4.4 Свойства оптимального побуквенного кода

Лемма 1

- ***Для оптимального кода с длинами кодовых слов*** $L_1,...,L_n$

  верно соотношение $L_1 \le L_2 \le ... \le L_n$

  если $p_1 \ge p_2 \ge ... \ge p_n$

Доказательство (от противного)

Пусть есть индексы *i* и *j* такие, что $L_i > L_j$ **при** $p_i > p_j$

Тогда

$$L_i p_i + L_j p_j = L_i p_i + L_j p_j + (L_i p_j + L_j p_i) - (L_i p_j + L_j p_i) =$$
$$= p_i (L_i - L_j) - p_j (L_i - L_j) + L_i p_j + L_j p_i =$$
$$= (p_i - p_j)(L_i - L_j) + L_i p_j + L_j p_i > L_i p_j + L_j p_i,$$

- **если поменяем местами L_i и L_j ,**
  то получим код, имеющий меньшую среднюю длину кодового слова, что противоречит оптимальности кода.

  Лемма 1 доказана

Лемма 2

- ***Пусть*** $\sigma = \langle a_1 \to b_1, \ldots, a_n \to b_n \rangle$
  схема оптимального префиксного кодирования для источника с распределением вероятностей

$$p_1 \geq p_2 \ldots \geq p_n > 0$$

Тогда среди элементарных кодов, имеющих максимальную длину, существуют два, которые различаются только в последнем разряде.

- Пусть $\sigma = \langle a_1 \to b_1, \ldots, a_n \to b_n \rangle$ -- схема кодирования для источника с распределением вероятностей

$$p_1 \geq p_2 \ldots \geq p_n > 0$$

- Рассмотрим новый источник с распределением $\{p'_1, p'_2, \ldots p'_{n-1}\}$

  причем

$$p'_1 = p_1 \quad \cdots \quad p'_{n-2} = p_{n-2} \qquad p'_{n-1} = p_{n-1} + p_n$$

- Построим код σ' по следующему правилу

$$\sigma' = \left\langle \begin{array}{l} a_1 \to b_1, \\ \ldots, \\ a_{n-2} \to b_{n-2} \\ a_{n-1} \to b'_{n-1} \end{array} \right\rangle$$

- где b'_{n-1} общая часть кодов b_{n-1} и b_n

Лемма 3

- ***Если схема кодирования*** σ' ***оптимальная,***

  то и схема σ ***оптимальная***

4.5 Оптимальный код Хаффмана

- **Метод оптимального побуквенного кодирования был разработан в 1952 г. Д. Хаффманом.**

- **Оптимальный двоичный код Хаффмана обладает минимальной средней длиной кодового слова среди всех побуквенных кодов для данного источника с алфавитом**

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

- **и вероятностями**

$$p_i = P(a_i) \qquad \sum_{i=1}^{n} p_i = 1 \qquad p_1 \geq p_2 \geq ... \geq p_n$$

- **Алгоритм построения оптимального кода Хаффмана основывается на утверждениях предыдущих лемм и заключается в следующем**

- Если $A = \{a_1, a_2\}$,
- то $a_1 \to 0 \quad a_2 \to 1$

- Если $A = \{a_1, a_2, ..., a_j, ..., a_n\}$

и известны коды $\langle a_j \to b_j \rangle \quad j = 1, \ldots, n$

то для алфавита $A' = \{a_1, a_2, ..., a'_j, a''_j ..., a_n\}$

с новыми символами a'_j и a''_j (вместо a_j)

и вероятностями $p_j = p'_j + p''_j$

код символа a_j заменяется на коды

$$a'_j \to b_j 0 \qquad a''_j \to b_j 1$$

Процесс построения кодов Хаффмана происходит в два этапа.

- На первом этапе складываются две наименьшие вероятности и суммарная вероятность включается на соответствующее место в упорядоченном массиве вероятностей так, чтобы массив остался упорядоченным.
- Это происходит до тех пор, пока в массиве не останется две вероятности.

Второй этап заключается в построении кодов символов.

- **Если в массиве вероятностей всего два значения, то символы источника кодируются 0 и 1.**
- **Если вероятность в массиве получилась в результате слияния двух наименьших вероятностей, то из имеющегося кода строится два кода, добавлением 0 и 1 справа, т.е. новые коды будут отличаться только последним битом.**

- **Утверждение.** *Код Хаффмана является префиксным.*

Пример

- Пусть источник имеет алфавит

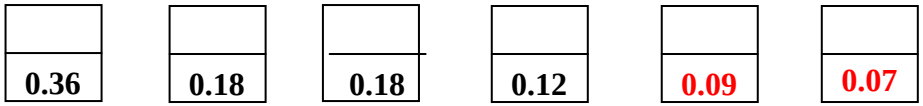
$$A = \{a_1, a_2, a_3, a_4, a_5, a_6\}$$

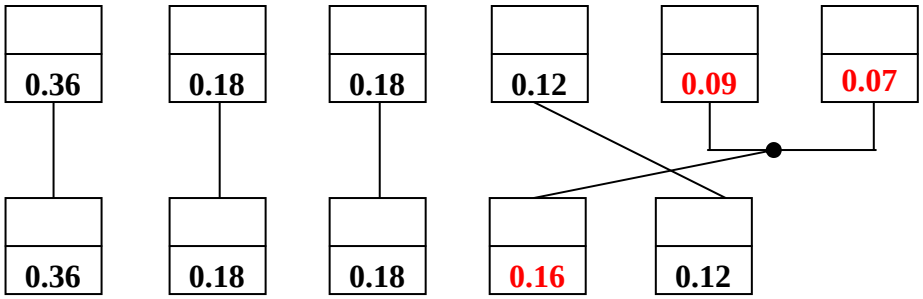
- с вероятностями

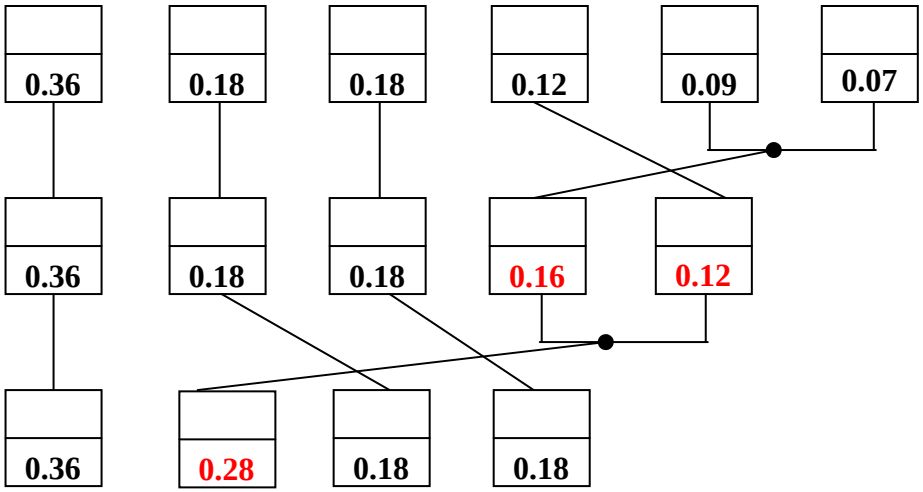
$$p_1 = 0.36 \qquad p_4 = 0.12$$

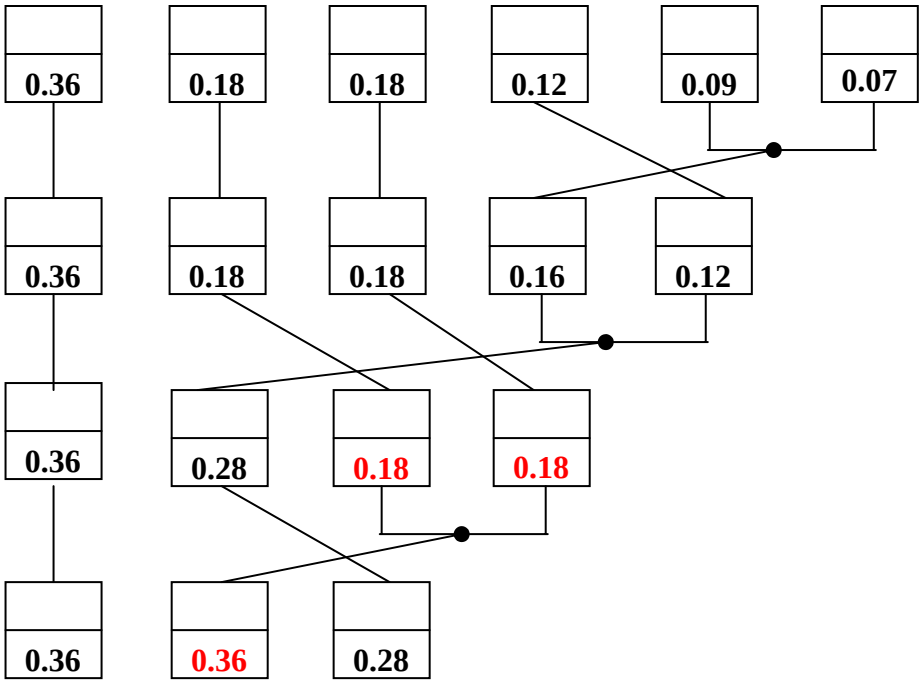
$$p_2 = 0.18 \qquad p_5 = 0.09$$

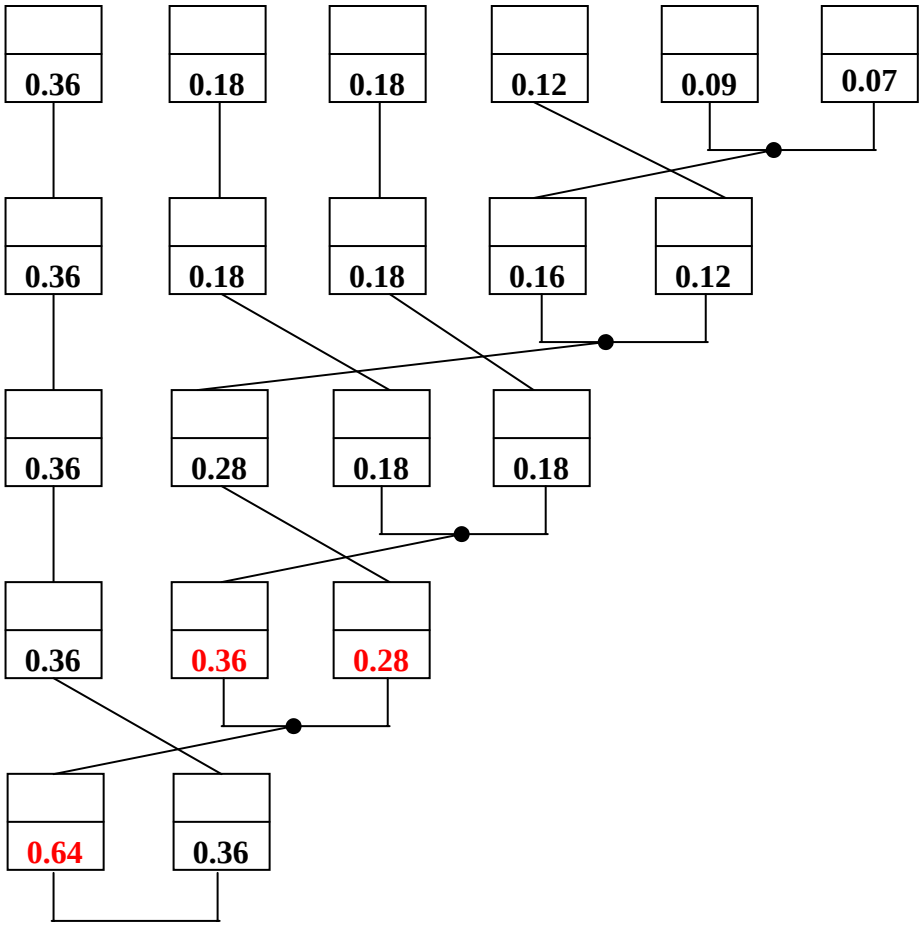
$$p_3 = 0.18 \qquad p_6 = 0.07$$

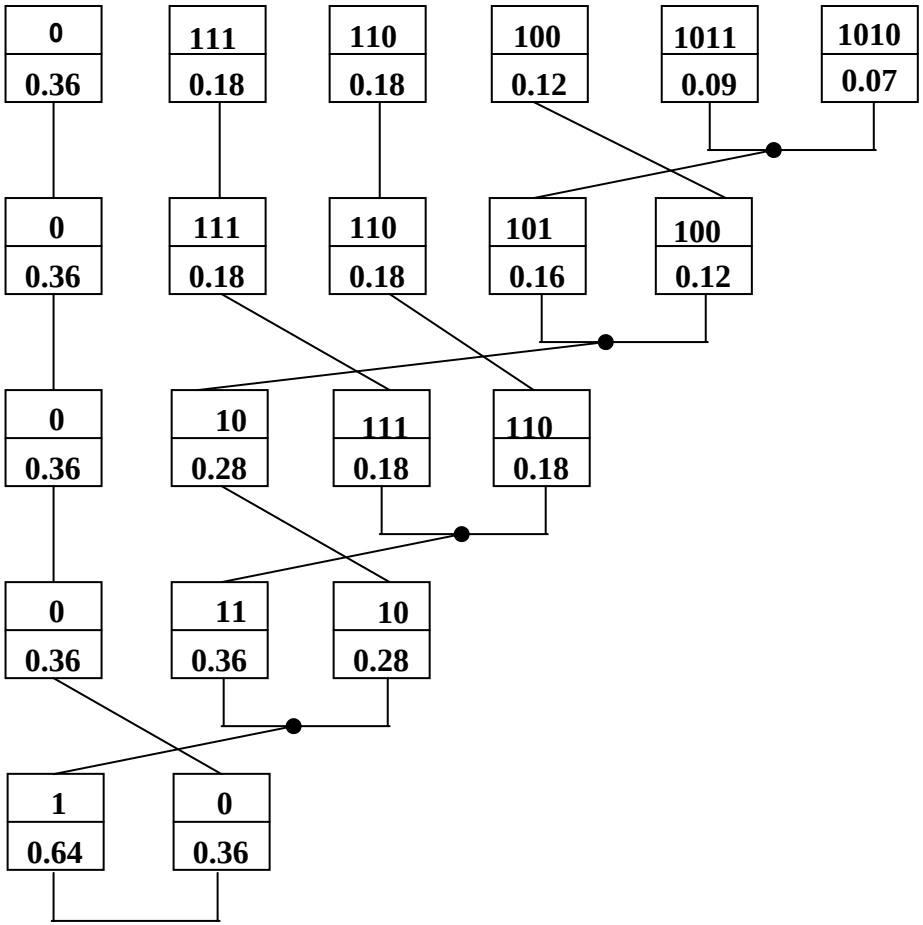

0.36
0.18
0.18
0.12
0.09
0.07


0.36
0.18
0.18
0.12
0.09
0.07
0.36
0.18
0.18
0.16
0.12


0.36
0.18
0.18
0.12
0.09
0.07
0.36
0.18
0.18
0.16
0.12
0.36
0.28
0.18
0.18


0.36
0.18
0.18
0.12
0.09
0.07
0.36
0.18
0.18
0.16
0.12
0.36
0.28
0.18
0.18
0.36
0.36
0.28


0.36
0.18
0.18
0.12
0.09
0.07
0.36
0.18
0.18
0.16
0.12
0.36
0.28
0.18
0.18
0.36
0.36
0.28
0.64
0.36


0
0.36
111
0.18
110
0.18
100
0.12
1011
0.09
1010
0.07
0
0.36
111
0.18
110
0.18
101
0.16
100
0.12
0
0.36
10
0.28
111
0.18
110
0.18
0
0.36
11
0.36
10
0.28
1
0.64
0
0.36

Код Хаффмана

a_i	p_i	L_i	кодовое слово
a_1	0.36	1	0
a_2	0.18	3	111
a_3	0.18	3	110
a_4	0.12	3	100
a_5	0.09	4	1011
a_6	0.07	4	1010

$$L_{cp}(P) = 0.36 \cdot 1 + 0.18 \cdot 3 + 0.18 \cdot 3 +$$
$$+ 0.12 \cdot 3 + 0.09 \cdot 4 + 0.07 \cdot 4 = 2.44 > 2.37$$

- **Для восстановления содержимого сообщения декодер должен знать таблицу частот, которой пользовался кодер.**
- **Следовательно, длина сжатого сообщения увеличивается на длину таблицы частот, которая должна посылаться впереди данных.**

- **Кроме того необходимость наличия полной частотной статистики перед началом собственно кодирования требует двух проходов по сообщению:**
  - **одного для построения модели сообщения (таблицы частот и дерева кодирования),**
  - **другого для собственно кодирования.**

Блочное кодирование

- **Для уменьшения избыточности кодирования используют принцип блочности**
- **Сообщение разбивается на последовательности одинаковой длины (блоки). Каждый блок кодируется отдельно.**

- Для известного бернуллиевского источника известны не только вероятности появления отдельных символов, но и вероятности появления всех последовательностей символов (как произведение вероятностей символов блока)
- Таким образом, блок длины L можно считать «буквой» нового алфавита с определенным вероятностным распределением

- **Можно получить более сильные результаты, если кодовые слова приписывать не отдельным буквам, а сообщениям (блокам из *L* букв) источника.**

Теорема

В случае бернуллиевского стационарного источника существует префиксный код для кодирования блоков длины L, такой, что для любого $\varepsilon > 0$

можно выбрать достаточно большое L, чтобы величина L_{cp}

удовлетворяла неравенствам:

$$H(p_1,\ldots,p_n) \le L_{cp} \le H(p_1,\ldots,p_n) + \varepsilon$$

Теорема

Пусть H_L ***– энтропия на букву в блоке длины L дискретного источника. Тогда существует префиксный код для кодирования блоков длины L, такой, что средняя длина кодового слова*** L_{cp} ***будет удовлетворять неравенствам:***

$$H_L \le L_{cp} < H_L + \frac{1}{L}$$

- Код для каждого блока строится с использованием методов побуквенного кодирования для алфавита блоков и вероятностного распределения блоков
- При этом избыточность кодовых символов распределяется между всеми буквами блока.

aababbababbb

Блоки	Вероятности	Коды	$p_i l_i$	средняя длина кода
a	0.25	0	0.25	1
b	0.75	1	0.75	

H≈0.8225

Избыточность 1-0.8225 ≈ 0.17

aa ba bb ab ab bb

Блоки	Вероятности	Коды	$p_i l_i$	средняя длина кода
aa	$(0.25)^2$	000	0.1875	0.84375
ab	$0.25 \cdot 0.75$	001	0.5625	r=0.021
ba	$0.75 \cdot 0.25$	01	0.375	
bb	$(0.75)^2$	1	0.5625	

aab abb aba bbb

Блоки	Вероятности	Коды	$p_i l_i$	средняя длина кода
aaa	$(0.25)^3$	00110	0.078125	0.825208
aab	$(0.25)^2 \cdot 0.75$	00111	0.234375	r=0.0027
aba	$(0.25)^2 \cdot 0.75$	00101	0.234375	
abb	$(0.75)^2 \cdot 0.25$	000	0.421875	
baa	$(0.25)^2 \cdot 0.75$	00100	0.234375	
bab	$(0.75)^2 \cdot 0.25$	010	0.421875	
bba	$(0.75)^2 \cdot 0.25$	011	0.421875	
bbb	$(0.75)^3$	1	0.421875	