3 ПОБУКВЕННОЕ КОДИРОВАНИЕ

3.1 Определения

- **Одной из основных задач теории информации является построение кодов для сообщений, порождаемых некоторым источником информации.**
- ***Кодирование* – это способ представления информации в удобном для хранения и передачи виде.**

- ***Кодирование дискретного источника*** **с алфавитом *A* заключается в сопоставлении сообщениям источника символам или группам символов в алфавите *B* (которые называются *кодовыми словами*).**

- **Алфавит *B* называется *кодовым алфавитом*.**
- ***Кодом* называется совокупность всех кодовых слов, применяемых для представления порождаемых источником символов.**

- ***Побуквенное кодирование*** **задается *таблицей кодовых слов:***

$$\sigma = \langle a_1 \to \beta_1, \ldots, a_n \to \beta_n \rangle$$

$a_i \in A$

β_i **– конечные последовательности в алфавите *B*.**

- **Количество букв в слове (последовательности символов) называется *длиной слова*.**
- **Пустое слово, т.е. слово, не содержащее ни одного символа обозначается Λ.**
- **Если $a=a_1a_2$, то a_1– *начало (префикс)* слова, a_2 – *окончание (постфикс)* слова.**

- **Далее будет рассматриваться *двоичное кодирование*, т.е. размер кодового алфавита равен 2.**
- **Конечную последовательность битов (0 или 1) назовем *кодовым словом*, а количество битов в этой последовательности – *длиной кодового слова*.**

- **Обратная процедура сопоставления кодовым словам в алфавите *B* символов алфавита *A* называется *декодированием*.**

Пример 1

- **Код ASCII (американский стандартный код для обмена информацией) каждому символу ставит в однозначное соответствие кодовое слово длиной 8 бит.**

- **Различают** ***два класса методов кодирования*** **дискретного источника информации: равномерное и неравномерное кодирование.**
- **Под** ***равномерным кодированием*** **понимается использование кодов со словами постоянной длины.**

- **Для того чтобы декодирование равномерного кода было возможным, разным символам алфавита источника должны соответствовать разные кодовые слова.**
- **При этом длина кодового слова должна быть не меньше** $\lceil \log_n m \rceil$ **символов,**

  где *m* – размер исходного алфавита,

  ***n* – размер кодового алфавита.**

- **При *неравномерном кодировании источника* используются кодовые слова разной длины.**
- **Причем кодовые слова обычно строятся так, чтобы часто встречающиеся символы кодировались более короткими кодовыми словами, а редкие символы – более длинными (за счет этого и достигается «сжатие» данных).**

Пример 2

- **Азбука Морзе является общеизвестным кодом из символов телеграфного алфавита, в котором буквам русского языка соответствуют кодовые слова (последовательности) из «точек» и «тире».**

- **При передаче закодированного сообщения по каналу связи могут возникать помехи (или шум), которые искажают сообщение, так что при декодировании приемник может получить изменённое сообщение.**
- **Для защиты сообщения от помех при передаче по каналу связи существуют специальные методы помехоустойчивого кодирования.**

- **Под *сжатием данных* понимается *компактное представление данных*, достигаемое за счет уменьшения избыточности информации, содержащейся в сообщениях.**
- **Большое значение для практического использования имеет *неискажающее сжатие*, позволяющее полностью восстановить исходное сообщение.**

- **При *неискажающем сжатии* происходит кодирование сообщения перед началом передачи или хранения, а после окончания процесса сообщение однозначно декодируется (это соответствует модели канала без помех).**

***Методы сжатия данных* можно разделить на две группы:**

- **статические методы**
- **адаптивные методы**

- ***Статические*** **методы сжатия данных предназначены для кодирования конкретных источников информации с известной статистической структурой, порождающих определенное множество сообщений.**

- **К наиболее известным статическим методам сжатия относятся коды Хаффмана, Шеннона, Фано, Гилберта-Мура, арифметический код и другие методы, которые используют известные сведения о вероятностях порождения источником различных символов или их сочетаний**

- **Если статистика источника информации неизвестна или изменяется с течением времени, то для кодирования сообщений такого источника применяются *адаптивные методы сжатия*.**

- Существует множество различных адаптивных методов сжатия данных. Наиболее известные из них – адаптивный код Хаффмана, код «стопка книг», интервальный и частотный коды, а также методы из класса Лемпела-Зива.

3.2 Префиксные и разделимые коды

- **Очевидным требованием к кодированию сообщений источника является условие однозначного декодирования,**

  т.е. после получения закодированного сообщения получатель должен иметь возможность прочитать исходное сообщение.

- **Рассматривается задача построения однозначно декодируемых кодов без учета статистики источника информации, т.е. можно считать, что сообщения источника независимы и равновероятны.**

- ***Кодирование сообщения*, которое порождает источник информации, будем понимать как сопоставление кодовой последовательности всему сообщению в целом или как построение кода сообщения из кодов его частей (побуквенное кодирование).**

- **Побуквенный код называется *разделимым* (или *однозначно декодируемым*), если любое сообщение из символов алфавита источника, закодированное этим кодом, может быть однозначно декодировано**
- **При разделимом кодировании любое кодовое слово единственным образом разлагается на элементарные коды.**

Пример

- **Пусть** $A = \{a_1, a_2, a_3\}$ $B = \{0,1\}$
- **Код**

$a_1 \to 1001$ $a_2 \to 0$ $a_3 \to 010$

- **не является разделимым, поскольку кодовое слово 010010 может быть декодируемо двумя способами:**

a_3a_3 или $a_2a_1a_2$

- **Побуквенный код называется *префиксным*, если в его множестве кодовых слов ни одно слово не является началом другого, т.е. элементарный код одной буквы не является префиксом элементарного кода другой буквы.**

Пример

- **Код из предыдущего примера не является префиксным,**

  поскольку элементарный код буквы a_2 является префиксом элементарного кода буквы a_3 .

- **Утверждение.**

  Префиксный код является разделимым.

Пример

- **Пусть *A*={*a,b*}, *B*={0,1}**
- **Разделимый код может быть не префиксным**

$$\sigma = \langle a \to 0, b \to 01 \rangle$$

3.3 Неравенство Крафта

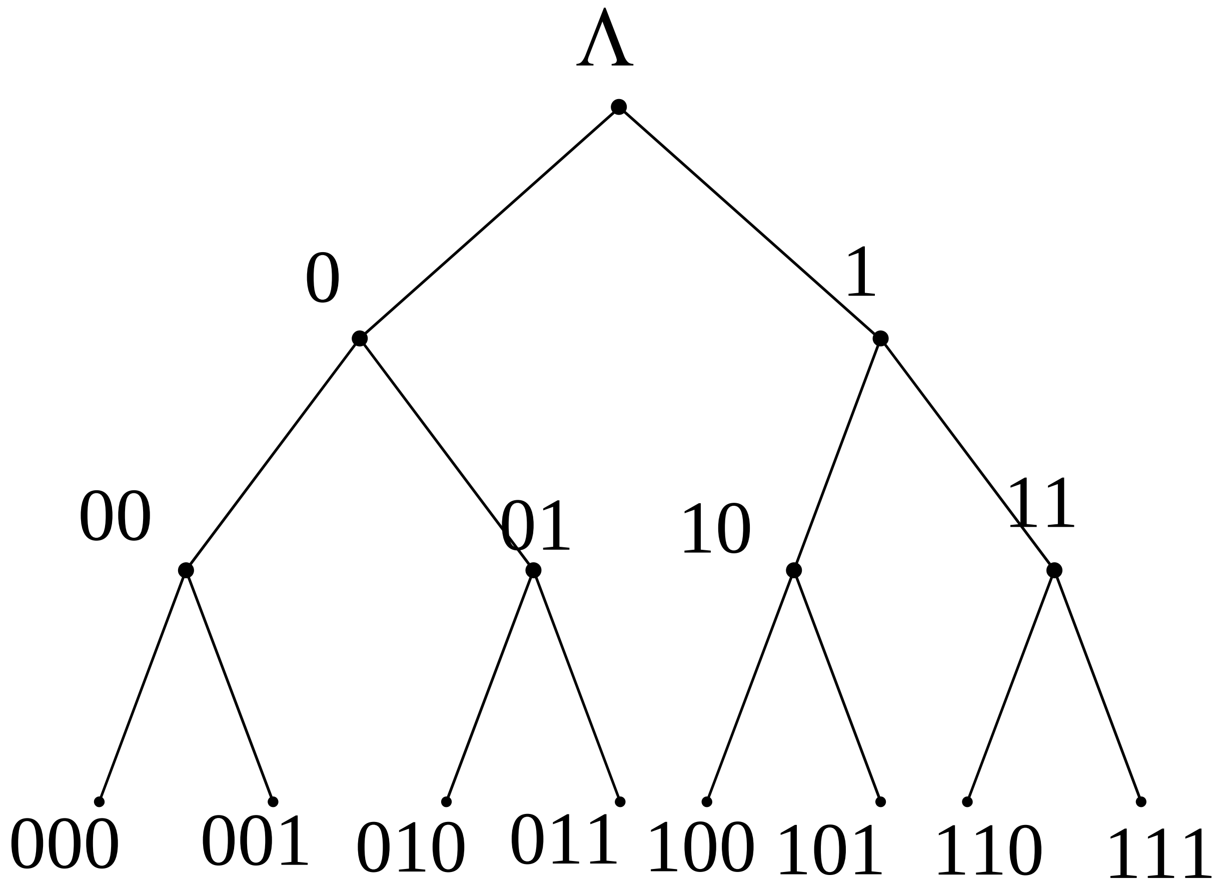
Теорема *(Крафт)*

Для того, чтобы существовал побуквенный двоичный префиксный код с длинами кодовых слов $L_1,...,L_n$ *необходимо и достаточно, чтобы*

$$\sum_{i=1}^{n} 2^{-L_i} \le 1$$

Доказательство

- Докажем необходимость.
- Пусть существует префиксный код с длинами $L_1, \ldots, L_n$
- Рассмотрим полное двоичное дерево высоты *h*.
- Каждая вершина закодирована последовательностью нулей и единиц (как показано на рисунке).


Λ
0
1
00
01
10
11
000
001
010
011
100
101
110
111

- В этом дереве выделим вершины, соответствующие кодовым словам.
- Тогда любые два поддерева, соответствующие кодовым вершинам дерева, не пересекаются, т.к. код префиксный.

- В полном дереве высоты h всего 2^h листьев.

В поддереве, соответствующем кодовому слову длины L_i всего листьев 2^{h-L_i}

Тогда $$\sum_{i=1}^{n} 2^{h-L_i} \le 2^h \qquad \sum_{i=1}^{n} 2^{-L_i} \le 1$$

- Докажем достаточность утверждения
- Пусть существует набор длин кодовых слов такой, что $\sum_{i=1}^{n} 2^{-L_i} \leq 1$

Рассмотрим полное двоичное дерево с помеченными вершинами.

Пусть длины кодовых слов упорядочены по возрастанию $L_1 \leq L_2 \leq ... \leq L_n$

- **Выберем в двоичном дереве вершину V_1 на уровне L_1.**
- **Уберем поддерево с корнем в вершине V_1 .**
- **В оставшемся дереве возьмем V_2 вершину на уровне L_2 и удалим поддерево с корнем в этой вершине и т.д.**

- Последовательности, соответствующие вершинам $V_1, V_2, \ldots, V_n$ образуют префиксный код.

Теорема доказана

Пример

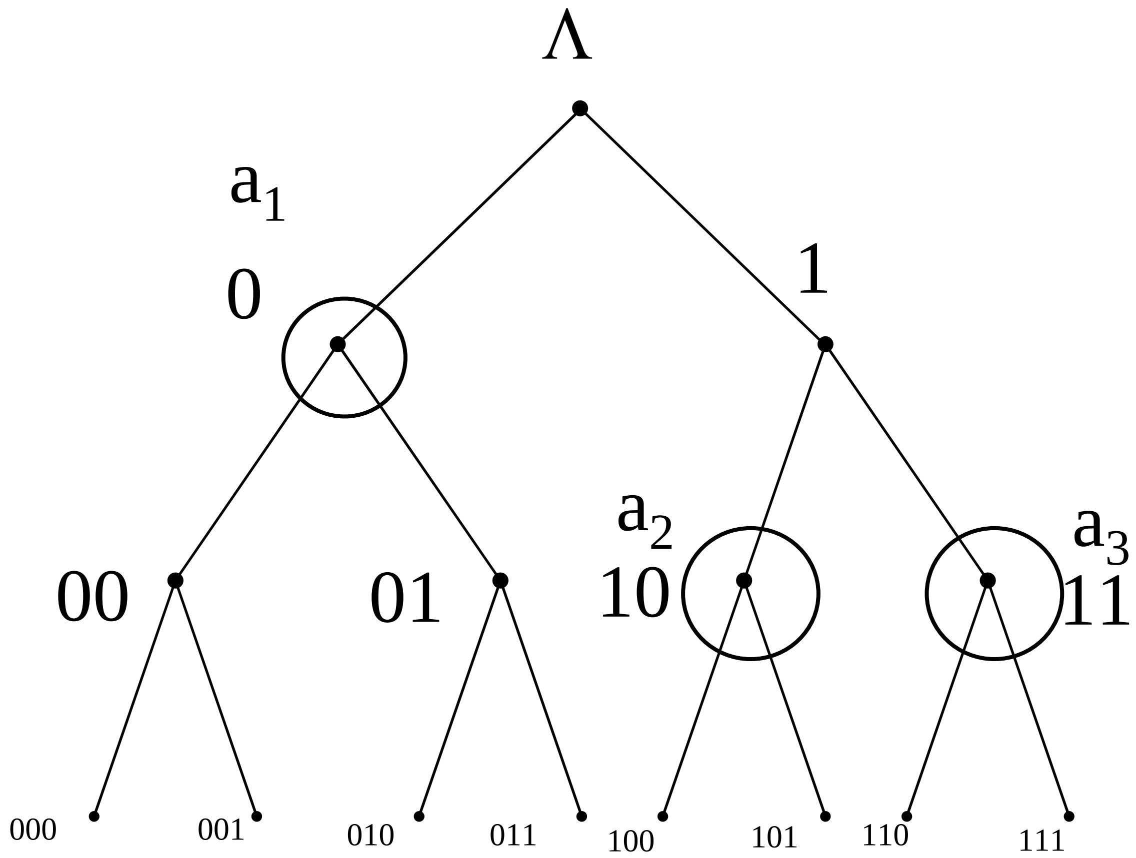
- Построить двоичный префиксный код с длинами $L_1=1$, $L_2=2$, $L_3=2$ для алфавита
$$A=\{a_1,a_2,a_3\}$$

Проверим неравенство Крафта для набора длин
$$\frac{1}{2^1}+\frac{1}{2^2}+\frac{1}{2^2}=1$$

Неравенство выполняется и, следовательно, префиксный код с таким набором длин кодовых слов существует.

- Рассмотрим полное двоичное дерево с 2^3 помеченными вершинами и выберем вершины дерева, как описано выше в доказательстве теоремы Крафта.
- Тогда элементарные коды могут быть такими:

$$a_1 \to 0 \qquad a_2 \to 10 \qquad a_3 \to 11$$



Λ
a_1
0
1
00
01
a_2
10
a_3
11
000
001
010
011
100
101
110
111

- Построить префиксный код с длинами $L_1=1,\ L_2=1,\ L_3=2$

- Процесс декодирования может быть организован следующим образом.
- Просматриваем полученное сообщение, двигаясь по дереву.
- Если попадем в кодовую вершину, то выдаем соответствующую букву и возвращаемся в корень дерева и т.д.

- Теорема Крафта, доказанная в предыдущем параграфе, может быть обобщена на случай разделимых кодов

Теорема *(МакМиллан)*

Для того, чтобы существовал побуквенный двоичный разделимый код с длинами кодовых слов $L_1, \ldots, L_n$ *необходимо и достаточно, чтобы*

$$\sum_{i=1}^{n} 2^{-L_i} \le 1$$