СПбГЭТУ «ЛЭТИ»
ПЕРВЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ

А.Б. Левина

Теория информации и теория кодирования

Сжатие данных

СПбГЭТУ «ЛЭТИ», 2022 г.

1 ВВЕДЕНИЕ

На данной лекции мы познакомимся с основными принципами сжатия информации, рассмотрим принципы неравномерного кодирования, код Шеннона, код Гильберта Мура, кодирование стационарного источника, арифметическое кодирование, кодирование дискретных источников, дискретный стационарный процесс, префиксный код, неравенство Крафта, код Хафманна.

2 КОДИРОВАНИЕ ДИСКРЕТНЫХ ИСТОЧНИКОВ

В этой части мы узнаем, что такое:

a. Дискретный стационарный процесс
b. Неравномерное побуквенное кодирование
c. Префиксный код
d. Неравенство Крафта
e. Код Хафманна

2.1 Дискретный стационарный процесс

Введем определение стационарности.

Определение: **Стационарность** — свойство процесса не менять свои характеристики со временем.

Стационарный процесс - это стохастический/случайный процесс, у которого не изменяется распределение вероятности при смещении во времени.

В теории вероятностей случайный процесс называется стационарным, если все его вероятностные характеристики не меняются с течением времени *t*.

Определение: Случайный процесс X(t) называется процессом **дискретным во времени**, если система, в которой он протекает, меняет свои состояния только в моменты времени $t_1,\ldots t_n$, число которых конечно или счётно.

Определение: Случайный процесс называется **процессом с непрерывным временем**, если переход из состояния в состояние может происходить в любой момент времени.

Определение: Случайный процесс называется **стационарным,** если все многомерные законы распределения зависят только от взаимного расположения моментов времени $t_1,...t_n$, но не от самих значений этих величин.

Так же случайный процесс называется **стационарным,** если его вероятностные закономерности неизменны во времени.

В противном случае, он называется ***нестационарным*.**

2.2 Энтропия на сообщение дискретного стационарного источника

После того как мы познакомились с дискретным стационарным источником введем такое важное понятие как энтропия, но для начала проговорим условие стационарности.

Пусть произвольный дискретный стационарный источник порождает последовательность $x=(x_1,.....x_t,.....)$, $x_t \in X_t=X$.

Условие стационарности:

1. Для любого t $p(x_t)=p(x)$ не зависит от t.
1. $H(X_t)=H(X)$не зависит от t, назовем ее *одномерной энтропией источника* и обозначим $H_1(X)$.

$H_1(X)$ – не учитывает зависимости букв!

Рассмотрим последовательность из n последовательных букв источника $\boldsymbol{x} = (x_1, ..., x_n) \in X_1X_2...X_n = X^n$. Для стационарного процесса энтропия распределения вероятностей на таких блоках $H(X_1...X_n) = H(X^n)$ не зависит от расположения блока во времени, ее называют *n-мерной энтропией* источника.

Величина $H(X^n)$ определяет среднее количество информации в последовательности из n букв.

Энтропия на букву последовательности длины n:

$$H_n(X) = \frac{H(X^n)}{n},$$

Другой подход к измерению информации:

- При передачи буквы x_n все предыдущие (x_1,….x_{n-1}) декодеру известны.
- Среднее количество подлежащей передаче информации об x_n определяется величиной условной энтропии

$$H(X_n|X_1, ..., X_{n-1}).$$

В силу стационарности конкретные значения индексов не играют роли, важна лишь длина предыстории, поэтому:

$$H(X_n|X_1, \ldots, X_{n-1}) = H(X|X^{n-1}).$$

Теорема 1.4 *Для дискретного стационарного источника*
A. $H(X|X^n)$ *не возрастает с увеличением* n*;*
B. $H_n(X)$ *не возрастает с увеличением* n*;*
C. $H_n(X) \geq H(X/X^{n-1})$*;*
D. $\lim_{n\to\infty} H_n(X) = \lim_{n\to\infty} H(X|X^n)$.

Из всего вышеизложенного мы можем сделать вывод, что для дискретного стационарного источника, выполнены следующие свойства:

1. Основной параметр – энтропия;
2. $H(X) \leq \log|X|$ равенство возможно когда все элементы X равновероятны;
3. Получение «теоретически» лучшего сжатия возможно или при работе с большими блоками или при знание «предыстории» всех символов.

1.3 НЕРАВНОМЕРНОЕ ПОБУКВЕННОЕ КОДИРОВАНИЕ

Рассмотрим алгоритм неравномерного побуквенного кодирования.

Постановка задачи:

Необходимо: Для дискретного источника X с {p(x), x€X} необходимо построить неравномерный двоичный код над алфавитом A={a} (A={0,1}).

Неравномерный побуквенный код $C = \{\boldsymbol{c}\}$ объема $|C| = M$ над алфавитом A определяется как произвольное множество последовательностей одинаковой или различной длины из букв алфавита A.

Код является *однозначно декодируемым*, если любая последовательность символов из A единственным способом разбивается на отдельные кодовые слова.

Введем определение префиксного кода.

Определение: Если ни одно кодовое слов не является началом другого код называется *префиксным*.

Префиксность - достаточное, но не необходимое условие однозначного декодирования.

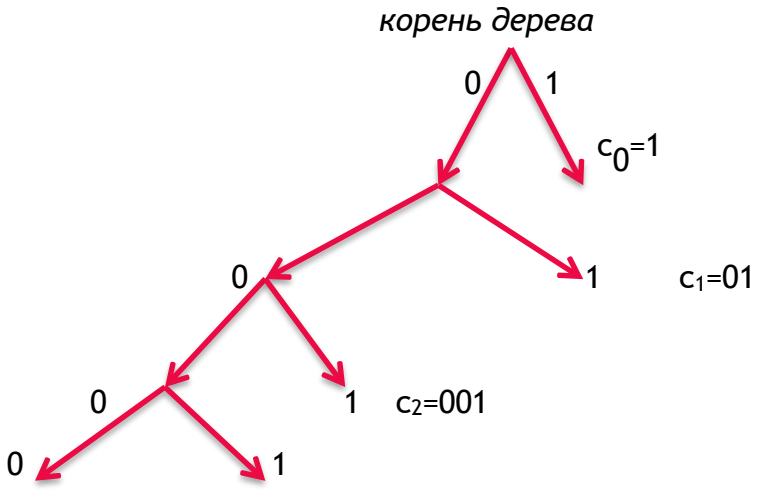
Рассмотрим пример: X= {0, 1, 2, 3} Какой код префиксный?

C_1= {00, 01, 10, 11}

C_2= {1, 01, 001, 000}

C_3= {1, 10, 100, 000}

C_4= {0, 1, 10, 01}



Определение: Код называется *древовидным*, если в качестве кодовых слов он содержит только кодовые слова, соответствующие концевым вершинам кодового дерева.

Древовидность кода = префиксность

Цель неравномерного кодирования=уменьшение затрат на передачу

В качестве критерия качества кода выберем среднюю длину кодового слова

Рассмотрим источник X= {1,….M} порождающий буквы с вероятностью $\{p_1,….p_M\}$, для кодирования взят код C=$\{c_1,…c_M\}$ с длиной слов $l_1….l_M$

Средней длиной кодовых слов называется величина

$$\bar{l} = \mathrm{M}[l_i] = \sum_{i=1}^{M} p_i l_i.$$

Задача побуквенного кодирования

Задача побуквенного неравномерного кодирования формулируется как задача построения однозначно декодируемого кода с наименьшей средней длиной кодовых слов при заданных ограничениях на сложность.

2.4 НЕРАВЕНСТОВ КРАФТА

Введем определение неравенства Крафта.

Теорема 2.1 *Необходимым и достаточным условием существования префиксного кода объема M с длинами кодовых слов $l_1, ..., l_M$ является выполнение неравенства Крафта*

$$\sum_{i=1}^{M} 2^{-l_i} \leq 1. \qquad (2.1)$$

С помощью неравенства Крафта мы можем доказать Теоремы побуквенного кодирования.

Теорема 2.2 *Для ансамбля $X = \{x, p(x)\}$ с энтропией H существует побуквенный неравномерный префиксный код со средней длиной кодовых слов $\bar{l} < H + 1$.*

Теорема 2.3 *Для любого однозначно декодируемого кода дискретного источника* $X = \{x, p(x)\}$ *с энтропией* H *средняя длина кодовых слов* $\bar{l}$ *удовлетворяет неравенству*

$$\bar{l} \geq H. \tag{2.2}$$

Следствие 2.4 *Для существования кода со средней длиной кодовых слов* $\bar{l} = H$ *необходимо и достаточно чтобы все вероятности сообщений* $x \in X$ *имели вид* $p(x) = 2^{-l(x)}$ *, где* $\{l(x)\}$ *– целые положительные числа.*

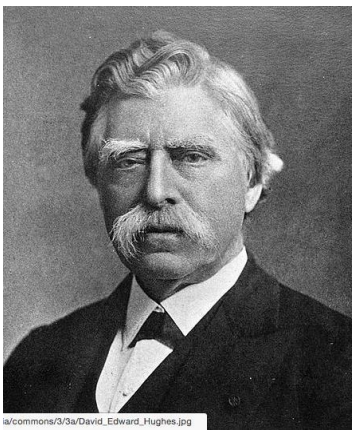
Теперь мы модем рассмотреть код, который является оптимальным побуквенным кодом.

1.4 Оптимальный побуквенный код - код Хаффмена

Алгоритм Хаффмана — жадный алгоритм оптимального префиксного кодирования алфавита с минимальной избыточностью.

Был разработан в 1952 году аспирантом Массачусетского технологического института Дэвидом Хаффманом при написании им курсовой работы. В настоящее время используется во многих программах сжатия данных.

Дэвид Хаффмен (1925-1999)



В 1955 году — Медаль Луиса Леви от Франклинского института за докторскую диссертацию о последовательно переключающихся схемах.

В 1973 году — Премию Уоллеса Макдауэлла.

В 1981 году — награду от компьютерного сообщества IEEE.

В 1998 году — золотую юбилейную награду за технологические новшества от IEEE.

В 1999 году — Медаль Ричарда Хэмминга за исключительный вклад в теорию информации.

Рассмотрим задачу построения оптимального кода - кода Хаффмана.

$X= \{1,….M\}$ с вероятностью сообщений $\{p_1,….p_M\}$, $p_1 \le p_2 \le ….. \le p_M$ (!!!)

Задача: построение оптимального кода, т.е. кода с наименьшей длиной кодовых конструкций.

Пусть код $C=\{c_1,…c_M\}$ с длиной кодовых слов $\{l_1….l_M\}$ оптимален для рассматриваемого ансамбля сообщений.

Свойства:

1. Если $p_i<p_j$ то $l_i \ge l_j$
2. Не менее двух кодовых слов имеют одинаковую длину $l_M=\max_m l_m$
3. Среди кодовых слов длины $l_M=\max_m l_m$ найдутся два слова, отличающиеся только в одном последнем символе.

Для ансамбля $X=\{1,…M\}$ и кода C удовлетворяющего свойствам 1-3 введем ансамбль $X'=\{1,…,M-1\}$ с вероятностью $\{p'_1,…,.p'_{M-1}\}$: $p'_1=p_1,….,p'_{M-2}=p_{M-2}$, $p'_{M-1}=p_{M-1}+p_M$

Из кода C построим C' для ансамбля X', приписав $x'_1,…,x'_{M-2}$ те же кодовые слова, что и в C, а $x'_{M-1}=c'_{M-1}$, где c'_{M-1} общая часть c_{M-1} и c_M.

Свойства:

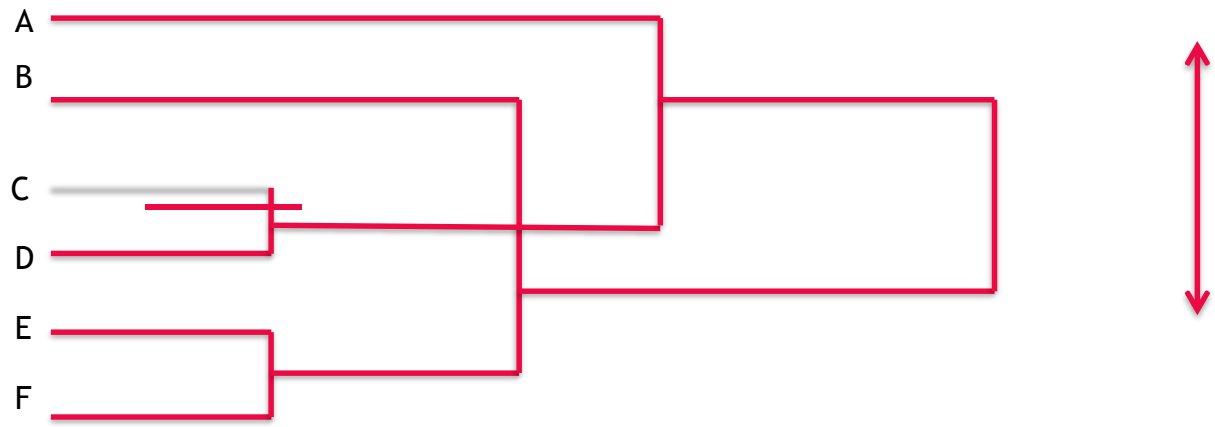
1. Если код C' оптимальный для X', то код C оптимален для X;
2. Задача построения кода объемом M сводится к задаче построения кодов объема $M'=M-1$, т.е. Получаем рекуррентное правило построения кодового дерева оптимального неравномерного кода.

Пример:

X={a,b,c,d,e,f} с вероятностью {0,35; 0,2; 0,15; 0,1; 0,1; 0,1}

Задание:

Посчитать длину кодового слова.



a	11
b	01
c	101
d	100
e	001
f	000

Метод кодирования состоит из двух основных этапов:

- Построение оптимального кодового дерева.
- Построение отображения код-символ на основе построенного дерева.

Записать сам алгоритм можно следующим образом:

1. Символы входного алфавита образуют список свободных узлов. Каждый лист имеет вес, который может быть равен либо вероятности, либо количеству вхождений символа в сжимаемое сообщение.
2. Выбираются два свободных узла дерева с наименьшими весами.
3. Создается их родитель с весом, равным их суммарному весу.
4. Родитель добавляется в список свободных узлов, а два его потомка удаляются из этого списка.

5 Одой дуге, выходящей из родителя, ставится в соответствие бит 1, другой — бит 0. Битовые значения ветвей, исходящих от корня, не зависят от весов потомков.

6 Шаги, начиная со второго, повторяются до тех пор, пока в списке свободных узлов не останется только один свободный узел. Он и будет считаться корнем дерева.

3. СЖАТИЕ ДАННЫХ

Рассмотрим алгоритмы, используемые для произведения сжатия данных, в этой части будет рассмотрено:

1. Код Шеннона
2. Код Гильберта Мура
3. Кодирование для стационарного источника
4. Арифметическое кодирование

3.1 КОД ШЕННОНА

В области сжатия данных **код Шеннона** — алгоритм сжатия данных без потерь с помощью построения префиксных кодов на основе набора символов и их вероятностей (расчётное или измеренное). Он является субоптимальным в том смысле, что не позволяет достичь минимально возможных кодовых длин как в кодировании Хаффмана.

Методика была использована для доказательства теоремы Шеннона о помехоустойчивом кодировании в 1948 в его статье «Математическая Теория связи».

Пусть X= {1,….M} с вероятностью сообщений $\{p_1,\ldots.p_M\}$, $p_1 \le p_2 \le \ldots.. \le p_M$

Каждой букве сопоставляем *кумулятивную вероятность*

$$q_1 = 0, \quad q_2 = p_1, \quad \ldots, \quad q_M = \sum_{i=1}^{M-1} p_i.$$

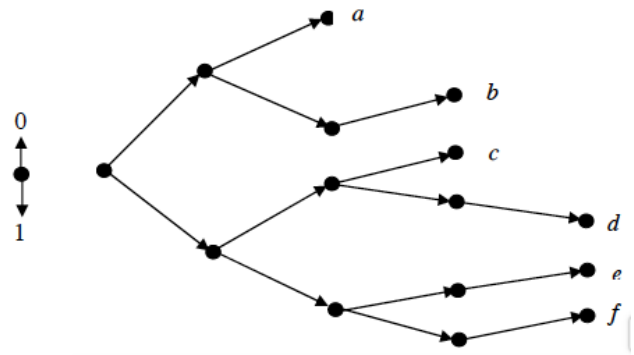
Кодовым словам кода Шенона для сообщения с номером m является двоичная последовательность, представляющая собой первые $l_m = |-\log p_m|$ разрядов после запятой в двоичной записи числа q_m

Пример:

X={a,b,c,d,e,f} с вероятностью {0,35; 0,2; 0,15; 0,1; 0,1; 0,1}

Найти: среднею длину кодовых слов и энтропию.

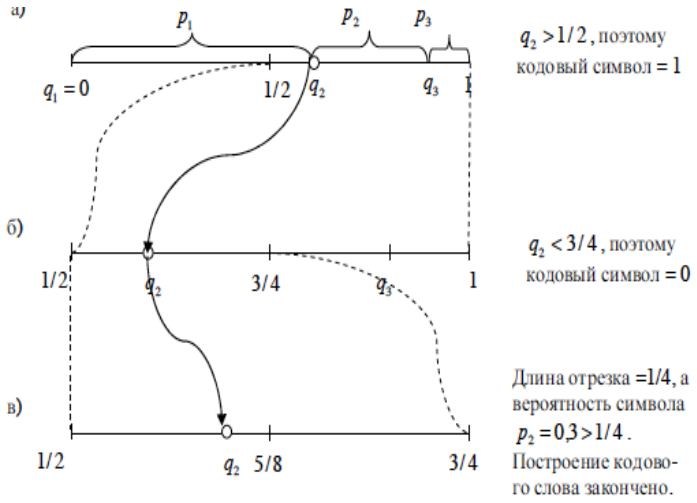
x	p_m	q_m	l_m	Двоичная запись q_m	Кодовое слово $\boldsymbol{c}_m$
a	0,35	0,00	2	0,00...	00
b	0,20	0,35	3	0,0101...	010
c	0,15	0,55	3	0,10001...	100
d	0,10	0,70	4	0,10110...	1011
e	0,10	0,80	4	0,11001...	1100
f	0,10	0,90	4	0,11100...	1110



Пример: M=3, p_1=0,6, p_2=0,3, p_3=0,1, соответственно:

q_1=0, q_2=0,6, q_3=0,9

Закодировать: m=2



3.2 КОД ГИЛЬБЕРТА МУРА

Рассмотрим источник, выбирающий буквы из алфавита X= {1,…M} с вероятностью {p_1,….p_M}, сопоставим каждой букве m=1,..M кумулятивную вероятность $q_m = \Sigma^{m-1}_{i=1} p_i$ и вычислим $\sigma_m = q_m + p_m/2$

Кодовым словом будет x_m первые $l_m = \lceil -\log(p_m/2) \rceil$ разрядов после запятой в записи σ_m.

x	$p(x)$	$q(x)$	$\sigma(x)$	$l(x)$	$c(x)$
a	0.35	0	0.175	3	001...
b	0.20	0.35	0.450	4	0111...
c	0.15	0.55	0.625	4	1010...
d	0.1	0.70	0.750	5	11000...
e	0.1	0.80	0.850	5	11011...
f	0.1	0.90	0.950	5	11110...

Задание: найти длину кодового слова.

3.3. Неравномерное кодирование для стационарного источника

Рассмотрим последовательность $x_1,....x_i$ принадлежащую X={x}.

Пусть указан некоторый способ кодирования, который для любых n для каждой последовательности $\boldsymbol{x} \in X^n$ на выходе источника строит кодовое слово $\boldsymbol{c}(\boldsymbol{x})$ длины $l(\boldsymbol{x})$. Тогда *средняя скорость кодирования для блоков длины* n определяется как

$$\bar{R}_n = \frac{1}{n}\mathbf{M}\left[l(\boldsymbol{x})\right]$$

Подбирая длину блоков, при которой средняя скорость будет наименьшей, получаем следующее определение для *средней скорости кодирования:* $\bar{R} = \inf_n \bar{R}_n.$

Рассмотрим FV кодирование - (fixed-to-variable).

Теорема 2.5 *Для дискретного стационарного источника с энтропией на сообщение H для любого FV-кодирования имеет место неравенство*

$$\bar{R} \geq H.$$

Теорема 2.6 *Для дискретного стационарного источника с энтропией на сообщение H и для любого $\delta > 0$ существует способ неравномерного FV-кодирования такой, для которого*

$$\bar{R} \leq H + \delta.$$

3.3 Арифметическое кодирование

Рассмотрим последнею часть данной лекции – арифметическое кодирование.

Рассмотрим дискретный постоянный источник:

1. X= {1,....M}
2. с вероятностью сообщений $\{p_1,....p_M\}$

3. $\{q_1,\ldots..q_M\}$ кумулятивная вероятность

Задача: Кодирование последовательности множества $X^n=\{x\}$, $x=(x_1,\ldots.x_n)$ а $x^i_j=(x_i,\ldots.x_j)$

Для последовательностей длины 1 (для отдельных сообщений из X) мы считаем, что сообщение с меньшим номером предшествуют сообщению с большим номером. Если, например, элементы X – числа, то $x \prec x'$, если $x < x'$, $x, x' \in X$.

Для двух последовательностей $\boldsymbol{x} = (x_1, ..., x_n)$, $\boldsymbol{y} = (y_1, ..., y_n)$ обозначим через i наименьший индекс такой, что $x_i \neq y_i$. Тогда $\boldsymbol{y} \prec \boldsymbol{x}$, если $y_i \prec x_i$.

Задача: Найти $q(\boldsymbol{x}) = \sum_{\boldsymbol{u} \prec \boldsymbol{x}} p(\boldsymbol{y})$, p(x) вычисляется по формуле $p(\boldsymbol{x}) = \prod_{i=1} p(x_i)$.

$$
\begin{aligned}
q(\boldsymbol{x}_1^n) &= \sum_{\boldsymbol{y}_1^n \prec \mathbf{x}_1^n} p(\boldsymbol{y}_1^n) = \\
&= \sum_{\boldsymbol{y}_1^{n-1} \prec \boldsymbol{x}_1^{n-1}} \sum_{y_n} p(\boldsymbol{y}_1^{n-1} y_n) + \sum_{\boldsymbol{y}_1^{n-1} = \mathbf{x}_1^{n-1}} \sum_{y_n \prec x_n} p(\boldsymbol{y}_1^{n-1} y_n) = \\
&= \sum_{\boldsymbol{y}_1^{n-1} \prec \boldsymbol{x}_1^{n-1}} p(\boldsymbol{y}_1^{n-1}) + \sum_{\boldsymbol{y}_1^{n-1} = \boldsymbol{x}_1^{n-1}} p(\boldsymbol{y}_1^{n-1}) \sum_{y_n \prec x_n} p(y_n) = \\
&= q(\boldsymbol{x}_1^{n-1}) + p(\boldsymbol{x}_1^{n-1}) q(x_n),
\end{aligned}
$$

Алгоритмически это можно записать следующим образом:

Input: Объем алфавита M
вероятности букв $p_i, i = 1, ..., M$
длина последовательности n
последовательность на выходе источника $(x_1, ..., x_n)$,
Output: Кодовое слово арифметического кода

`Кумулятивные вероятности:`
$q_1 = 0$;
for $i = 2$ to M **do**
  $q_i = q_{i-1} + p_{i-1}$;
end

`Кодирование:`
for $i = 1$ to n **do**
  $F \leftarrow F + q(x_i)G$;
  $G \leftarrow p(x_i)G$;.
end

`Формирование кодового слова:`
$\boldsymbol{c} \leftarrow$ первые $\lceil -\log G \rceil + 1$ разрядов после запятой в двоичной записи числа $F + G/2$.

Пример: $X=\{a,b,c\}$, $p_a=0,1$, $p_b=0,6$, $p_c=0,3$. Кодируем x=(bcbab)

Шаг i	x_i	$p(x_i)$	$q(x_i)$	F	G
0	-	-	-	0,0000	1,0000
1	b	0,6	0,1	0,1000	0,6000
2	c	0,3	0,7	0,5200	0,1800
3	b	0,6	0,1	0,5380	0,1080
4	a	0,1	0,0	0,5380	0,0108
5	b	0,6	0,1	0,5391	0,0065
6	Длина кодового слова $\lceil -\log G + 1 \rceil = 9$ Кодовое слово $F + G/2 = 0,5423... \rightarrow$ $\rightarrow \hat{F} = 0,541 \rightarrow 100010101$				

Input: Объем алфавита M
вероятности букв $\{p_1, ..., p_M\}$
кумулятивные вероятности букв $q_i, i = 1, ..., M$
длина декодируемой последовательности n
кодовое слово в виде числа $\hat{F}$.
Output: Декодированная последовательность букв $(x_1, ..., x_n)$

Инициализация: $q_{M+1} = 1$; $S = 0$; $G = 1$.

Декодирование:
for $i = 1$ to n do
  $j = 1$;
  while $S + q_{j+1}G < \hat{F}$ do
    $j \leftarrow j + 1$.
  end
  $S \leftarrow S + q_jG$;
  $G \leftarrow p_jG$;
  $x_i = j$.
end

Результат: последовательность $(x_1, ..., x_n)$;

Пример: X={a,b,c}, p_a=0,1, p_b=0,6, p_c=0,3.

Декодируем 0100010101

Шаг	S	G	Гипотеза x	$q(x)$	$S + qG$	Решение x_i	$p(x)$
0	100010101$\rightarrow \hat{F}$ =0,541						
1	0,0000	1,0000	a	0,0	0,0000< $\hat{F}$	b	0,6
			b	0,1	**0,1000**< $\hat{F}$		
			c	0,7	0,7000> $\hat{F}$		
2	0,1000	0,6000	a	0,0	0,1000< $\hat{F}$	c	0,3
			b	0,1	0,1600< $\hat{F}$		
			c	0,7	**0,5200**< $\hat{F}$		
3	0,5200	0,1800	a	0,0	0,5200< $\hat{F}$	b	0,6
			b	0,1	**0,5380**< $\hat{F}$		
			c	0,7	0,6460> $\hat{F}$		
4	0,5380	0,1080	a	0,0	**0,5380**< $\hat{F}$	a	0,1
			b	0,1	0,5488> $\hat{F}$		
5	0,5380	0,0108	a	0,0	0,5380< $\hat{F}$	b	0,6
			b	0,1	**0,5391**< $\hat{F}$		
			c	0,7	0,5456> $\hat{F}$		

4 ЗАКЛЮЧЕНИЕ

На данной лекции были рассмотрены основные принципами сжатия информации, принципы неравномерного кодирования, код Шеннона, код Гильберта Мура, кодирование стационарного источника, арифметическое кодирование, кодирование дискретных источников, дискретный стационарный процесс, префиксный код, неравенство Крафта, код Хафманна.