ОПТИМАЛЬНОЕ ПОБУКВЕННОЕ КОДИРОВАНИЕ

4.1 *Определения*

- **При кодировании сообщений считается, что символы сообщения порождаются некоторым *источником информации*.**

- **Если вероятностный источник с алфавитом** $A = \{a_1, a_2, ..., a_n\}$

  порождает символы алфавита независимо друг от друга, т.е. знание предшествующих символов не влияет на вероятность последующих, то такой источник называется *бернуллиевским*.

- **Для любого сообщения** $x_1 x_2 \ldots x_L$ **порождаемого бернуллиевским источником, выполняется равенство:**

$$P(x_1 x_2 \ldots x_L) = P(x_1) \cdot P(x_2) \cdot \ldots \cdot P(x_L)$$

где $P(x)$ **– вероятность появления символа *x*,**

$P(x_1 x_2 \ldots x_L)$ **– вероятность появления последовательности** $x_1 x_2 \ldots x_L$

- **Пусть имеется дискретный вероятностный источник без памяти, порождающий символы алфавита**

$$A = \{a_1, a_2, ..., a_n\}$$

с вероятностями

$$p_i = P(a_i) \qquad \sum_{i=1}^{n} p_i = 1$$

- **Пусть имеется разделимый двоичный побуквенный код для бернуллиевского источника, порождающего символы алфавита**

$$A = \{a_1, a_2, ..., a_n\}$$

с вероятностями $p_i = P(a_i)$

состоящий из *n* кодовых слов

с длинами $L_1, ..., L_n$

Средней длиной кодового слова

называется величина $L_{cp} = \sum_{i=1}^{n} p_i L_i$

которая показывает среднее число кодовых букв на одну букву источника.

Пример

- **Пусть имеются два источника с одним и тем же алфавитом** $A = \{a_1, a_2, a_3\}$ **но с разными вероятностными распределениями**

  $P = \{1/3, 1/3, 1/3\}$ **и** $Q = \{1/4, 1/4, 1/2\}$ **,**

  которые кодируются одним и тем же кодом $\sigma = \langle a_1 \to 10, a_2 \to 000, a_3 \to 01 \rangle$

- **Средняя длина кодового слова для разных источников будет различной**

$$L_{cp}(P) = \frac{1}{3} \cdot 2 + \frac{1}{3} \cdot 3 + \frac{1}{3} \cdot 2 = \frac{7}{3} \approx 2.33$$

$$L_{cp}(Q) = \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 3 + \frac{1}{2} \cdot 2 = \frac{9}{4} = 2.25$$

- **Побуквенный разделимый код называется *оптимальным*, если средняя длина кодового слова *минимальна* среди всех побуквенных разделимых кодов для данного распределения вероятностей символов**

4.2 Теоремы Шеннона

- **Взаимосвязь между средней длиной кодового слова и энтропией дискретного вероятностного источника при побуквенном кодировании выражает следующая теорема.**

Прямая теорема кодирования

Для бернуллиевского источника с алфавитом $A = \{a_1, a_2, ..., a_n\}$

и вероятностями $p_i = P(a_i)$ $\sum_{i=1}^{n} p_i = 1$

существует разделимый побуквенный код, у которого средняя длина кодового слова превосходит энтропию не больше, чем на единицу

$$L_{cp} < H(p_1, ..., p_n) + 1.$$

Обратная теорема кодирования

- ***Для бернуллиевского источника с алфавитом*** $A = \{a_1, a_2, ..., a_n\}$

  и вероятностями $p_i = P(a_i) \quad \sum_{i=1}^{n} p_i = 1$

  и любого разделимого побуквенного кода средняя длина кодового слова всегда не меньше энтропии источника

$$L_{cp} \geq H(p_1, ..., p_n)$$

Доказательство обратной теоремы

- Поскольку код разделимый, то для него верно неравенство МакМиллана

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

- Применим известное неравенство

$$\ln x \leq x\text{-}1$$

$$H(p_1,\ldots,p_n) - L_{cp} = -\sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{n} p_i L_i = \sum_{i=1}^{n} p_i(-\log p_i + \log 2^{-L_i}) =$$

$$= \sum_{i=1}^{n} p_i \log \frac{2^{-L_i}}{p_i} \leq \sum_{i=1}^{n} p_i \left(\frac{2^{-L_i}}{p_i} - 1\right) \log e = \left(\sum_{i=1}^{n} 2^{-L_i} - 1\right) \log e \leq 0$$

$$H(p_1,\ldots,p_n) \leq L_{cp}$$

- Второе неравенство доказывается конструктивно с помощью кода, построенного по методу Шеннона
- Каждому символу источника будет соответствовать кодовое слово из $\left\lceil \log \frac{1}{p_i} \right\rceil$ двоичных цифр числа $q_i = \sum_{j=1}^{i-1} p_j$

Вероятности упорядочены по убыванию

- ***Избыточностью кода*** **называется разность между средней длиной кодового слова и предельной энтропией источника сообщений**

$$r = L_{ср} - H(p_1,\ldots,p_n)$$

Следствие

- ***Для существования разделимого кода с нулевой избыточнотью*** $r = 0$
- ***для бернуллиевского источника с алфавитом*** $A = \{a_1, a_2, ..., a_n\}$
  и вероятностями $p_i = P(a_i)$
  необходимо и достаточно, чтобы все вероятности сообщений источника
- ***имели вид*** $p_i = 2^{-L_i}$
  где $\{L_i\}$ ***целые положительные числа***

4.3 *Почти оптимальное кодирование*

Метод Шеннона

- **Метод кодирования, предложенный К. Шенноном, позволяет построить почти оптимальный двоичный префиксный код.**

- **Пусть имеется дискретный вероятностный источник, порождающий символы алфавита** $A = \{a_1, a_2, ..., a_n\}$

  с вероятностями $p_i = P(a_i)$

  при этом символы исходного алфавита упорядочены по убыванию их вероятностей, т.е. $p_1 \geq p_2 \geq ... \geq p_n$

Код Шеннона строится следующим образом:

- 1. Вычисляются кумулятивные вероятности

$$Q_i = \sum_{j=1}^{i-1} p_j \qquad i = 0,\ldots,n-1$$

- 2. В качестве кода символа a_i берут $L_i = \lceil -\log p_i \rceil$ первых двоичных цифр после запятой числа Q_{i-1} $i = 1,\ldots,n$

- **Утверждение**

  Код Шеннона является префиксным.

доказательство прямой теоремы

- $$L_{cp} = \sum p_i L_i = \sum p_i \left\lceil \log \frac{1}{p_i} \right\rceil < \sum p_i (\log \frac{1}{p_i} + 1) = H(p_1, ..., p_n) + 1$$

Пример

- Пусть источник имеет алфавит

$$A = \{a_1, a_2, a_3, a_4, a_5, a_6\}$$

- с вероятностями

$$p_1 = 0.36 \qquad p_4 = 0.12$$

$$p_2 = 0.18 \qquad p_5 = 0.09$$

$$p_3 = 0.18 \qquad p_6 = 0.07$$

- Вычислим кумулятивные вероятности

$$Q_0 = 0$$

$$Q_1 = p_1 = 0.36$$

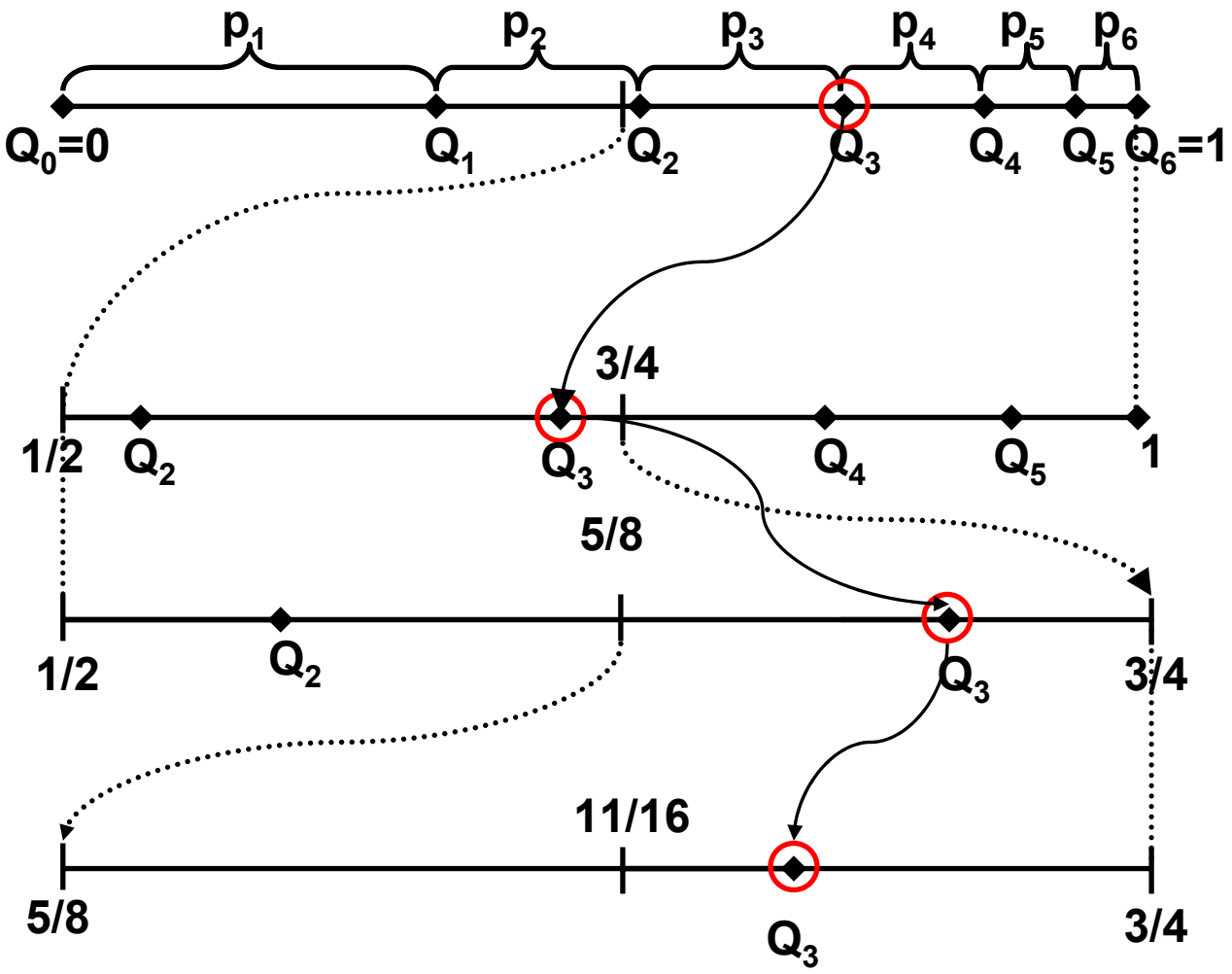
$$Q_2 = p_1 + p_2 = 0.54$$

$$Q_3 = p_1 + p_2 + p_3 = 0.72$$

$$Q_4 = p_1 + p_2 + p_3 + p_4 = 0.84$$

$$Q_5 = p_1 + p_2 + p_3 + p_4 + p_5 = 0.93$$

Кодирование символа a_4



Q_3>1/2, поэтому кодовый символ 1

Q_3<3/4, поэтому кодовый символ 0

Q_3>5/8, поэтому кодовый символ 1

Q_3>11/16, поэтому кодовый символ 1

- **Для вероятностей, представленных в виде десятичных дробей, удобно определять длину кодового слова L_i из соотношения**

$$\frac{1}{2^{L_i}} \le p_i < \frac{1}{2^{L_i-1}} \qquad i = 1,\ldots,n$$

,

$$\frac{1}{2^{L_1}} \le 0.36 < \frac{1}{2^{L_1-1}} \qquad L_1 = 2 \qquad \frac{1}{2^{L_4}} \le 0.12 < \frac{1}{2^{L_4-1}} \qquad L_4 = 3$$

$$\frac{1}{2^{L_2}} \le 0.18 < \frac{1}{2^{L_2-1}} \qquad L_2 = 3 \qquad \frac{1}{2^{L_5}} \le 0.09 < \frac{1}{2^{L_5-1}} \qquad L_5 = 4$$

$$\frac{1}{2^{L_3}} \le 0.18 < \frac{1}{2^{L_3-1}} \qquad L_3 = 3 \qquad \frac{1}{2^{L_6}} \le 0.07 < \frac{1}{2^{L_6-1}} \qquad L_6 = 4$$

Код Шеннона

a_i	P_i	Q_{i-1}	L_i	кодовое слово
a_1	$1/2^2 \le 0.36 < 1/2$	0	2	00
a_2	$1/2^3 \le 0.18 < 1/2^2$	0.36	3	010
a_3	$1/2^3 \le 0.18 < 1/2^2$	0.54	3	100
a_4	$1/2^4 \le 0.12 < 1/2^3$	0.72	4	1011
a_5	$1/2^4 \le 0.09 < 1/2^3$	0.84	4	1101
a_6	$1/2^4 \le 0.07 < 1/2^3$	0.93	4	1110

$$L_{cp} = 0.36 \cdot 2 + (0.18 + 0.18) \cdot 3 +$$
$$+ (0.12 + 0.09 + 0.07) \cdot 4 = 2.92 < 2.37 + 1$$

Код Фано

- **Пусть имеется дискретный вероятностный источник, порождающий символы алфавита** $A = \{a_1, a_2, ..., a_n\}$

  с вероятностями $p_i = P(a_i)$

  при этом символы исходного алфавита упорядочены по убыванию их вероятностей, т.е. $p_1 \geq p_2 \geq ... \geq p_n$

- **Упорядоченный по убыванию вероятностей список букв алфавита источника разбивается на две части.**
- **Кодам символов из первой части списка приписывается 0, а символам из второй части – 1. Далее каждую из частей списка разбивают на две части и т.д.**
- **Процесс продолжается до тех пор, пока весь список не разобьется на части, содержащие по одному символу.**

Пример

- Пусть источник имеет алфавит

$$A = \{a_1, a_2, a_3, a_4, a_5, a_6\}$$

- с вероятностями

$$p_1 = 0.36 \qquad p_4 = 0.12$$

$$p_2 = 0.18 \qquad p_5 = 0.09$$

$$p_3 = 0.18 \qquad p_6 = 0.07$$

Разбиение множества символов

1-ый блок	$S1 = \sum p_i$	2-ой блок	$S2 = \sum p_j$	$\Delta S = \lvert S1-S2 \rvert$
$a_1\ a_2\ a_3\ a_4\ a_5$	0,93	a_6	0,07	0,86
$a_1\ a_2\ a_3\ a_4$	0,84	$a_5\ a_6$	0,16	0,68
$a_1\ a_2\ a_3$	0,72	$a_4\ a_5\ a_6$	0,28	0,44
$a_1\ a_2$	0,54	$a_3\ a_4\ a_5\ a_6$	0,46	0,08
a_1	0,36	$a_2\ a_3\ a_4\ a_5\ a_6$	0,64	0,28

Разбиение множества $a_3a_4a_5a_6$

1-ый блок	$S1 = \sum p_i$	2-ой блок	$S2 = \sum p_j$	$\Delta S = \lvert S1-S2 \rvert$
$a_3\ a_4\ a_5$	0,39	a_6	0,07	0,32
$a_3\ a_4$	0,30	$a_5\ a_6$	0,16	0,14
a_3	0,18	$a_4\ a_5\ a_6$	0,28	0,10

Разбиение множества $a_4 a_5 a_6$.

1-ый блок	$S1 = \sum p_i$	2-ой блок	$S2 = \sum p_j$	$\Delta S = \lvert S1-S2 \rvert$
$a_4\ a_5$	0,21	a_6	0,07	0,14
a_4	0,12	$a_5\ a_6$	0,16	0,04

a_i	P_i	кодовое слово				L_i
a_1	0.36	0	0			2
a_2	0.18	0	1			2
a_3	0.18	1	0			2
a_4	0.12	1	1	0		3
a_5	0.09	1	1	1	0	3
a_6	0.07	1	1	1	1	4

- Полученный код является префиксным и почти оптимальным со средней длиной кодового слова

$$L_{cp} = 0.36 \cdot 2 + 0.18 \cdot 2 + 0.18 \cdot 2 + 0.12 \cdot 3 + 0.09 \cdot 4 + 0.07 \cdot 4 = 2.44$$

Алфавитный код Гилберта – Мура

- Пусть символы алфавита источника некоторым образом упорядочены

$$a_1 \le a_2 \le \ldots \le a_n$$

- Код σ называется *алфавитным*, если кодовые слова лексикографически упорядочены, т.е. $\sigma(a_1) \le \sigma(a_2) \le \ldots \le \sigma(a_n)$

- Вычисляются величины $q_i \quad i = 1,\ldots,n$

$$q_1 = Q_0 + \frac{p_1}{2}$$

$$q_2 = Q_1 + \frac{p_2}{2}$$

$$q_3 = Q_2 + \frac{p_3}{2}$$

$$\ldots$$

$$q_n = Q_{n-1} + \frac{p_n}{2}$$

- где Q_i кумулятивные вероятности $i = 1,\ldots,n$

- В качестве кода символа $a_i \quad i = 0,\ldots,n-1$
- берут $L_i = \left\lceil -\log \frac{p_i}{2} \right\rceil$ первых двоичных цифр после запятой числа $q_i \quad i = 1,\ldots,n$

- **Утверждение**

  Код Гилберта-Мура является алфавитным и префиксным.

- **Утверждение** Средняя длина кодового слова кода Гилберта-Мура удовлетворяет соотношению

$$L_{cp} < H(p_1,\ldots,p_n) + 2$$

- **Доказательство.** Действительно,

$$L_{cp} = \sum_{i=1}^{n} L_i p_i = \sum_{i=1}^{n} \left\lceil -\log \frac{p_i}{2} \right\rceil p_i <$$

$$< \sum_{i=1}^{n} (-\log p_i + 1 + 1) p_i = H(p_1, \ldots, p_n) + 2$$

Пример

- Пусть источник имеет алфавит

$$A = \{a_1, a_2, a_3, a_4, a_5, a_6\}$$

- с вероятностями

$$p_1 = 0.07 \qquad p_4 = 0.12$$

$$p_2 = 0.18 \qquad p_5 = 0.09$$

$$p_3 = 0.18 \qquad p_6 = 0.36$$

Вычислим величины

$$q_1 = Q_0 + \frac{p_1}{2} = 0 + 0.035 = 0.035$$

$$q_2 = Q_1 + \frac{p_2}{2} = 0.7 + 0.9 = 0.16$$

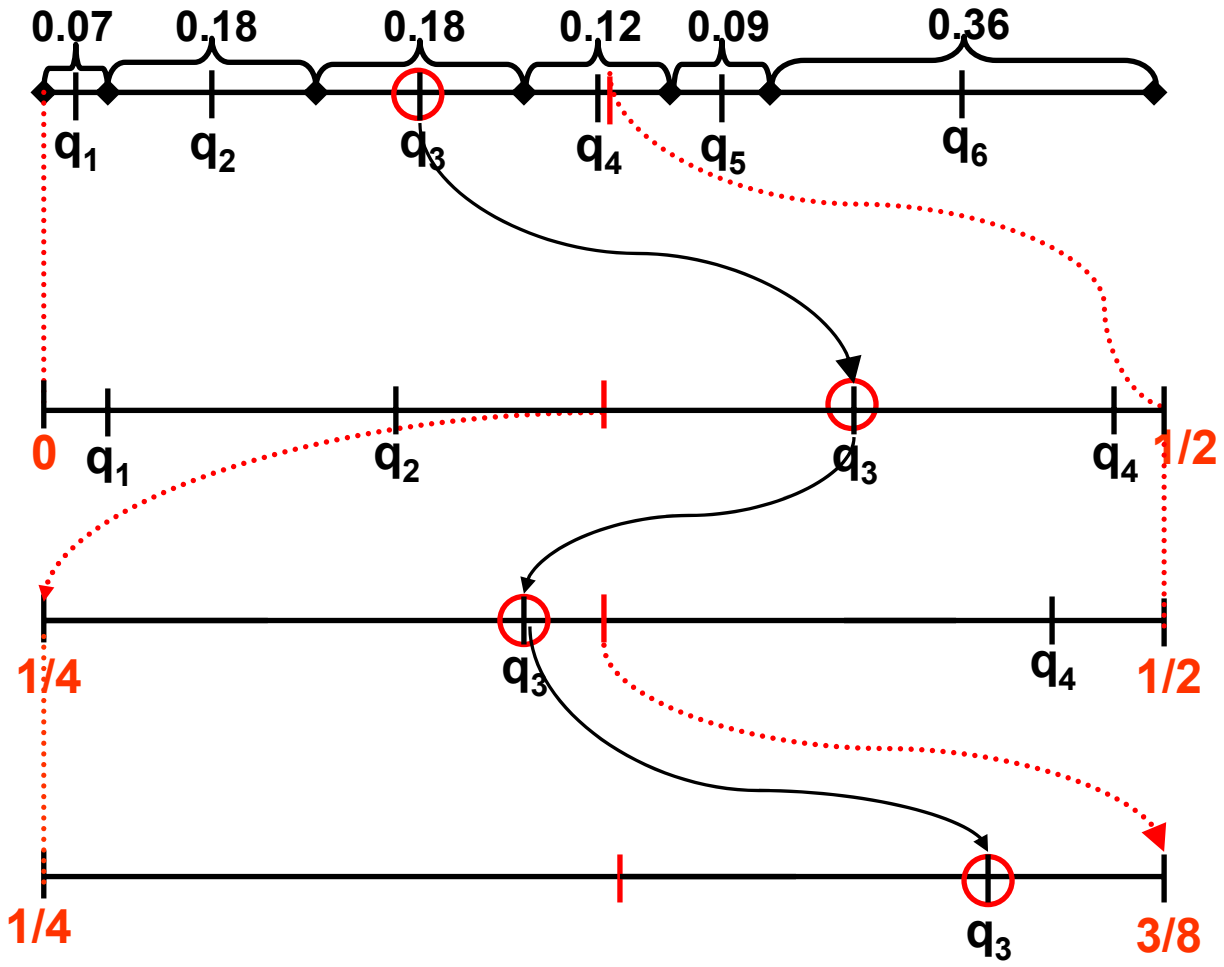
$$q_3 = Q_2 + \frac{p_3}{2} = 0.25 + 0.09 = 0.34$$

$$q_4 = Q_3 + \frac{p_4}{2} = 0.43 + 0.06 = 0.49$$

$$q_5 = Q_4 + \frac{p_5}{2} = 0.55 + 0.045 = 0.595$$

$$q_6 = Q_5 + \frac{p_6}{2} = 0.64 + 0.18 = 0.82$$

Кодирование символа a_3



q_3<1/2, поэтому кодовый символ 0

q_3>1/4, поэтому кодовый символ 1

q_3<3/8, поэтому кодовый символ 0

Q_5

q_3>5/16, поэтому кодовый символ 1

Код *Гилберта-Мура*

a_i	q_i	L_i	кодовое слово
a_1	0.035	5	00001
a_2	0.16	4	0010
a_3	0.34	4	0101
a_4	0.49	5	01111
a_5	0.595	5	10001
a_6	0.82	3	110

- *Lcp*=4.0.18+4.0.18+3.0.36+5.0.07+
  - +5.0.09+5.0.12=3.92<2.37+2