Лекция 14. Алфавитные коды. Оптимальные коды и их свойства. Метод Хаффмана построения оптимального кода.

Лектор — Селезнева Светлана Николаевна
selezn@cs.msu.ru

факультет ВМК МГУ имени М.В. Ломоносова

Лекции на сайте https://mk.cs.msu.ru

Длина кода сообщения

Пусть $A = \{a_1, \ldots, a_r\}$ — исходный алфавит и $C_\varphi = \{B_1, \ldots, B_r\}$ — разделимый алфавитный код в кодирующем алфавите B, причем $|B_i| = l_i$, $i = 1, \ldots, r$.

Пусть $\alpha \in A^*$ — сообщение, $|\alpha| = m$, и в слове α буква a_i встречается m_i раз, $i = 1, \ldots, r$.

Отметим, что $\sum_{i=1}^{r} m_i = m$.

Найдем длину кода сообщения α при кодировании φ:

$$|\varphi(\alpha)| = \sum_{i=1}^{r} m_i l_i.$$

Отношение длин кода и сообщения

Посмотрим, как соотносится длина кода сообщения с длиной сообщения:

$$\frac{|\varphi(\alpha)|}{|\alpha|} = \frac{\sum_{i=1}^{r} m_i l_i}{m} = \sum_{i=1}^{r} \frac{m_i}{m} \cdot l_i.$$

Положим $p_i = \frac{m_i}{m}$ — частота появления буквы a_i в слове α, $i = 1, \ldots, r$. Тогда:

$$\frac{|\varphi(\alpha)|}{|\alpha|} = \sum_{i=1}^{r} p_i l_i.$$

Отношение длин кода и сообщения

Итак,

$$\frac{|\varphi(\alpha)|}{|\alpha|} = \sum_{i=1}^{r} p_i l_i.$$

Допустим, мы хотим уменьшить это отношение насколько это возможно при условии, что код обязан остаться разделимым.

Значит, нужно подобрать однозначный алфавитный код, для которого сумма в правой части минимальна.

Задача об оптимальном кодировании

Рассмотрим следующую задачу.

Заданы исходный и кодирующий алфавиты A и B.

Предположим, что частоты букв исходного алфавита известны.

При каком однозначном кодировании φ из A в B отношение длины кода к длине сообщения будет наименьшим?

Если найдется однозначное кодирование φ^* из A в B, при котором достигается эта наименьшая граница отношения, то его назовем оптимальным.

Стоимость кода

Пусть $A = \{a_1, \ldots, a_r\}$ — исходный алфавит и B — кодирующий алфавит.

Пусть $P = (p_1, \ldots, p_r)$ — набор частот появления букв исходного алфавита, где

1) $p_i \in \mathbb{R}_+$,
2) $p_i > 0$,
3) $\sum\limits_{i=1}^{r} p_i = 1$.

Пусть $C_\varphi = \{B_1, \ldots, B_r\}$ — алфавитный код, $|B_i| = l_i$ для всех $i = 1, \ldots, r$.

Стоимостью (или **избыточностью**) кода C_φ назовем величину

$$c(\varphi) = \sum_{i=1}^{r} p_i l_i.$$

Оптимальный код

Однозначный код C_{φ^*} назовем **оптимальным** (или кодом **с минимальной избыточностью**) (при заданных A, B, P), если

$$c(\varphi^*) = \inf_{\varphi} c(\varphi),$$

где инфимум берется по всем однозначным алфавитным кодам.

Существование оптимального кода

Предложение 14.1. *При любых заданных* A, $|A| = r$, B, $|B| = q$, *и* $P = (p_1, \dots, p_r)$ *найдется оптимальный код* C_{φ^*}.

Доказательство. Рассмотрим равномерный (а значит, однозначный) алфавитный код $C_{\varphi'} = \{B'_1, \dots, B'_r\}$, где $|B'_i| = \lceil \log_q r \rceil$ для всех $i = 1, \dots, r$.

Его стоимость равна: $c(\varphi') = \lceil \log_q r \rceil$.

При поиске оптимального кода можно рассматривать только однозначные коды с не большей стоимостью.

Существование оптимального кода

Доказательство. Пусть $C_\varphi = \{B_1, \ldots, B_r\}$ — однозначный алфавитный код с не большей стоимостью, где $|B_i| = l_i$ для всех $i = 1, \ldots, r$.

Значит,

$$\sum_{i=1}^{r} p_i l_i \leqslant \lceil \log_q r \rceil,$$

откуда

$$p_i l_i \leqslant \lceil \log_q r \rceil, \text{ и } l_i \leqslant \frac{\lceil \log_q r \rceil}{p_i}$$

для всех $i = 1, \ldots, r$.

Существование оптимального кода

Доказательство. Итак,

$$l_i \leqslant \frac{\lceil \log_q r \rceil}{p_i}$$

для всех $i = 1, \ldots, r$.

Но найдется только конечное число алфавитных кодов с такими ограничениями длин кодовых слов.

Следовательно, в определениии оптимального кода инфимум берется по конечному множеству.

А значит, всегда найдется какой-то элемент этого множества, на котором этот инфимум достигается.

$\square$

Оптимальный код

По предложению 14.1 можно уточнить определение оптимального кода.

Однозначный код C_{φ^*} называется **оптимальным** (или кодом **с минимальной избыточностью**) (при заданных A, B, P), если

$$c(\varphi^*) = \min_{\varphi} c(\varphi),$$

где минимум берется по всем однозначным алфавитным кодам.

Существование префиксного оптимального кода

Предложение 14.2. *При любых заданных A, B и P найдется* **префиксный** *оптимальный код* C_{φ^*}.

Доказательство. По теореме 11.3 для любого однозначного кода найдется *префиксный* код с теми же длинами кодовых слов.

А значит, для оптимального кода найдется префиксный код с теми же длинами кодовых слов.

Из того, что в определении стоимости кода участвуют только длины кодовых слов, получаем, что этот префиксный код также является оптимальным.

□

Префиксные оптимальные коды

Из предложения 14.2 следует, что при поиске оптимального кода можно ограничиться только префиксными кодами.

Как найти оптимальный код, если известны A, B и P?

Сначала докажем некоторые свойства оптимальных кодов.

Частым буквам — более короткие слова

Лемма 14.1. *Пусть заданы* A, $|A| = r$, B *и* $P = (p_1, \dots, p_r)$, *причем* $p_i > p_j$. *Если* $C_\varphi = \{B_1, \dots, B_r\}$ *— оптимальный код, то* $|B_i| \leqslant |B_j|$.

Доказательство. Пусть $l_1, \dots, l_r$ — длины кодовых слов $B_1, \dots, B_r$ и, для определенности, $i < j$.

Докажем от обратного: предположим, что $l_i > l_j$.

Рассмотрим код $C_{\varphi'}$, где

$$C_{\varphi'} = \{B_1, \dots, B_{i-1}, B_j, B_{i+1}, \dots, B_{j-1}, B_i, B_{j+1}, \dots, B_r\}.$$

Код $C_{\varphi'}$ получен из однозначного кода C_φ перестановкой кодовых слов B_i и B_j. Значит, код $C_{\varphi'}$ — также однозначен.

Частым буквам — более короткие слова

Доказательство. Получаем:

$$\begin{aligned} c(\varphi') - c(\varphi) &= p_i(l_j - l_i) + p_j(l_i - l_j) = \\ &= (p_i - p_j)(l_j - l_i) < 0, \end{aligned}$$

т. к. $p_i > p_j$ и $l_i > l_j$.

Значит, $c(\varphi') < c(\varphi)$, чего не может быть, т. к. C_φ — оптимальный код.

Следовательно, $l_i \leqslant l_j$.

$\square$

Слова с наибольшей длиной

Лемма 14.2. *Пусть заданы* A, $|A| = r$, $r \geqslant 2$, $B = \{0, 1\}$ *и* $P = (p_1, \ldots, p_r)$. *Если* $C_\varphi = \{B_1, \ldots, B_r\}$ *— оптимальный* **префиксный** *код и* B_i *— кодовое слово с наибольшей длиной, причем* $B_i = B_i' b$, *где* $B_i' \in B^*$, $b \in B$, *то в коде* C_φ *найдется кодовое слово* $B_j = B_i' \bar{b}$.

Слова с наибольшей длиной

Доказательство. Если $r = 2$, то лемма верна. Пусть $r \geqslant 3$ и $l_1, \ldots, l_r$ — длины кодовых слов $B_1, \ldots, B_r$. Отметим, что $l_i \geqslant 2$.

Докажем от обратного: предположим, что слово $B_i'\bar{b}$ в коде C_φ не встречается.

Рассмотрим код $C_{\varphi'}$, где

$$C_{\varphi'} = \{B_1, \ldots, B_{i-1}, B_i', B_{i+1}, \ldots, B_r\}.$$

Код $C_{\varphi'}$ получен из префиксного кода C_φ удалением последней буквы из самого длинного кодового слова B_i. Значит, код $C_{\varphi'}$ также является префиксным.

Слова с наибольшей длиной

Доказательство. Получаем:

$$c(\varphi') - c(\varphi) \quad = \quad p_i(l_i - 1) - p_i l_i = -p_i < 0,$$

т. к. $p_i > 0$.

Значит, $c(\varphi') < c(\varphi)$, чего не может быть, т. к. C_φ — оптимальный код.

Следовательно, в коде C_φ найдется кодовое слово $B_i'\bar{b}$.

□

Две наименьшие частоты

Лемма 14.3. *Пусть заданы* A, $|A| = r$, $r \geqslant 2$, $B = \{0, 1\}$ *и* $P = (p_1, \ldots, p_r)$, *причем*

$$p_1 \geqslant p_2 \geqslant \ldots \geqslant p_{r-1} \geqslant p_r.$$

Тогда найдется такой оптимальный префиксный код, что кодовые слова, сопоставляемые буквам с частотами p_{r-1} *и* p_r, *являются самыми длинными и отличаются только последней буквой.*

Две наименьшие частоты

Доказательство. Пусть $C_\varphi = \{B_1, \dots, B_r\}$ — какой-то оптимальный префиксный код и $l_1, \dots, l_r$ — длины кодовых слов $B_1, \dots, B_r$.

Пусть B_i — кодовое слово с наибольшей длиной в коде C_φ, $B_i = B_i' b$, где $B_i' \in B^*$, $b \in B$.

По лемме 14.2 в коде C_φ найдется кодовое слово $B_j = B_i' \bar{b}$. Пусть, для определенности, $i < j$.

Рассмотрим код $C_{\varphi'}$, где

$$C_{\varphi'} = \{B_1, \dots, B_{i-1}, B_{r-1}, B_{i+1}, \dots, B_{j-1}, B_r, B_{j+1}, \dots, B_{r-2}, B_i, B_j\}.$$

Код $C_{\varphi'}$ получен из префиксного кода C_φ перестановкой кодовых слов. Значит, код $C_{\varphi'}$ также является префиксным.

Две наименьшие частоты

Доказательство. Получаем:

$$\begin{aligned} c(\varphi') - c(\varphi) &= p_i(l_{r-1} - l_i) + p_j(l_r - l_j) + \\ &\quad + p_{r-1}(l_i - l_{r-1}) + p_r(l_j - l_r) = \\ &= (p_i - p_{r-1})(l_{r-1} - l_i) + (p_j - p_r)(l_r - l_j). \end{aligned}$$

Теперь если $p_i = p_{r-1}$, то $(p_i - p_{r-1})(l_{r-1} - l_i) = 0$.

Если же $p_i > p_{r-1}$, то по лемме 14.1 верно $l_i \leqslant l_{r-1}$. Но l_i — наибольшая длина среди кодовых слов, поэтому $l_i = l_{r-1}$, откуда $(p_i - p_{r-1})(l_{r-1} - l_i) = 0$

Аналогично устанавливаем, что $(p_j - p_r)(l_r - l_j) = 0$.

Значит, $c(\varphi') = c(\varphi)$. Но C_φ — оптимальный код, поэтому $C_{\varphi'}$ — также оптимальный код.

Код $C_{\varphi'}$ — искомый. $\square$

Поиск оптимального кода

Если заданы A, B и P, то как найти оптимальный код?

Мы покажем (см. теорему редукции), что задачу поиска оптимального кода можно свести к такой же задаче, но для исходного алфавита с меньшим числом букв.

Два префиксных кода

Лемма 14.4. *Пусть* $B = \{0, 1\}$ *— кодирующий алфавит, заданы два исходных алфавита* A, A' *и соответствующие наборы частот* P, P' *и алфавитные коды* C_φ, $C_{\varphi'}$*:*

$$\begin{array}{ll} A = \{a_1, \ldots, a_{r-1}, a_r\}, & A' = \{a_1, \ldots, a_{r-1}, a', a''\}, \\ P = (p_1, \ldots, p_{r-1}, p_r), & P' = (p_1, \ldots, p_{r-1}, p', p''), \\ C_\varphi = \{B_1, \ldots, B_{r-1}, B_r\}, & C_{\varphi'} = \{B_1, \ldots, B_{r-1}, B_r0, B_r1\}, \end{array}$$

где $r \geqslant 2$. *Тогда если один из этих кодов префиксный, то и другой префиксный, причем*

$$c(\varphi') = c(\varphi) + p_r.$$

Доказательство проведите самостоятельно.

Теорема редукции

Теорема 14.4 (редукции). *Пусть $B = \{0, 1\}$ — кодирующий алфавит, заданы два исходных алфавита A, A' и соответствующие наборы частот P, P' и алфавитные коды C_φ, $C_{\varphi'}$:*

$$
\begin{array}{ll}
A = \{a_1, \ldots, a_{r-1}, a_r\}, & A' = \{a_1, \ldots, a_{r-1}, a', a''\}, \\
P = (p_1, \ldots, p_{r-1}, p_r), & P' = (p_1, \ldots, p_{r-1}, p', p''), \\
C_\varphi = \{B_1, \ldots, B_{r-1}, B_r\}, & C_{\varphi'} = \{B_1, \ldots, B_{r-1}, B_r0, B_r1\},
\end{array}
$$

где $r \geqslant 2$. Тогда:

1) *если $C_{\varphi'}$ — оптимальный префиксный код, то и C_φ — оптимальный префиксный код;*

2) *если C_φ — оптимальный префиксный код и*

$$p_1 \geqslant p_2 \geqslant \ldots \geqslant p_{r-1} \geqslant p' \geqslant p'',$$

то и $C_{\varphi'}$ — оптимальный префиксный код.

Теорема редукции

Доказательство. По лемме 14.4 если один из кодов φ, φ' префиксный, то и другой префиксный, причем

$$c(\varphi') = c(\varphi) + p_r.$$

Теорема редукции

Доказательство. 1. Пусть $C_{\varphi'}$ — оптимальный префиксный код. Предположим, что код C_{φ} не является оптимальным.

Значит, найдется оптимальный префиксный код $C_{\varphi_1} = \{D_1, \ldots, D_{r-1}, D_r\}$. Отметим, что $c(\varphi_1) < c(\varphi)$.

Рассмотрим префиксный код $C_{\varphi_1'} = \{D_1, \ldots, D_{r-1}, D_r 0, D_r 1\}$.

Получаем:

$$\begin{aligned} c(\varphi_1') - c(\varphi') &= (c(\varphi_1) + p_r) - (c(\varphi) + p_r) = \\ &= c(\varphi_1) - c(\varphi) < 0. \end{aligned}$$

Значит, $c(\varphi_1') < c(\varphi')$, чего не может быть, т. к. $C_{\varphi'}$ — оптимальный код.

Следовательно, код C_{φ} — оптимальный.

Теорема редукции

Доказательство. 2. Пусть теперь C_φ — оптимальный префиксный код и $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_{r-1} \geqslant p' \geqslant p''$. Предположим, что код $C_{\varphi'}$ не является оптимальным.

Значит, по лемме 14.3 найдется оптимальный префиксный код $C_{\varphi'_1}$, имеющий вид $\{D_1, \ldots, D_{r-1}, D_r0, D_r1\}$. Отметим, что $c(\varphi'_1) < c(\varphi')$.

Рассмотрим префиксный код $C_{\varphi_1} = \{D_1, \ldots, D_{r-1}, D_r\}$.

Получаем:

$$\begin{aligned} c(\varphi_1) - c(\varphi) &= (c(\varphi'_1) - p_r) - (c(\varphi') - p_r) = \\ &= c(\varphi'_1) - c(\varphi') < 0. \end{aligned}$$

Значит, $c(\varphi_1) < c(\varphi)$, чего не может быть, т. к. C_φ — оптимальный код.

Следовательно, код $C_{\varphi'}$ — оптимальный.

Метод Хаффмана построения оптимального кода

По теореме редукции задачу поиска оптимального кода можно свести к такой же задаче, но с исходным алфавитом с числом букв, меньшим на единицу, и с набором частот, получающимся из первоначального сложением двух наименьших частот.

Так можно уменьшать число букв в исходных алфавитах до тех пор, пока не получим алфавит из двух букв.

А для исходного алфавита из двух букв при любом наборе частот в кодирующем алфавите $B = \{0, 1\}$ оптимальным является код $C_\varphi = \{0, 1\}$.

Построение оптимального кода

Алгоритм построения оптимального кода в кодирующем алфавите $B = \{0, 1\}$.

Вход: набор частот $P = (p_1, \ldots, p_r)$, $p_i \in \mathbb{R}_+$, $p_i > 0$ для всех $i = 1, \ldots, r$, $\sum\limits_{i=1}^{r} p_i = 1$, $r \geqslant 2$.

Выход: дерево D_{φ^*} какого-то оптимального префиксного кода $C_{\varphi^*} = \{B_1, \ldots, B_r\}$ для набора частот P.

Построение оптимального кода

Описание алгоритма.

1. Положить: $H_1 = (V_1, E_1)$, где $V_1 = \{u_1, \ldots, u_r\}$, $E_1 = \emptyset$, и $p(u_i) = p_i$ для всех $i = 1, \ldots, r$, $W_1 = V_1$.

2. Цикл: для всех $k = 1, \ldots, r-1$ повторить:

выбрать в множестве W_k две такие вершины w' и w'', что

$$p(w') \leqslant p(w), \;\; p(w'') \leqslant p(w)$$

для любой вершины $w \in W_k$, $w \neq w'$, $w \neq w''$, положить:

$$H_{k+1} = (V_{k+1}, E_{k+1}),$$

где $V_{k+1} = V_k \cup \{v_k\}$, $E_{k+1} = E_k \cup \{(v_k, w'), (v_k, w'')\}$, и

$$p(v_k) = p(w') + p(w''), \;\; W_{k+1} = (W_k \cup \{v_k\}) \setminus \{w', w''\},$$

ребру (v_k, w') приписать 0, ребру (v_k, w'') приписать 1.

3. Положить: $D_{\varphi^*} = H_r$ с корнем v_{r-1}.

Окончание описания алгоритма.

Построение оптимального кода

Пример. Построим оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$:

Построение оптимального кода

Пример. Построим оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$:

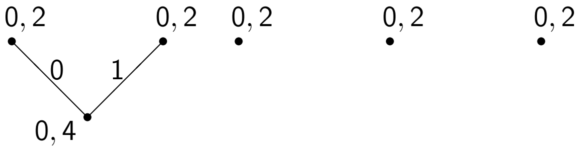
Построение оптимального кода

Пример. Построим оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$:

$0,2 \quad 0,2 \quad 0,2 \quad 0,2 \quad 0,2$

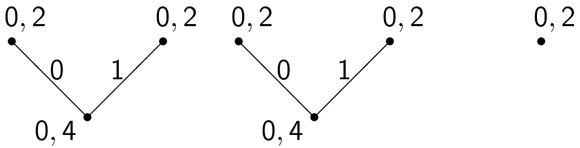
Построение оптимального кода

Пример. Построим оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$:



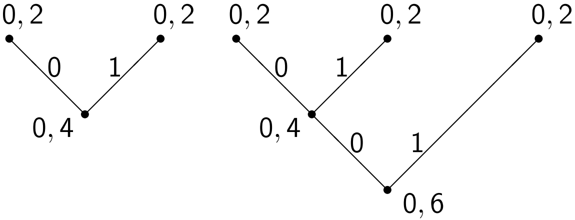
Построение оптимального кода

Пример. Построим оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$:



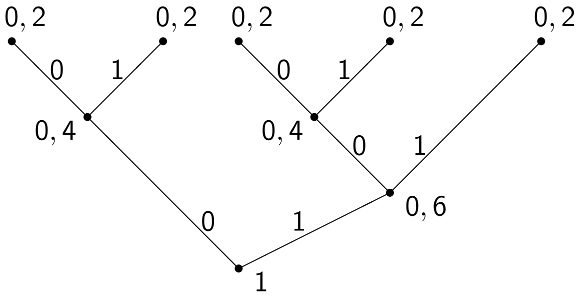
Построение оптимального кода

Пример. Построим оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$:



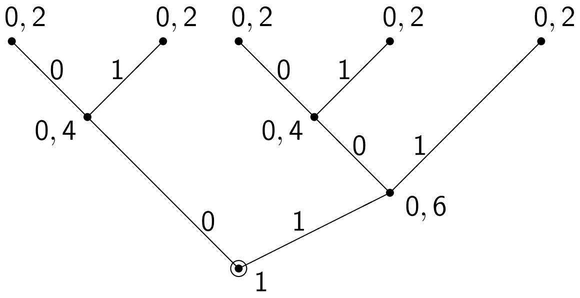
Построение оптимального кода

Пример. Построим оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$:



Построение оптимального кода

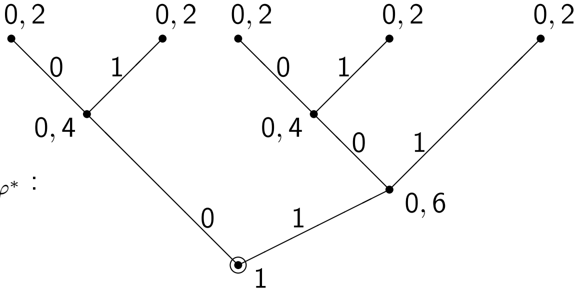
Пример. Построим оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$:



Построение оптимального кода

Пример. Построим оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$:

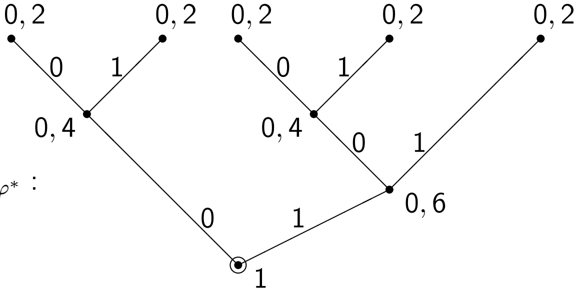
D_{φ^*} :



Построение оптимального кода

Пример. Построим оптимальный префиксный код в кодирующем алфавите $B = \{0,1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$:

D_{φ^*} :

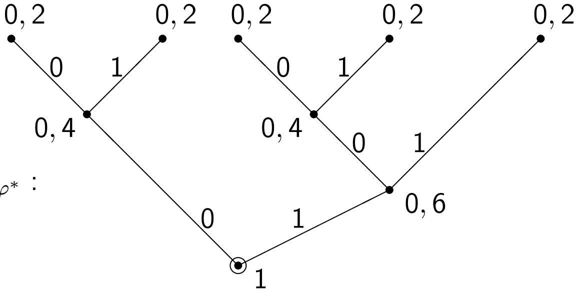


Получаем: $C_{\varphi^*} = \{00, 01, 100, 101, 11\}$.

Построение оптимального кода

Пример. Построим оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$:

D_{φ^*} :



Получаем: $C_{\varphi^*} = \{00, 01, 100, 101, 11\}$. Кроме того,

$$c(\varphi^*) = 3 \cdot 2 \cdot 0,2 + 2 \cdot 3 \cdot 0,2 = 2,4.$$

Задачи для самостоятельного решения

1. Докажите лемму 14.4.

2^*. Покажите, что в кодирующем алфавите $B = \{0, 1\}$ найдется оптимальный код с набором $(l_1, \dots, l_r)$ длин кодовых слов тогда и только тогда, когда

$$\sum_{i=1}^{r} \frac{1}{2^{l_i}} = 1.$$

Литература к лекции

1. Алексеев В. Б. Лекции по дискретной математике. М.: Инфра-М, 2012. С. 50–55.

2. Яблонский С. В. Введение в дискретную математику. М.: Высшая школа, 2001. С. 276–288.

3. Гаврилов Г. П., Сапоженко А. А. Задачи и упражнения по дискретной математике. М.: Физматлит, 2004. Гл. VIII 2.1, 2.10.