4.5. Оценки стоимости кодирования разделимого кода

Теорема 3 (нижн. гр. стоим-ти кодир-я разд. кода):
Для любого разделимого кода φ источника S: $C(\varphi, S) \geq H(S)$

Д-во: Пусть $l_i = |\varphi(x_i)|$ $i=\overline{1,N}$. $C(\varphi,S) = \sum\limits^{N} p_i l_i = \sum p_i \log 2^{l_i}$
Рассмотрим:
$$H(S) - C(\varphi, S) = \sum p_i \log \frac{1}{p_i} + \sum p_i \log 2^{-l_i} = \sum p_i \log \frac{2^{-l_i}}{p_i}$$
Восп-ся нер-вом: $\ln x \leq x - 1$. Тогда:
$$H(S) - C(\varphi, S) \leq \log e \sum p_i \left(\frac{2^{-l_i}}{p_i} - 1\right) = \log e \left(\left(\sum 2^{-l_i}\right) - \left(\sum p_i\right)\right) = \log e \left(\sum 2^{-l_i} - 1\right) \underset{\text{по т. 2}}{\leq} 0$$

Зам $C(\varphi, S) = H(S) \iff p_i = 2^{-l_i}$ $\forall i = \overline{1,N}$

Опр Оптимальное кодир-е — то, чья стоимость наименьшая

Теорема 4 (верхняя граница стоим-ти оптимал. кодир-я разд. кода):
$\exists$ разд. код такой, что: $C(\varphi, S) < H(S) + 1$

Д-во: Пусть s_i – наим. целое число такое, что: $\log \frac{1}{p_i} \leq s_i < \log \frac{1}{p_i} + 1$
Имеем: $\sum\limits^{N} 2^{-s_i} \leq \sum 2^{-\log \frac{1}{p_i}} = \sum p_i = 1 \Rightarrow$ числа s_i м/б длинами код. слов для разд. код. S (по т. 2)

Оценим стоимость такого кодир-я:
$$C(\varphi, S) = \sum p_i s_i < \sum p_i \left(\log \frac{1}{p_i} + 1\right) = H(S) + 1$$

Зам Кодир-е в Т.4 — это кодирование Шеннона

Опр Код. Шеннона:
Занумеруем буквы $x_1, \dots, x_N$ так, что $p_1 \geq p_2 \geq \dots \geq p_N$
Опр-м σ_i: $\sigma_1 = 0$, $\sigma_{i+1} = \sigma_i + p_i$
Положим $\varphi(x_i)$ равным первым $\lceil \log \frac{1}{p_i} \rceil$ двоичн. знакам числа σ_i

Ex $A = \{a, b, c, d, e, f, g\}$ $P = \{0.5, 0.2, 0.1, 0.07, 0.05, 0.05, 0.03\}$
$\Rightarrow$ $\sigma = \{0, 0.5, 0.7, 0.8, 0.87, 0.92, 0.97\}$
Ищем длины код. слов:

$2^{-1} \leq 0.5 < 1$ (1) $\leftarrow \lceil \log \frac{1}{p_i} \rceil$
$2^{-3} \leq 0.2 < 2^{-2}$ (3)
$2^{-4} \leq 0.1 < 2^{-3}$ (4)
$2^{-4} \leq 0.07 < 2^{-3}$ (4)
$2^{-5} \leq 0.05 < 2^{-4}$ (5)
$2^{-5} \leq 0.05 < 2^{-4}$ (5)
$2^{-6} \leq 0.03 < 2^{-5}$ (6)

Вспомогат. табличка:
$2^{-1} = 0.5$
$2^{-2} = 0.25$
$2^{-3} = 0.125$
$2^{-4} = 0.0625$
$2^{-5} = 0.03125$
$2^{-6} = 0.015625$

тут можно ещё и сократить коды

0.0	0.5	0.7	и т.д.
$\boxed{0}$ 0	1 0	1 4	
0 0	0 0	0 8	
	0 0	1 6	
		1 2	

Упр Код Шеннона префиксный.

4.6. Энтропийное кодирование

Из 4.5 знаем, что $H(S) \leq C(\hat{A}, S) < H(S) + 1$

$x_i \to \dots$ алфавитное код-е

$(x_{i_1}, \dots, x_{i_N}) \to \dots$ блочное код-е (на самом деле алф-е, но с большим алфавитом)

Утв S' — источник $A' = \{(x_{i_1}, \dots, x_{i_N})\}$ $A = \{x_i\}$

$H(S)$ $H(S') = n \cdot H(S)$

Тогда из Т3,4 для S' $\exists$ код: $n \cdot H(S) \leq C(\hat{A}, S') < n \cdot H(S) + 1$

$$H(S) \leq \frac{C(\hat{A}, S')}{n} < H(S) + \frac{1}{n} \xrightarrow[n \to \infty]{} H(S)$$

скорость на букву

Ура блочному кодированию! Солнышко стремится к $H(S)$! (лучиками)

4.7. Код Хаффмана

Опр Занумеруем буквы так, что $p_1 \geq p_2 \geq \dots \geq p_N$

Склеиваем буквы с наим. вероятн-ми:

$$S \longrightarrow S'$$

$$p_1 \dots p_N \qquad \underset{\overset{\shortparallel}{p_1}}{\tilde{p}_1}, \dots, \underset{\overset{\shortparallel}{p_{N-2}}}{\tilde{p}_{N-2}}, \quad \tilde{p}_{N-1} = p_{N-1} + p_N$$

$$A \longrightarrow A'$$

Строим дерево от листьев. Листья — буквы $x_1, \dots, x_N$

Склейка 2х букв соотв-ет такому действию в дереве: x_{N-1}, x_N → $\tilde{x}_{N-1}$

Вновь переупорядочиваем вершины $\tilde{A}$ так, что: $\tilde{p}_1 \geq \tilde{p}_2 \geq \dots \geq \tilde{p}_{N-1}$

и проделываем те же шаги.

За N-1 шаг дерево построено. В листьях — буквы, путь к ним — код.

Зам Код Хаффмана префиксный

Теорема 5

Код Хаффмана оптимален.

Схема д-ва:

исп-ся след. утв-я:

Утв 1 В любом опт. коде код. слово, соотв. наименее вероятному сообщ-ю, имеет наиб. длину

Утв 2 В оптимальн. двоичн. префиксн. коде 2 наименее вероятн. сообщ-я код-ся словами один. длины, одно из ктр заканч. нулём, а другое — единицей

Утв 3 Если оптимален дв. пр. код для $\hat{S}$, то опт. и получ. из него код для S.

5. Сжатие текстов: сжатие без потерь

5.1. Постановка задачи

В п.4 — код-е источника (не знаем, какой текст будет порождаться источником, => готовимся к любому тексту)

В п.5 — код-е текста (имеем дело с конкр. текстом и не важно, каким источником был порождён)

S. $A = \{a, b\}$ $\quad D = \left\{\frac{253}{256}, \frac{3}{256}\right\}$ $\quad H(S) \approx 0.091976$

Рассмотрим текст из алфавита источника длины 256:
$256 \cdot H(S) \approx 24$ бита ← при исп. блочн. код-я
А если опт. код Хаффмана, то 256 битов!

Если текст $\underbrace{ab}_{1}\underbrace{ab}_{2} \ldots \underbrace{ab}_{128}$ $\quad$ 100...0, 01 $\quad$ т.е. всего 10 битов!

Колмогоровская сложность

? Какова max степень сжатия сообщ-я, не завися от источника, с кот ктр оно порождалось, и от выч. мощностей входных данных

Абсолютный код сообщения

Опр Колмогоровская сложность — наим. длина программы, порожд. сообщ-е

Утв Колм. сложн-ть алг. невычислима

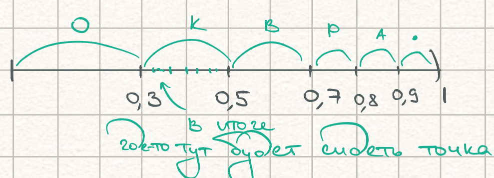
5.2. Арифметическое кодирование

Текст кодируем точкой на интервале $[0, 1)$

Ex КОВ.КОРОВА

Посчитали:

О	3	0.3
К	2	0.2
В	2	0.2
Р	1	0.1
А	1	0.1
.	1	0.1



Выберем точку из середины посл. интервала. В двоичн. предст. (за запятой) мы берём ровно столько знаков, чтобы соотв. коду число достоверно попадало в этот интервал

т.е. делаем, пока есть однозначность. если никто не влез, пора заканчивать

Опр $k(\ell)$ — миним. число такое, что $\ell > \frac{1}{2^{k(\ell)-1}}$, где $\ell = p_{i_1} \cdot \ldots \cdot p_{i_N}$

↑ длина арифм. кода сообщ-я → $0.\boxed{1011}$