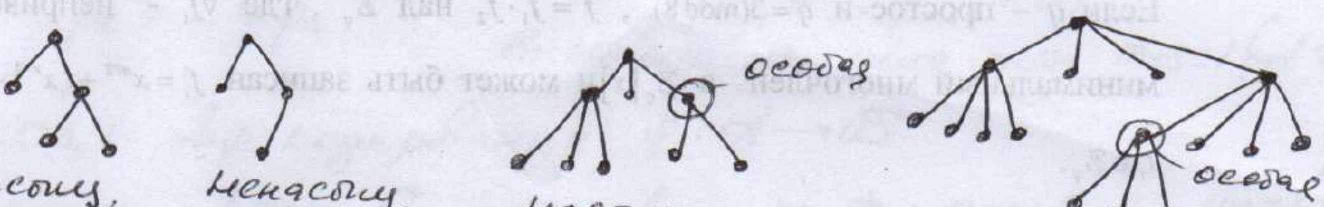
Построение оптимальных кодов по алг. Фано и Хаффмана.

ОПР. Говорят, что корневое дерево (D-ичн.) G явл. насыщенным, если $\forall$ неконцевая вершина имеет ровно D потомков, за искл., м/б, одной вершины на предпоследнем уровне, наз. особой вершиной, им. D_0 потомков $2 \le D_0 < D$.

Замеч.: При $D=2$ в насыщ. дереве нет особых вершин.

Пример:



насыщ. $D=2$ ненасыщ. $D=2$ насыщ. $D=3$ насыщ. $D=4$

ОПР. Префиксное алф. кодир. φ наз. приведенным, если соответств. код. дерево G насыщено, и кроме того,

1) если в дереве G на предпослед. уровне есть особая верш., то ее D_0 потомкам соотв. символы алф. A с D_0 наименьшими вер. из набора $p_1, \ldots, p_m$

2) если в дереве G нет особой вершины, то имеется неконцевая верш. на предпоследнем уровне, D потомкам кот. соответств. символы алф. A с D наименьш. вер. из набора $p_1, \ldots, p_m$.

Редукция привед. кодир-я.

Пусть $m > D$, $\varphi: A \to B^*$ – приведенное D-ичное префиксное алф. кодир-е, $p_1 \ge p_2 \ge \ldots \ge p_m$. Предположим, что код. слова $\underbrace{\varphi(a_{m-s+1}), \ldots, \varphi(a_m)}$ для s симв. $a_{m-s+1}, \ldots, a_m$ с наименьш. вер. $p_{m-s+1}, \ldots, p_m$

им. одинак. длину L,
общий префикс $b_{j_1} \ldots b_{j_{L-1}}$
и различаются лишь последним символом.

Рассм. алфавит $A^{(1)} = \{a_1 \ldots a_{m-s}, \sigma\} = \left(A \setminus \{a_{m-s+1}, \ldots, a_m\}\right) \cup \{\overset{a_{m-s+1} \vee \ldots \vee a_m}{\overset{\shortparallel}{\sigma}}\}$

с распред. вероятн. $P^{(1)} = (p_1, \ldots, p_{m-s}, \gamma)$, $\gamma = p_{m-s+1} + \ldots + p_m$

Зададим кодирование соотнош.: $\varphi_1 : A^{(1)} \to B^* : \varphi_1(a_i) = \varphi(a_i)$
$\varphi_1(\sigma) = b_{j_1} \ldots b_{j_{L-1}}$

φ_1 наз. редуцированным по отнош. к σ.

Теорема: Для любых $m = \#\mathcal{A}$, $\mathcal{D} = \#\mathcal{B} \geq 2$ алгоритм Хаффмана строит оптимальное приведенное $\mathcal{D}$-ичное префиксное алфавитное кодирование.
(без д-ва).

Алгоритм Хаффмана ($\mathcal{D} \geq 2$)

Вход:
$\mathcal{A} = \{a_1, \dots, a_m\}$, $m \equiv 1 \bmod (\mathcal{D}-1)$
(если не выполняется, добавить симв. $a_{m+1}, \dots, a_{m+s}$ с вероятн. $p_{m+1} = \dots = p_{m+s} = 0$)
s - наименьшее целое, чтобы $m+s \equiv 1 \pmod{\mathcal{D}-1}$

$p = (p_1, \dots, p_m)$

$p_1 \geq p_2 \geq \dots \geq p_m$.

Выход: Алф. кодирование $\varphi : \mathcal{A} \to \mathcal{B}^*$.

Алгоритм: I шаг. Слияние. Если $m = \mathcal{D} \Rightarrow$ перейти к шагу II.

Если $m > \mathcal{D}$:

Выбрать $\mathcal{D}$ симв. с наименьшей вер.

Постр. слияние $\mathcal{A} \to \mathcal{A}^{(1)} = \{a_1, \dots, a_{m-\mathcal{D}}, \sigma = a_{m-\mathcal{D}+1} \cup \dots \cup a_m\}$

$p^{(1)} = (p_1, \dots, p_{m-\mathcal{D}}, \sigma = p_{m-\mathcal{D}+1} + \dots + p_m)$

Упоряд. символы $\mathcal{A}^{(1)}$ по невозраст. вер. $p^{(1)}$, вернуться на шаг I с алфавитом $(\mathcal{A}^{(1)}, p^{(1)})$

II шаг. Имеем: $\mathcal{A} = \mathcal{A}^{(0)} \to \mathcal{A}^{(1)} \to \dots \to \mathcal{A}^{(k-1)}$

$|\mathcal{A}^{(k-1)}| = \mathcal{D}$

$\exists\, \varphi_{k-1} : \mathcal{A}^{(k-1)} \to \mathcal{B}$ - произвольное вз. однозначное отображение

$\exists$ заданы кодирования

$\varphi_{k-1}, \varphi_{k-2}, \dots, \varphi_{i+1}$. Зададим $\varphi_i : \mathcal{A}^{(i)} \to \mathcal{B}^*$

$\divideontimes$ слияние $\mathcal{A}^{(i)} \to \mathcal{A}^{(i+1)}$

$a_1', \dots a_s' \mapsto \sigma'$

Было: $\varphi_{i+1}(\sigma') = w \in \mathcal{B}^*$

$\Rightarrow \varphi_i(a_j') = w b_j$, $1 \leq j \leq s$.

…чание. Очевидно, что на i-м шаге количество букв уменьшается на $\mathcal{D}-1$ (слияние $\mathcal{D}$ букв в одну), а на последнем шаге должна остаться одна буква. Следовательно, $(\mathcal{D}-1) \mid (m-1)$ – должно выполняться. Иначе построение код. дерева будет не корректно.

Однако буквы с вер. 0 можно удалить из получ. дерева, тогда их родительская вершина станет особой в насыщенном дереве.

Пример.

$\mathcal{A}=\{a, b, c, d, e, f, g, h\}$ $\quad \mathcal{B}=\{0,1,2\}$

$p=(0{,}2;\ 0{,}14;\ 0{,}13;\ 0{,}13;\ 0{,}12;\ 0{,}12;\ 0{,}08;\ 0{,}08)$. $\quad \mathcal{D}=\#\mathcal{B}=3.$

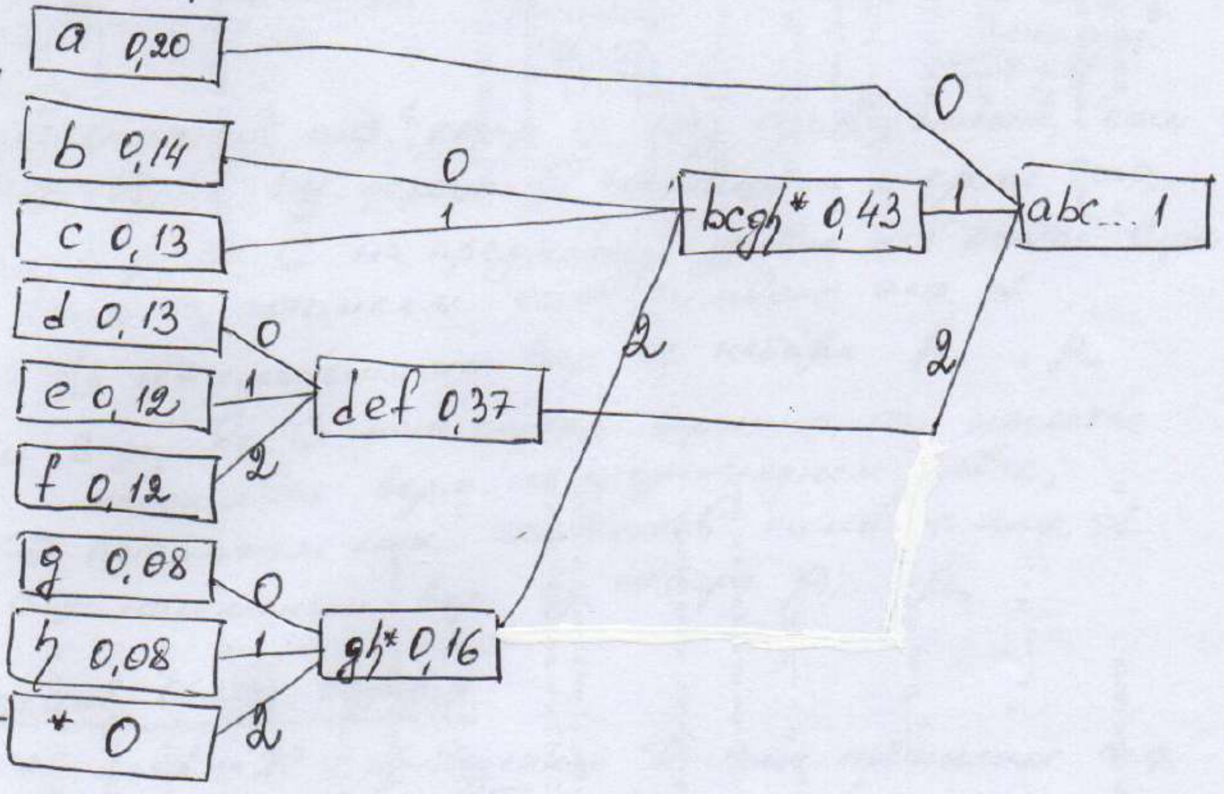
$m=8$

$8 \not\equiv 1 \bmod \mathcal{D}-1$

$s=(1-m) \bmod \mathcal{D}-1 =$

$= -7 \bmod 2 =$

$= 1.$



$\varphi(a)=0$

$\varphi(b)=10$

$\varphi(c)=11$

$\varphi(d)=20$

$\varphi(e)=21$

$\varphi(f)=22$

$\varphi(g)=120$

$\varphi(h)=121$

$$l^{\varphi}=0{,}2\cdot 1+(0{,}14+0{,}13+0{,}13+0{,}12+0{,}12)\cdot 2+(0{,}08+0{,}08)\cdot 3=1{,}96$$

– средняя длина оптимального кода.