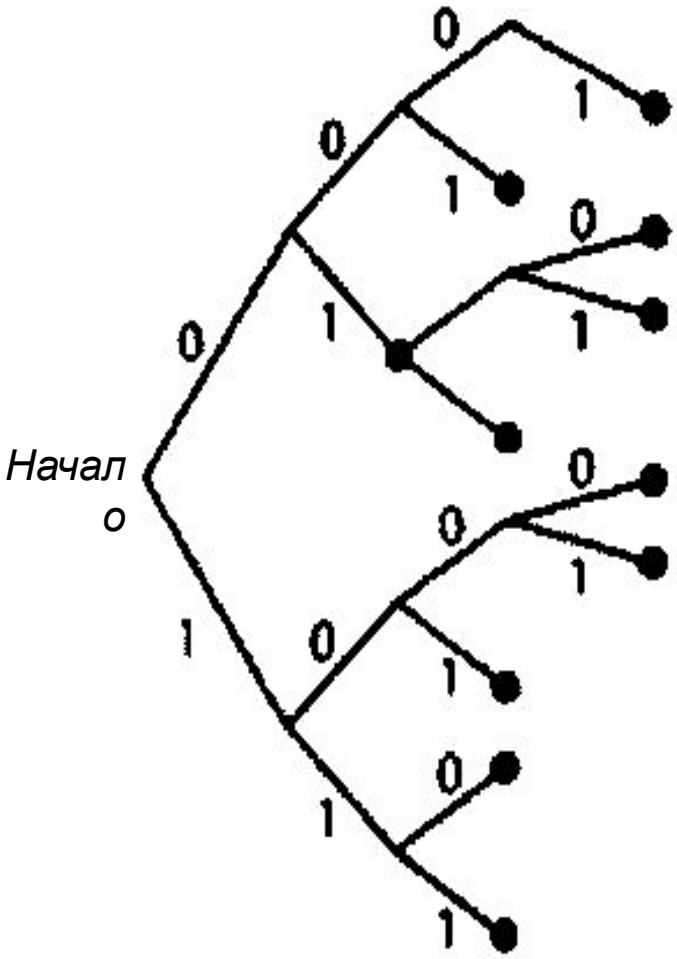
КОДОВЫЕ ДЕРЕВЬЯ

Рассмотрим **деревья для двоичного кода**, состоящего из двух знаков 0 и 1.

- **Кодовое дерево** имеет **точку начала отсчета**, которая представляет собой один **узел**.
- От этого узла идет одна или две ветви, каждая из которых заканчивается другим узлом.
- **Две ветви**, исходящие из узла, маркируются **0** и **1**, что означает направление вверх или вниз соответственно.
- Каждая жирная точка в конкретном дереве представляет собой **кодовое слово**, заданное последовательностью нулей и единиц на пути от первоначального узла к этому узлу.

Уральский федеральный университет

ДВОИЧНОЕ КОДОВОЕ ДЕРЕВО



0001
001
01
0100
0101
011
1000
1001
101
110
111

Пример. Кодовое дерево.

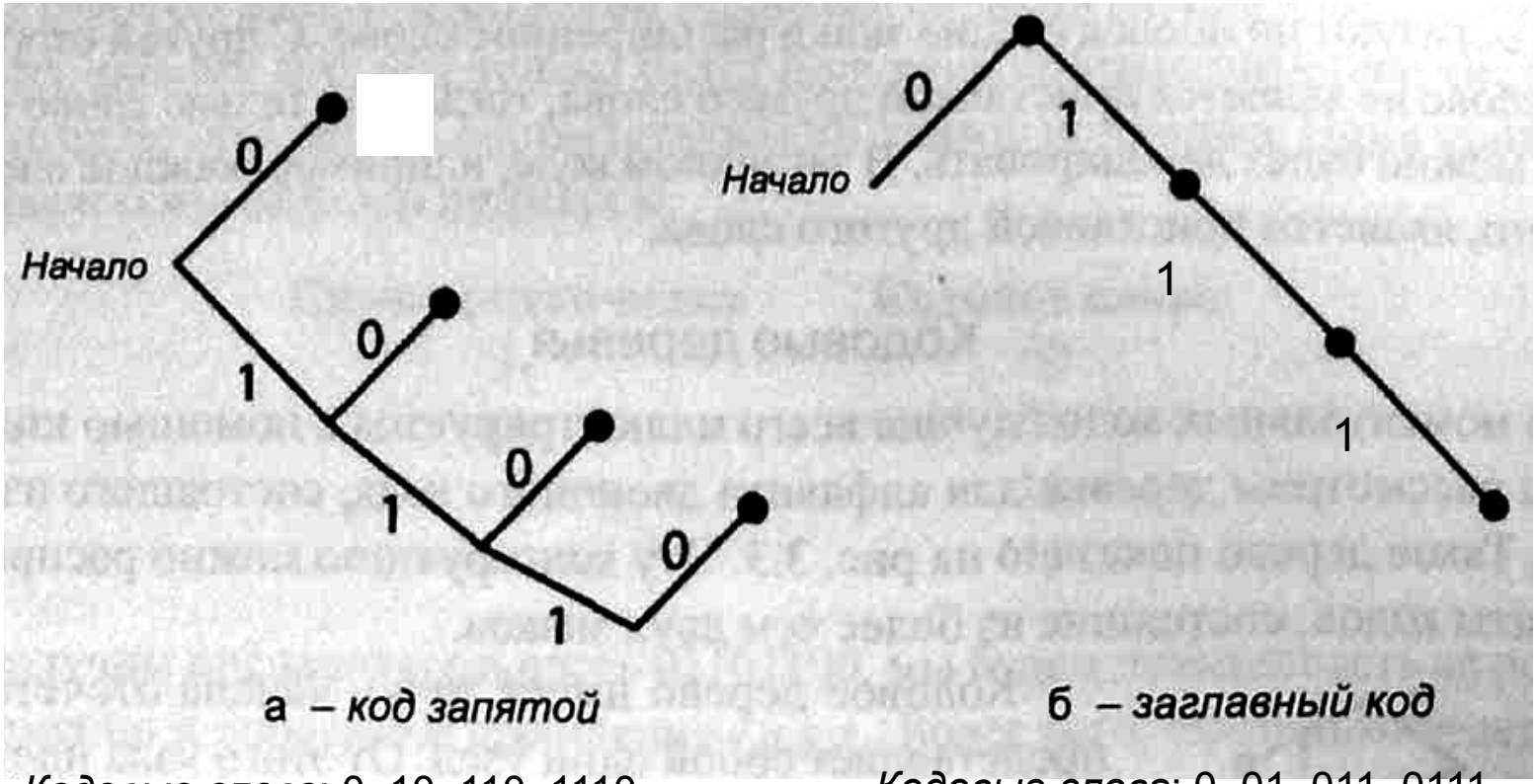
Верхняя точка соответствует слову 0001.

Дерево содержит 11 кодовых слов.

Является ли код, представленный на данном дереве, моментальным?

Свойство моментального кода: все кодовые слова моментального кода соответствуют конечным узлам кодового дерева.

Пример. Два кодовых дерева: код запятой и заглавный код.



Кодовые слова: 0, 10, 110, 1110

Кодовые слова: 0, 01, 011, 0111

Уральский федеральный университет

НЕРАВЕНСТВО КРАФТА

Неравенство Крафта. Моментальный код может быть построен с помощью данных длин кодовых слов $l_1, l_2, \ldots, l_m$ если и только если

$$\sum_{i=1}^{m} r^{-l_i} \leq 1,$$

где r – это число знаков кодового алфавита, а m – это число символов источника.

Для двоичного кода, когда $r = 2$, неравенство Крафта имеет следующий вид:

$$\sum_{i=1}^{m} 2^{-l_i} \leq 1.$$

Доказательство неравенства Крафта основывается на свойстве моментального кода, заключающееся в том, что каждое кодовое слово является конечным узлом дерева.

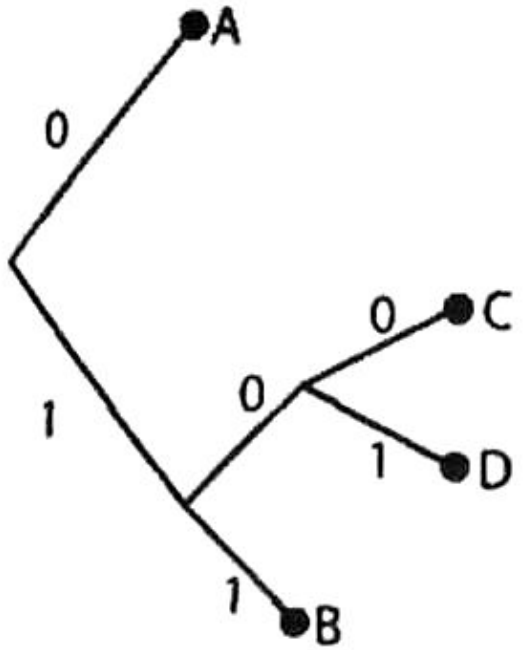


Иллюстрация неравенства Крафта. Моментальный двоичный код для четырех символов A, B, C, D.
Кодом для A используется половина имеющихся кодовых слов. Кодом B – одна четвертая, кодами C и D – одна восьмая.

$$1/2 + 1/4 + 1/8 + 1/8 = 1$$

Уральский федеральный университет

ПРИМЕР 1

Блок-коды. Предположим, что мы хотим построить моментальный код для m символов при помощи двоичного кода $(0,1)$, в котором все кодовые слова имеют равную длину l.

Неравенство Крафта требует, чтобы

$$\sum_{i=1}^{m} 2^{-l} \equiv m2^{-l} \leq 1$$

Отсюда $m \leq 2^l$.

Уральский федеральный университет

ПРИМЕР 2

Два кода. Рассмотрим два двоичных кода:

Символ источника	Код 1	Код 2
s_1	0	0
s_2	10	10
s_3	110	110
s_4	111	11

Длина слов **для первого кода**: 1, 2, 3, 3.

Неравенство Крафта: $\sum_{i=1}^{m} 2^{-l} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} \leq 1$

Следовательно, код 1 – моментальный.

(Продолжение.)

Второй код – это действительный несингулярный код.

Длина слов: 1, 2, 3, 2.

Неравенство Крафта : $$\sum_{i=1}^{m} 2^{-l} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{4} = 1\frac{1}{8}.$$

Таким образом, код 2 не является моментальным.

(Фактически, он не является однозначно декодируемым, поскольку 110 = s_3 = $s_4 s_1$.)

Рассмотрим Лемму.

Лемма 1. Пусть $p_i, i = 1, 2, \dots, m$ и $q_i, i = 1, 2, \dots, m$ удовлетворяют $\sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1.$

При этом все $p_i > 0, q_i > 0.$

Тогда $$\sum_{i=1}^{m} p_i \log p_i \geq \sum_{i=1}^{m} p_i \log q_i$$

причем равенство имеет место в случае, если и только если $p_i = q_i$ для каждого i.

Уральский федеральный университет

СРЕДНЯЯ ДЛИНА КОДА И ЭНТРОПИЯ

Первое соотношение между энтропией и средней длиной кода.

Неравенство длины кода. Средняя длина L двоичного моментального кода удовлетворяет следующему неравенству:

$$L \geq H,$$

где H– это энтропия источника.

Доказательство:

Пусть $l_1, l_2, \ldots, l_m$ – это значения длины слова моментального кода. Рассмотрим числа

$$q_i = \frac{2^{-l_i}}{\sum_{j=1}^{m} 2^{-l_j}}$$

Эти q_i являются положительными и суммируются до 1.

(Продолжение на след. слайде.)

Применяя неравенства леммы, получаем:

$$H = -\sum_{i=1}^{m} p_i \log p_i \le -\sum_{i=1}^{m} p_i \log q_i$$

$$= -\sum_{i=1}^{m} p_i \left[\log 2^{-l_i} - \log \sum_{j=1}^{m} 2^{-l_j} \right]$$

$$= \sum_{i=1}^{m} p_i \left[l_i + \log \sum_{j=1}^{m} 2^{-l_j} \right] \le \sum_{i=1}^{m} p_i l_i = L$$

На последнем этапе используется неравенство Крафта $\sum_{i=1}^{m} 2^{-l_i} \le 1$, что означает $\log \sum_{j=1}^{n} 2^{-l_j} \le 0$.

Отсюда $H \le L$.

Уральский федеральный университет

ПРИМЕР 3

Вероятности в половинной степени.

Предположим, что вероятности символа источника все имеют вид $p_i = \left(\frac{1}{2}\right)^{k_i}$ для различных целых k_i.

Конечно, вероятности должны суммироваться до 1. Например, вероятности могут быть 1/2, 1/4, 1,4 или 1/4, 1/4, 1/8, 1/8, 1/8, 1/16, 1/32, 1/32.

В таких случаях можно задать $l_i = k_i$. Эти значения длин являются действительными вариантами, потому что сумма с левой стороны неравенства Крафта становится:

$$\sum_{i=1}^{m} 2^{-l_i} = \sum_{i=1}^{m} 2^{-k_i} = 1$$

Энтропия источника равняется:

$$H = -\sum_{i=1}^{m} p_i \log p_i = \sum_{i=1}^{m} 2^{-k_i} \log 2^{k_i} = \sum_{i=1}^{m} 2^{-k_i} k_i$$

С другой стороны средняя длина слова равна:

$$L = \sum_{i=1}^{m} p_i l_i = \sum_{i=1}^{n} 2^{-k_i} l_i$$

Поскольку $l_i = k_i$, имеем $H = L$.

Отсюда **для источников с вероятностями, которые представляют собой степени одной второй, моментальные коды существуют при $L = H$.**

ПРИМЕР 4

Символ источника	Вероятность	Кодовое слово
s_1	1/2	0
s_2	1/4	10
s_3	1/8	110
s_4	1/8	111

$$H = -\sum_{i=1}^{m} p_i \log p_i = \frac{1}{2} + \frac{1}{4}2 + \frac{1}{8}3 + \frac{1}{8}3 = \frac{7}{4}$$

$$L = \sum_{i=1}^{m} p_i l_i = \frac{1}{2} + \frac{1}{4}2 + \frac{1}{8}3 + \frac{1}{8}3 = \frac{7}{4}$$

Уральский федеральный университет

КОДИРОВАНИЕ ШЕННОНА

Если вероятности источника имеют вид $p_i = \left(\frac{1}{2}\right)^{k_i}$,

то $l_i = k_i$. Или по-другому:

$$l_i' = \log \frac{1}{p_i}$$

Если вероятности источника не являются половинными степенями, то числа l_i' могут не быть целыми, но они все равно удовлетворяют неравенству Крафта. Вычисляя сумму, получаем:

$$\sum_{i=1}^{m} 2^{-l_i'} = \sum_{i=1}^{m} 2^{\log p_i} = \sum_{i=1}^{m} p_i = 1$$

Из идеи Шеннона следует, что каждое из этих l_i' увеличивается до следующего самого большого целого числа (т.е. округляется), и новые величины обозначаются как l_i.

Отсюда, имеет место моментальный код и длина слова

$$l_i < \log(1/p_i) + 1 \quad \text{для каждого } i.$$

Отсюда, средняя длина слова удовлетворяет неравенству:

$$L < \sum_{i=1}^{n} p_i \left(\log \frac{1}{p_i} + 1 \right) = H + 1$$

Уральский федеральный университет

ПРЕДЕЛЫ СРЕДНЕЙ ДЛИНЫ

Итак, мы нашли **верхний и нижний предел** средней длины кода.

Пределы средней длины. Источник с энтропией H может кодироваться с помощью моментального двоичного кода средней длины L, удовлетворяющего неравенству

$$H \leq L < H + 1.$$

Уральский федеральный университет

ПЕРВАЯ ТЕОРЕМА ШЕННОНА

Предположим, что вместо S кодируется S^n («большой пакет»).

Можно найти моментальный код для S^n, имеющий среднюю длину кодового слова L, удовлетворяющий

$$H(S^n) \leq L \leq H(S^n) + 1.$$

Это эквивалентно следующему:

$$nH(S) \leq L \leq nH(S) + 1.$$

В результате деления на n получаем:

$$H(S) \leq L/n \leq H(S) + 1/n.$$

Первая теорема Шеннона. Кодируя последовательность независимых символов (в S^n), можно построить декодируемые коды таким образом, что

$$\lim_{n\to\infty}\frac{L_n}{n} = H$$

где H – энтропия источника S,

n – длина последовательностей символов,

L_n – средняя длина кодовых слов, соответствующая S^n.

Отсюда можно приблизить среднюю длину кодового слова L/n настолько, насколько нужно, к H.

Цена, которую нужно заплатить за такое улучшение, заключается в увеличении сложности кодирования в силу увеличения размера источника и увеличения задержки в процессе декодирования, поскольку вся последовательность должна обрабатываться одновременно.

Уральский федеральный университет

ПРИМЕР 5

Два дня погоды в Калифорнии

Предположим, что погода в Калифорнии солнечная (S) с вероятностью 7/8 и облачная (C) с вероятностью 1/8. Предположим также, что погодные условия в последующие дни являются независимыми.

Для того чтобы отправить прогноз погоды **на один день** с помощью двоичного кода, потребуется код с длиной $L = 1$ (один символ на каждое состояние погоды).

Уральский федеральный университет

ПРИМЕР 5 (ПРОДОЛЖЕНИЕ)

Если будут отправляться прогнозы **на два дня вместе**, может быть использован модифицированный код запятой, как показано ниже.

Символ источника	Вероятность	Код
SS	49/64	0
SC	7/64	10
CS	7/64	110
CC	1/64	111

$$L = \frac{49}{64} + \frac{7}{64}2 + \frac{7}{64}3 + \frac{1}{64}3 = \frac{87}{64}$$

Средняя длина на один ежедневный прогноз поэтому составит:

$$L/2 = \frac{87}{128} = 0{,}68$$

Таким образом, дополнительная гибкость значительно сократила среднюю длину кода и приблизила ее к энтропии источника (в данном примере $H = 0{,}54$ из предыдущей лекции).

Путем рассмотрения более длинных последовательностей ($n > 2$) можно достичь ещё большего сокращения средней длины.

Уральский федеральный университет

ВЫВОДЫ

1. События источника информации (это могут быть, например, буквы в тексте) передаются посредством кодовых слов с использованием нулей и единиц. (Моментальные двоичные коды).
2. По первой теореме Шеннона средняя длина кодового слова кода в символах на одно событие должна быть больше или равна энтропии источника.
3. Уменьшение средней длины кодового слова требует, чтобы кодирование применялось к длинным последовательностям, а не к единичным событиям.
4. Кодирование и декодирование в этом случае может стать сложным, и между передачей и декодированием кодовых слов может происходить продолжительная задержка.

Уральский федеральный университет

УПРАЖНЕНИЯ

1. (Значения длины кода.) Какие из следующих значений длины кода являются реальными для построения моментального двоичного кода для пяти символов?

(а) 2 2 2 3 3

(б) 1 2 2 4 5

(в) 1 2 3 4 4

(г) 1 2 3 3 8

Уральский федеральный университет

2. (Декодируемость.) Является ли следующий код однозначно декодируемым?

A 0 1

B 1 0

C 0 1 1

D 1 0 1

3. (Шрифт Брайля.) В шрифте Брайля каждый знак состоит из рисунка точек, воз вышающихся над поверхностью. В знаке имеется шесть позиций, каждая из которых может быть либо плоской, либо выпуклой. В результате получается общее количество 2^6 или 64 возможные буквы, которые могут быть описаны одним знаком Брайля.

Для стандартного английского языка имеется более 64 символов, подлежащих описанию; от *a* до *z* как в верхнем, так и в нижнем регистре, от 0 до 9 и стандартная пунктуация (пробел, запятая и т.д.). Таким образом, некоторые знаки требуют более одного знака Брайля.

Предположим, что 8 процентов букв – это заглавные буквы и 12 процентов букв – это разряды (цифры).

(а) В шрифте Брайля 1-й ступени строчные буквы (буквы нижнего регистра) и стандартная пунктуация требуют одного знака. Имеется также специальный знак, указывающий, что следующий знак будет давать описание заглавной буквы. Аналогичным образом имеется специальный знак, указывающий, что следующий знак будет описывать цифру (разряд).

Какое ориентировочное количество знаков Брайля используется для описания стандартной английской буквы?

(б) В шрифте Брайля 2-й ступени, кроме знаков, обозначающих заглавные буквы и цифры, имеются одинарные знаки, использующиеся для обозначения общих групп букв (например, *the, and, ing).* Предположим, что каждая группа букв имеет длину в три знака и что 20 процентов стандартного английского текста происходит из одной из общих групп букв, используемых в шрифте Брайля 2-й ступени.

Какое ориентировочное количество знаков Брайля используется для описания стандартной английской буквы?

Уральский федеральный университет

4. (Двойной код.) Предположим, что в качестве простой меры безопасности от подслушивания каждому из символов источника S присваивается два кодовых слова, и в процессе передачи передаваемое слово выбирается произвольно, каждое с вероятностью 50 процентов. Конечно, код все равно должен быть однозначно декодируемым. С точки зрения энтропии $H(S)$ какой нижний предел средней длины кодового слова кода?

5. (Знаки из трех слов.) Рассмотрим указанный ниже источник.

s_1	1/3	s_5	1/27
s_2	1/3	s_6	1/27
s_3	1/9	s_7	1/27
s_4	1/9		

(а) При помощи логарифмов с основанием 3 рассчитайте энтропию данного источника.

(б) Сделав допущение, что в алфавите кода имеется три знака 0,1,2, найдите код для источника, имеющего среднюю длину, равную его энтропии с основанием 3.

Уральский федеральный университет

Спасибо за внимание!