Занятие 9. Алфавитные коды. Однозначность алфавитного кода. Оптимальные коды. Метод Хаффмана построения оптимального кода.

Селезнева Светлана Николаевна
selezn@cs.msu.ru

факультет ВМК МГУ имени М.В. Ломоносова

Страница курса на сайте http://mk.cs.msu.ru

Для разбора домашнего задания

Кодирование

Пусть заданы два конечных алфавита A и B.

Алфавит A назовем **исходным**, алфавит B — **кодирующим**.

Кодированием (из A в B) называется произвольное отображение

$$\varphi : A^* \to B^*.$$

При кодировании φ любое слово $\alpha \in A^*$ называется (**исходным**) **сообщением**, а слово $\beta = \varphi(\alpha) \in B^*$ — его **кодом** (или **закодированным сообщением**).

Можно рассматривать кодирования вида $\varphi : S \to B^*$, где $S \subseteq A^*$ — множество исходных сообщений.

Код

Если $\varphi : A^* \to B^*$ — кодирование, то множество кодов всех слов из A^* назовем **кодом** C_φ, т. е.

$$C_\varphi = \{\varphi(\alpha) \mid \alpha \in A^*\} \subseteq B^*.$$

Т. е. код C_φ — множество кодов всех сообщений.

Разделимость кодирования

Кодирование $\varphi : A^* \to B^*$ называется **однозначным** (или **разделимым**), если для любых слов $\alpha_1, \alpha_2 \in A^*$ из $\alpha_1 \neq \alpha_2$ следует $\varphi(\alpha_1) \neq \varphi(\alpha_2)$.

Т. е. кодирование φ — разделимо, если оно разным сообщениям сопоставляет различные коды.

Другими словами, кодирование φ — однозначно, если любое слово $\beta \in B^*$ является кодом не более одного сообщения.

Алфавитное кодирование

Пусть $A = \{a_1, \ldots, a_r\}$ — исходный алфавит,
$B = \{b_1, \ldots, b_q\}$ — кодирующий алфавит.

Кодирование $\varphi : A^* \to B^*$ называется **алфавитным** (или **побуквенным**), если оно описывается следующей схемой:

1) заданы различные непустые коды букв алфавита A:

$$\begin{aligned} \varphi(a_1) &= B_1,\ B_1 \in B^*, \\ \varphi(a_2) &= B_2,\ B_2 \in B^*, \\ &\ldots, \\ \varphi(a_r) &= B_r,\ B_r \in B^*, \end{aligned}$$

2) слова в алфавите A кодируются побуквенно, т. е. если $\alpha \in A^*$, $\alpha = a_{i_1} a_{i_2} \ldots a_{i_m}$, где $m \geqslant 2$, то

$$\varphi(\alpha) = \varphi(a_{i_1})\varphi(a_{i_2}) \ldots \varphi(a_{i_m}) = B_{i_1} B_{i_2} \ldots B_{i_m}.$$

Алфавитный код

Пусть φ — алфавитное кодирование из A в B, т. е.

$$\varphi(a_1) = B_1,\ \varphi(a_2) = B_2,\ \ldots,\ \varphi(a_r) = B_r.$$

Коды букв алфавита A, т. е. слова $B_1, \ldots, B_r$, называются **кодовыми словами**.

Множество всех кодовых слов при кодировании φ назовем **алфавитным кодом** C_φ, т. е.

$$C_\varphi = \{B_1, \ldots, B_r\}.$$

Отметим, что код C_φ однозначно определяет алфавитное кодирование φ (при заданном порядке букв из A).

Алфавитный код C_φ назовем **однозначным** (или **разделимым**), если кодирование φ — разделимо.

Декодирование

Пусть $C_\varphi = \{B_1, \ldots, B_r\} \subseteq B^*$ — алфавитный код и $\beta \in B^*$.

Декодировать слово β означает разбить его на последовательность кодовых слов (если это возможно), т. е. представить в виде:

$$\beta = B_{i_1} B_{i_2} \ldots B_{i_m},$$

где $B_{i_1}, \ldots, B_{i_m} \in C_\varphi$.

Если код C_φ является разделимым, то для любого слова $\beta \in B^*$ найдется не более одного декодирования.

Равномерный алфавитный код

Алфавитный код $C = \{B_1, \ldots, B_r\} \subseteq B^*$ называется **равномерным**, если длины всех его кодовых слов одинаковы, т. е.

$$|B_1| = |B_2| = \ldots = |B_r|.$$

Предложение. *Любой равномерный алфавитный код является разделимым*.

Префиксный алфавитный код

Алфавитный код $C = \{B_1, \ldots, B_r\} \subseteq B^*$ называется **префиксным**, если никакое его кодовое слово не является префиксом никакого другого его кодового слова, т. е.

$$\nexists\, B_i, B_j \in C : B_i = B_j\beta_2 \text{ для некоторого слова } \beta_2 \in B^*.$$

Предложение. *Любой префиксный алфавитный код является разделимым*.

Суффиксный алфавитный код

Алфавитный код $C = \{B_1, \ldots, B_r\} \subseteq B^*$ называется **суффиксным**, если никакое его кодовое слово не является суффиксом никакого другого его кодового слова, т. е.

$$\nexists\, B_i, B_j \in C : B_i = \beta_1 B_j \text{ для некоторого слова } \beta_1 \in B^*.$$

Предложение. *Любой суффиксный алфавитный код является разделимым.*

Граф алфавитного кода

Пусть $C_\varphi = \{B_1, \ldots, B_r\} \subseteq B^*$ — алфавитный код.

Построим *орграф* $G_\varphi = (V_\varphi, E_\varphi)$ кода C_φ.

1. Множество вершин V_φ, $V_\varphi \subseteq B^*$, состоит из пустого слова Λ и всех тех слов в алфавите B, которые являются собственным префиксом некоторого кодового слова и одновременно собственным суффиксом некоторого кодового слова (другого или, возможно, того же) и не являются никаким кодовым словом, т. е.

$$V_\varphi = \{\beta \in B^* \mid \begin{array}{l} 1)\ \exists B_i \in C_\varphi : B_i = \beta\beta',\ \beta' \neq \Lambda; \\ 2)\ \exists B_j \in C_\varphi : B_j = \beta''\beta,\ \beta'' \neq \Lambda; \\ 3)\ \beta \neq B_k,\ k = 1, \ldots, r\}. \end{array}$$

Граф алфавитного кода

Итак, $C_\varphi = \{B_1, \ldots, B_r\} \subseteq B^*$ — алфавитный код.

2. Опишем множество дуг E_φ: если $\beta', \beta'' \in V_\varphi$, то $(\beta', \beta'') \in E_\varphi$, если найдется такое кодовое слово B_i и такая последовательность D кодовых слов $B_{i_1}, \ldots, B_{i_k}$, что

$$B_i = \beta' B_{i_1} \ldots B_{i_k} \beta'',$$

причем если $\beta' = \beta'' = \Lambda$, то $k \geqslant 2$; если $\beta' \neq \Lambda$ или $\beta'' \neq \Lambda$, то $k \geqslant 1$; если $\beta', \beta'' \neq \Lambda$, то $k \geqslant 0$.

При этом дуге $(\beta', \beta'') \in E_\varphi$ приписываем пометку D, где $D = B_{i_1}, \ldots, B_{i_k}$.

Критерий разделимости алфавитного кода

Теорема. *Алфавитный код C_φ является разделимым тогда и только тогда, когда в графе G_φ отсутствуют ориентированные циклы (в том числе, и петли), проходящие через вершину Λ.*

Проверка разделимости алфавитного кода

Алгоритм проверки разделимости алфавитного кода

Вход: алфавитный код $C = \{B_1, \ldots, B_r\} \subseteq B^*$ в кодирующем алфавите B.

Выход: «да», если код C является разделимым, и «нет» и слово $\beta \in B^*$, допускающее не менее двух декодирований, в обратном случае.

Проверка разделимости алфавитного кода

Описание алгоритма.

1. Построить орграф G кода C.

2. Если граф G не содержит петель или направленных циклов, проходящих через «пустую» вершину, то выдать «да» и остановиться.

3. Иначе, пусть $\beta_0, \beta_1, \ldots, \beta_m, \beta_0$ — направленный цикл в G, где $\beta_i \in B^*$, $i = 1, \ldots, m$, $\beta_0 = \Lambda$, причем дуга (β_{i-1}, β_i) помечена последовательностью D_i, $i = 1, \ldots, m$, а дуга (β_m, β_0) помечена последовательностью D_{m+1}. Тогда выдать «нет» и

$$\beta = D_1\beta_1 D_2\beta_2 \ldots \beta_m D_{m+1} \in B^*$$

и остановиться.

Окончание описания алгоритма.

Гл. 7, 1.2(1)

Гл. 7, 1.2(1). Проверить, является ли алфавитный код $C = \{01, 201, 112, 122, 0112\}$ однозначным.

Гл. 7, 1.2(1)

Гл. 7, 1.2(1). Проверить, является ли алфавитный код $C = \{01, 201, 112, 122, 0112\}$ однозначным.

Решение. Построим граф $G = (V, E)$.

Гл. 7, 1.2(1)

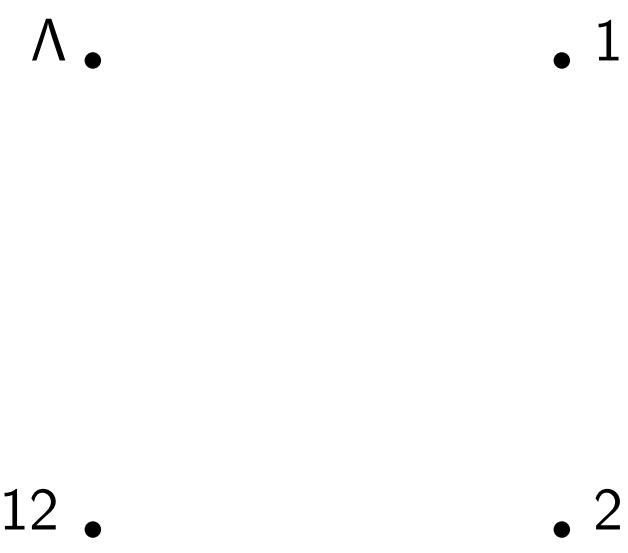
Гл. 7, 1.2(1). Проверить, является ли алфавитный код $C = \{01, 201, 112, 122, 0112\}$ однозначным.

Решение. Построим граф $G = (V, E)$. Получаем: $V = \{\Lambda, 1, 2, 12\}$.

Гл. 7, 1.2(1)

Гл. 7, 1.2(1). Проверить, является ли алфавитный код $C = \{01, 201, 112, 122, 0112\}$ однозначным.

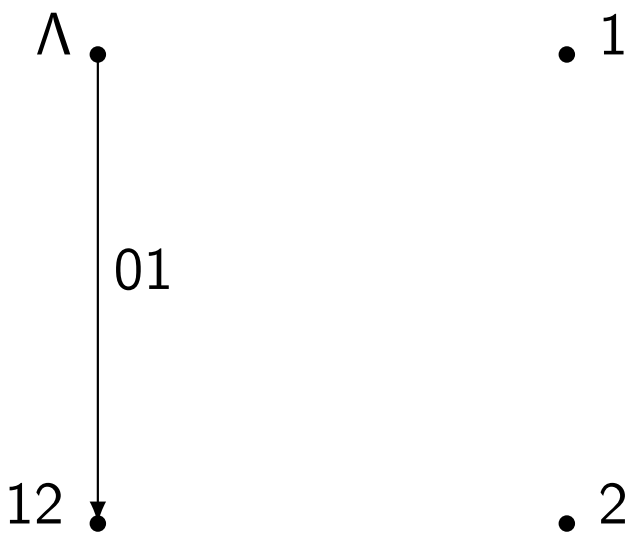
Решение. Построим граф $G = (V, E)$. Получаем: $V = \{\Lambda, 1, 2, 12\}$.



Гл. 7, 1.2(1)

Гл. 7, 1.2(1). Проверить, является ли алфавитный код $C = \{01, 201, 112, 122, 0112\}$ однозначным.

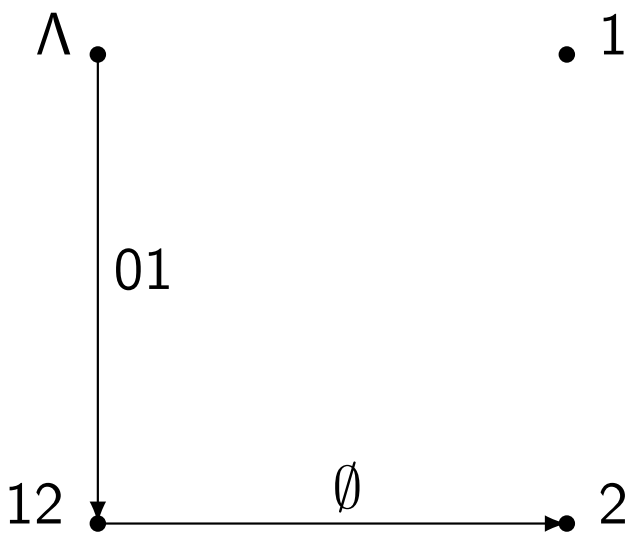
Решение. Построим граф $G = (V, E)$. Получаем: $V = \{\Lambda, 1, 2, 12\}$.



Гл. 7, 1.2(1)

Гл. 7, 1.2(1). Проверить, является ли алфавитный код $C = \{01, 201, 112, 122, 0112\}$ однозначным.

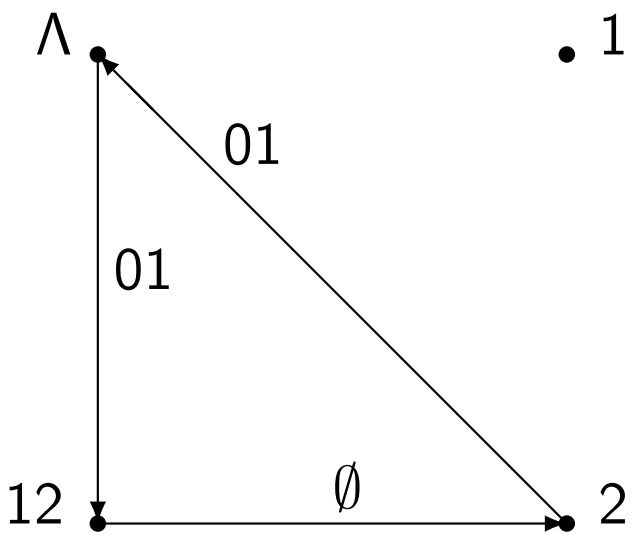
Решение. Построим граф $G = (V, E)$. Получаем: $V = \{\Lambda, 1, 2, 12\}$.



Гл. 7, 1.2(1)

Гл. 7, 1.2(1). Проверить, является ли алфавитный код $C = \{01, 201, 112, 122, 0112\}$ однозначным.

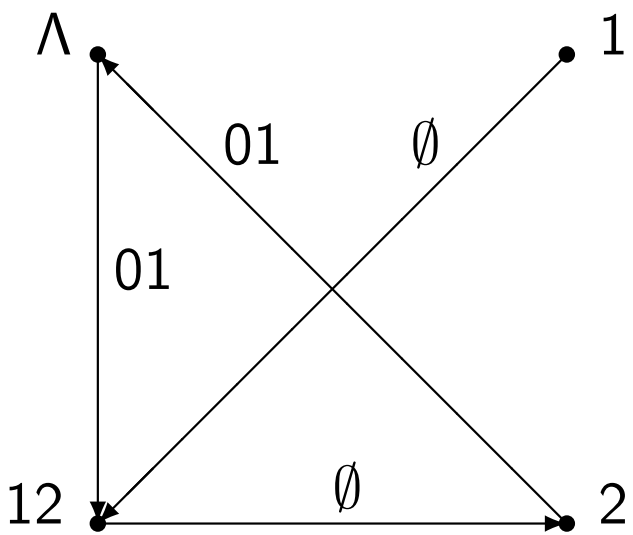
Решение. Построим граф $G = (V, E)$. Получаем: $V = \{\Lambda, 1, 2, 12\}$.



Гл. 7, 1.2(1)

Гл. 7, 1.2(1). Проверить, является ли алфавитный код $C = \{01, 201, 112, 122, 0112\}$ однозначным.

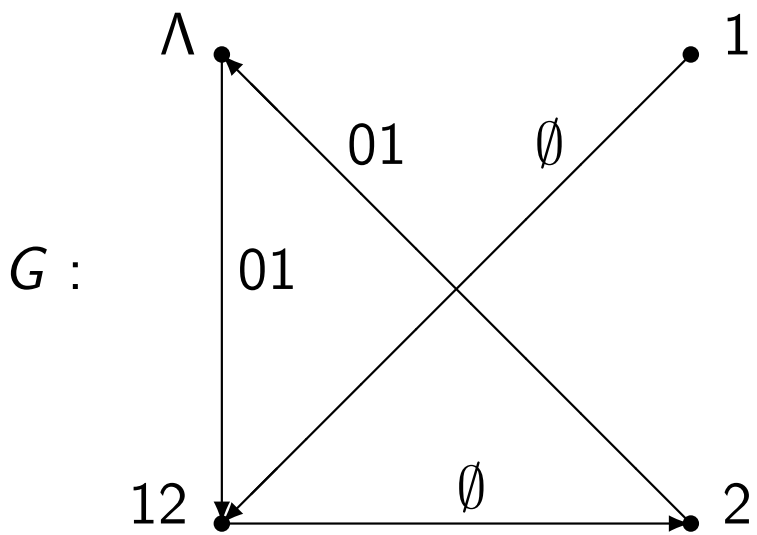
Решение. Построим граф $G = (V, E)$. Получаем: $V = \{\Lambda, 1, 2, 12\}$.



Гл. 7, 1.2(1)

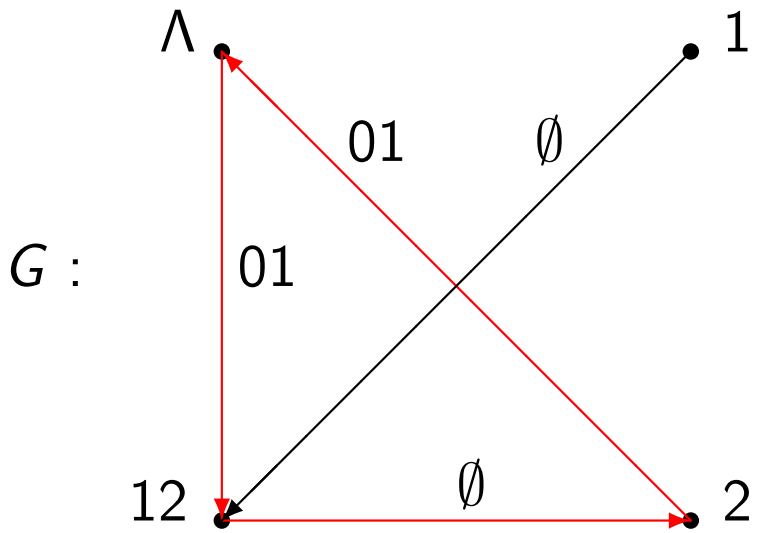
Гл. 7, 1.2(1). Проверить, является ли алфавитный код $C = \{01, 201, 112, 122, 0112\}$ однозначным.

Решение. Построим граф $G = (V, E)$. Получаем: $V = \{\Lambda, 1, 2, 12\}$.



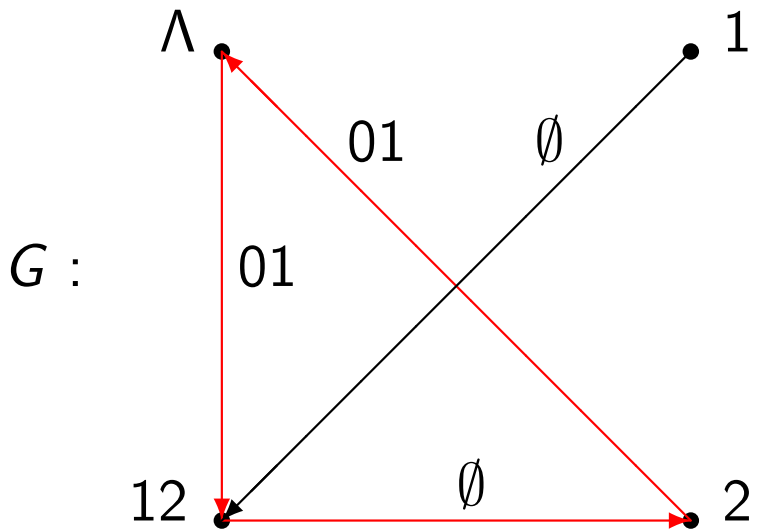
Гл. 7, 1.2(1)

Решение (продолжение). В графе G кода C найдется цикл, проходящий через вершину Λ. Значит, код C — не однозначный.



Гл. 7, 1.2(1)

Решение (продолжение). В графе G кода C найдется цикл, проходящий через вершину Λ. Значит, код C — не однозначный.



Кроме того,

$$\beta = 0112201 = 0112 + 201 = 01 + 122 + 01.$$

Для самостоятельного разбора: гл. 7, 1.2

Гл. 7, 1.2(2, 3, 7). Проверить, является ли алфавитный код $C = \{01, 201, 112, 122, 0112\}$ однозначным:

2) $C = \{001, 021, 102, 201, 001121, 01012101\}$;

3) $C = \{0, 01, 0010001001\}$;

7) $C = \{01, 12, 021, 0102, 10112\}$.

Для решения задач

Неравенство Макмиллана

Теорема (неравенство Макмиллана). *Пусть* $C_\varphi = \{B_1, \ldots, B_r\}$ *— алфавитный код в кодирующем алфавите* B, $|B| = q$, *и* $|B_i| = l_i$, $i = 1, \ldots, r$. *Если код* C_φ *— разделим, то верно неравенство:*

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

Префиксный код с заданными длинами кодовых слов

Теорема (о существовании префиксного кода с заданными длинами кодовых слов). *Пусть $q, l_1, \ldots, l_r$ — такие натуральные числа, что выполняется неравенство:*

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

Тогда существует такой **префиксный** *код $C = \{B_1, \ldots, B_r\}$ в любом кодирующем алфавите из q букв, что $|B_i| = l_i$ для всех $i = 1, \ldots, r$.*

Префиксные коды

Теорема (о существовании префиксного кода с теми же длинами кодовых слов). *Если* $C = \{B_1, \ldots, B_r\}$ — *разделимый алфавитный код в кодирующем алфавите* B, *то найдется такой* **префиксный** *код* $C' = \{B'_1, \ldots, B'_r\}$ *в том же алфавите* B, *что* $|B'_i| = |B_i|$ *для всех* $i = 1, \ldots, r$.

Гл. 7, 1.7

Гл. 7, 1.7. Найдется ли разделимый алфавитный код в кодирующем алфавите из $q = 3$ букв с длинами кодовых слов:

$$1,\ 1,\ 2,\ 2,\ 2,\ 2?$$

Гл. 7, 1.7

Гл. 7, 1.7. Найдется ли разделимый алфавитный код в кодирующем алфавите из $q = 3$ букв с длинами кодовых слов:

$$1,\ 1,\ 2,\ 2,\ 2,\ 2?$$

Решение. Проверим выполнение неравенства Макмиллана для такого кода:

$$\frac{2}{3^1} + \frac{4}{3^2} = \frac{2}{3} + \frac{4}{9} = \frac{10}{9} > 1.$$

Гл. 7, 1.7

Гл. 7, 1.7. Найдется ли разделимый алфавитный код в кодирующем алфавите из $q = 3$ букв с длинами кодовых слов:

$$1,\ 1,\ 2,\ 2,\ 2,\ 2?$$

Решение. Проверим выполнение неравенства Макмиллана для такого кода:

$$\frac{2}{3^1} + \frac{4}{3^2} = \frac{2}{3} + \frac{4}{9} = \frac{10}{9} > 1.$$

Если бы нашелся такой разделимый код, то сумма в левой части не превосходила бы единицу, что не так. Значит, такой разделимый код не найдется.

Гл. 7, 1.7

Гл. 7, 1.7. Найдется ли разделимый алфавитный код в кодирующем алфавите из $q = 3$ букв с длинами кодовых слов:

$$1,\ 2,\ 2,\ 3,\ 3,\ 3?$$

Гл. 7, 1.7

Гл. 7, 1.7. Найдется ли разделимый алфавитный код в кодирующем алфавите из $q = 3$ букв с длинами кодовых слов:

$$1,\ 2,\ 2,\ 3,\ 3,\ 3?$$

Решение. Проверим выполнение неравенства Макмиллана для такого кода:

$$\frac{1}{3^1} + \frac{2}{3^2} + \frac{3}{3^3} = \frac{1}{3} + \frac{2}{9} + \frac{3}{27} = \frac{2}{3} \leqslant 1.$$

Гл. 7, 1.7

Гл. 7, 1.7. Найдется ли разделимый алфавитный код в кодирующем алфавите из $q = 3$ букв с длинами кодовых слов:

$$1,\ 2,\ 2,\ 3,\ 3,\ 3?$$

Решение. Проверим выполнение неравенства Макмиллана для такого кода:

$$\frac{1}{3^1} + \frac{2}{3^2} + \frac{3}{3^3} = \frac{1}{3} + \frac{2}{9} + \frac{3}{27} = \frac{2}{3} \leqslant 1.$$

Построим префиксный код с такими длинами кодовых слов в кодирующем алфавите $B = \{0, 1, 2\}$:

$$B_1 = 0,\ B_2 = 10,\ B_3 = 11,$$
$$B_4 = 120,\ B_5 = 121, B_6 = 122.$$

Для самостоятельного разбора: гл. 7, 1.7, 1.8

Гл. 7, 1.7(1, 3, 5). Проверить, найдется ли разделимый алфавитный код в кодирующем алфавите из q букв с заданным набором длин кодовых слов:

1) $L = (1, 2, 2, 3)$, $q = 2$;

3) $L = (2, 2, 2, 4, 4, 4)$, $q = 2$;

5) $L = (1, 1, 2, 2, 3, 3, 3)$, $q = 3$.

Гл. 7, 1.6(1, 3, 4). Построить разделимый алфавитный код в кодирующем алфавите из $q = 2$ букв с заданным набором длин кодовых слов:

1) $L = (1, 2, 3, 3)$;

3) $L = (2, 2, 3, 3, 4, 4, 4, 4)$;

4) $L = (2, 2, 2, 4, 4, 4)$.

Для решения задач

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.

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

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.

Гл. 7, 2.1

Гл. 7, 2.1(3). Построить оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$.

Гл. 7, 2.1

Гл. 7, 2.1(3). Построить оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$.

Решение.

• • • • •

Гл. 7, 2.1

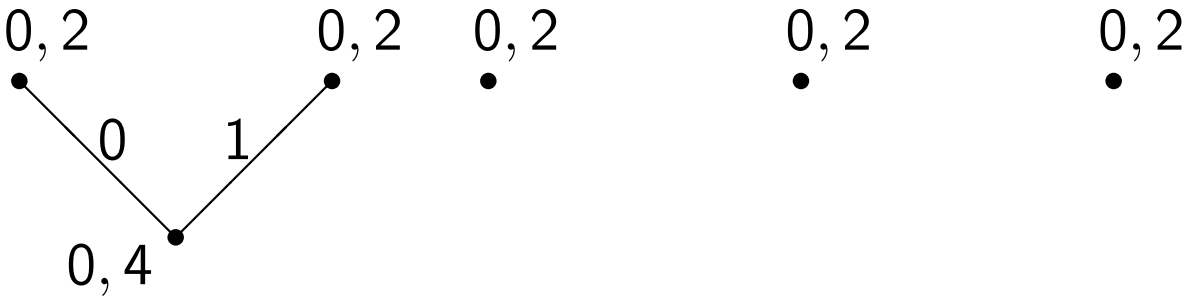
Гл. 7, 2.1(3). Построить оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$.

Решение.

Гл. 7, 2.1

Гл. 7, 2.1(3). Построить оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$.

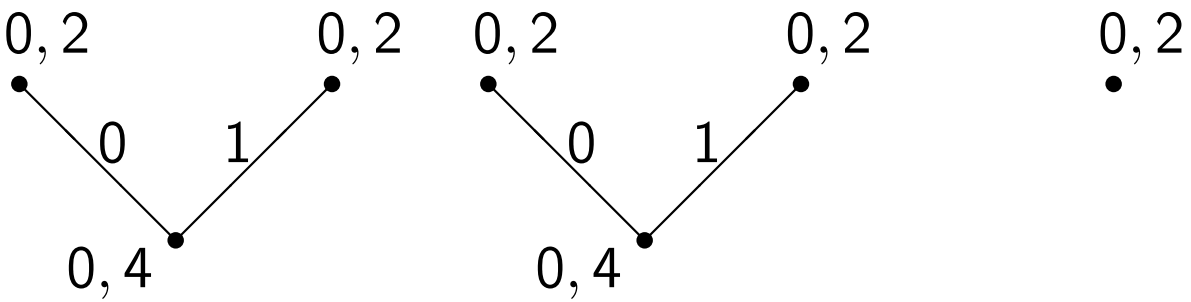
Решение.



Гл. 7, 2.1

Гл. 7, 2.1(3). Построить оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$.

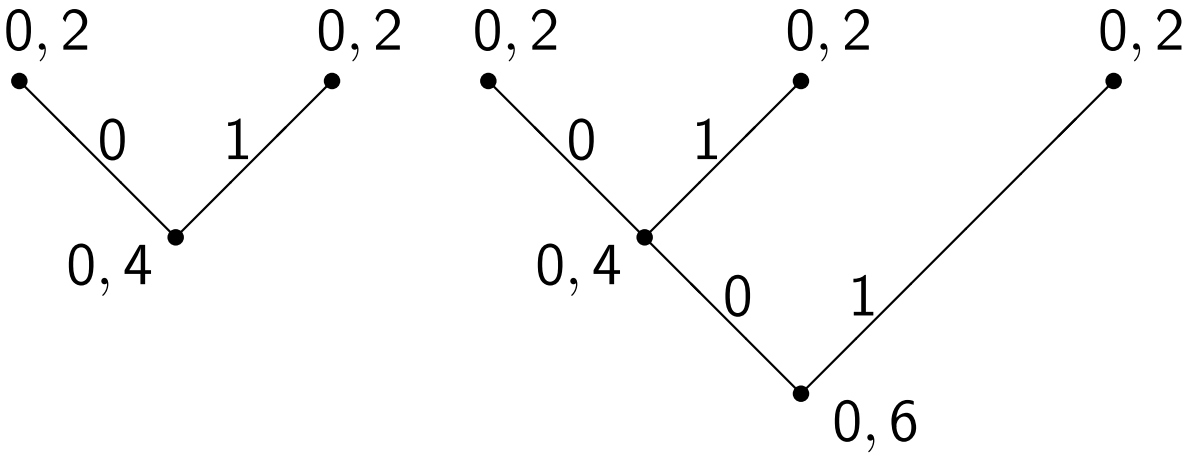
Решение.



Гл. 7, 2.1

Гл. 7, 2.1(3). Построить оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$.

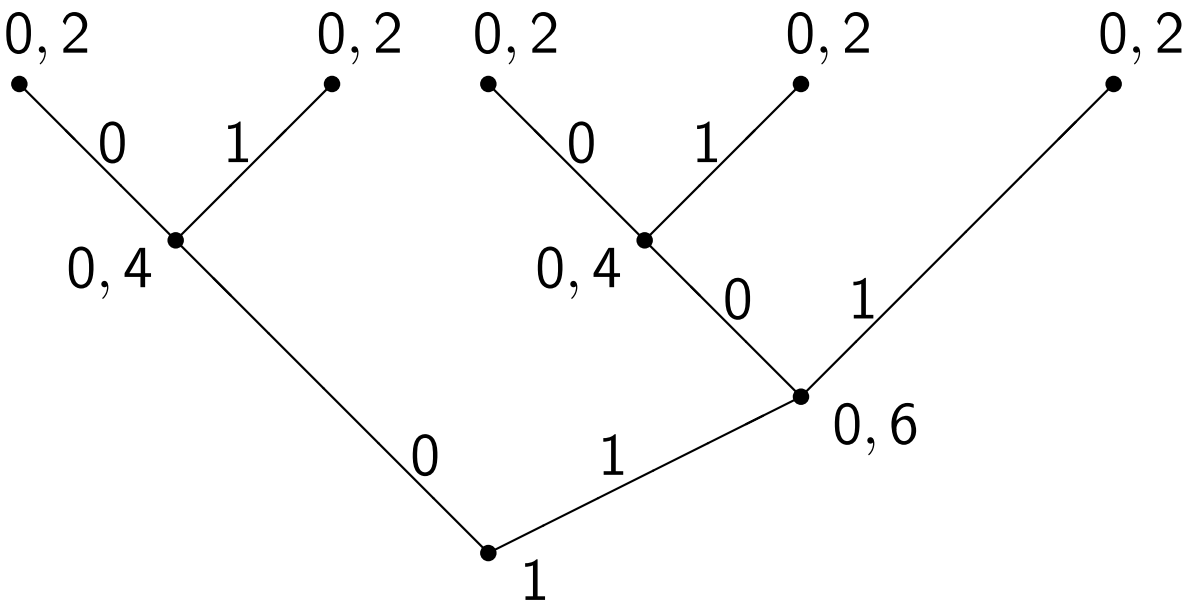
Решение.



Гл. 7, 2.1

Гл. 7, 2.1(3). Построить оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$.

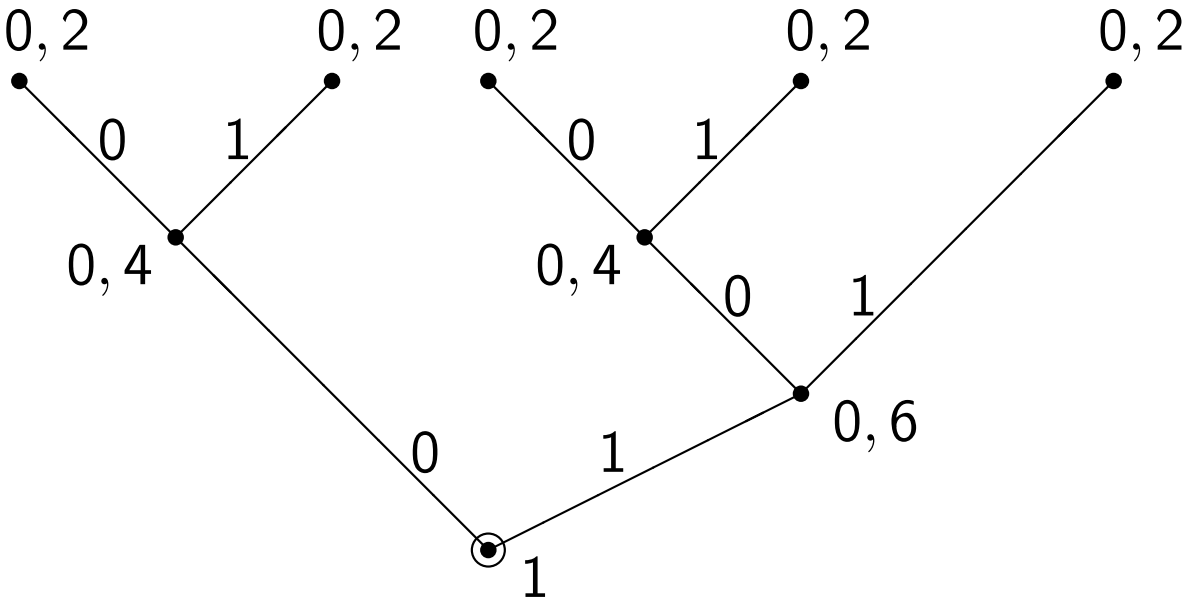
Решение.



Гл. 7, 2.1

Гл. 7, 2.1(3). Построить оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$.

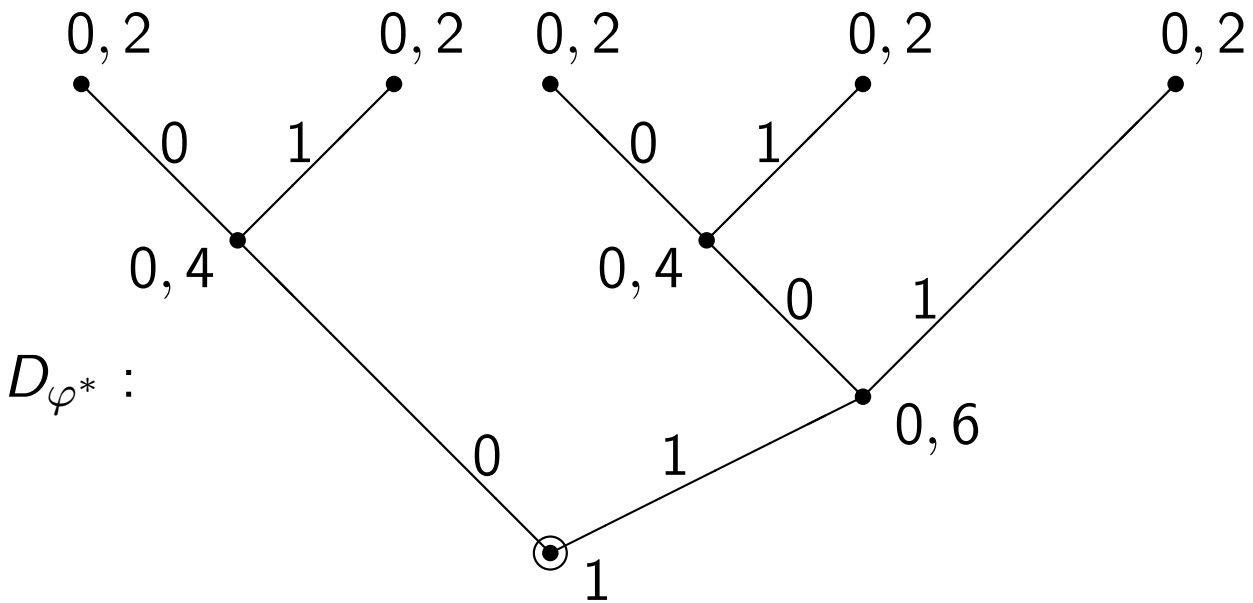
Решение.



Гл. 7, 2.1

Гл. 7, 2.1(3). Построить оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$.

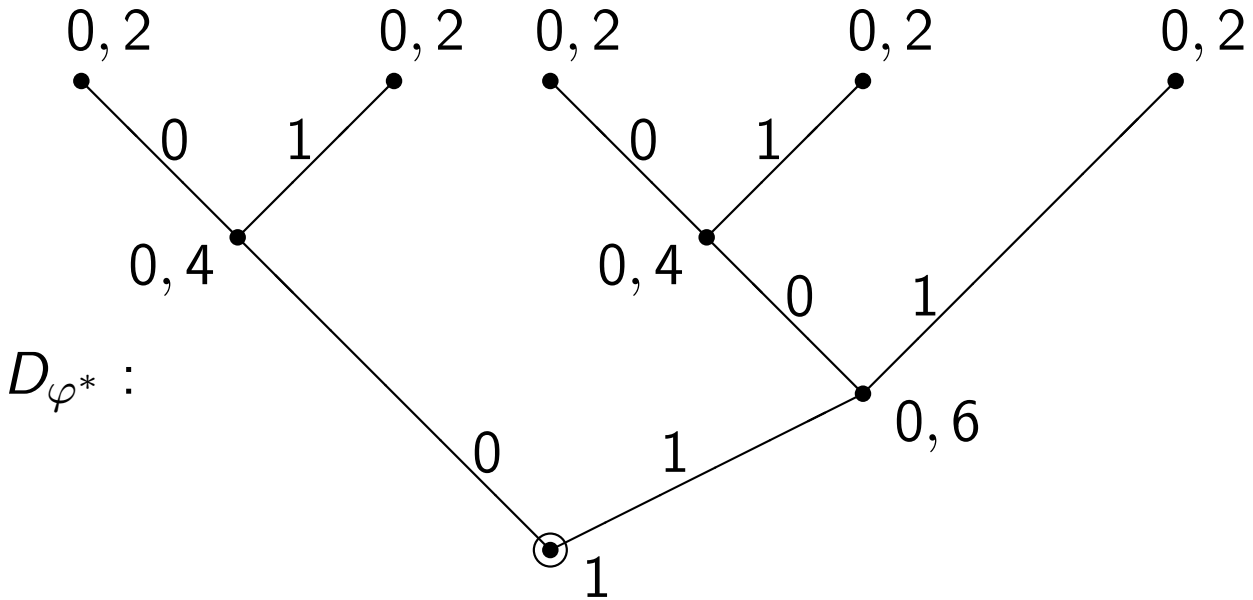
Решение.



Гл. 7, 2.1

Гл. 7, 2.1(3). Построить оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$.

Решение.

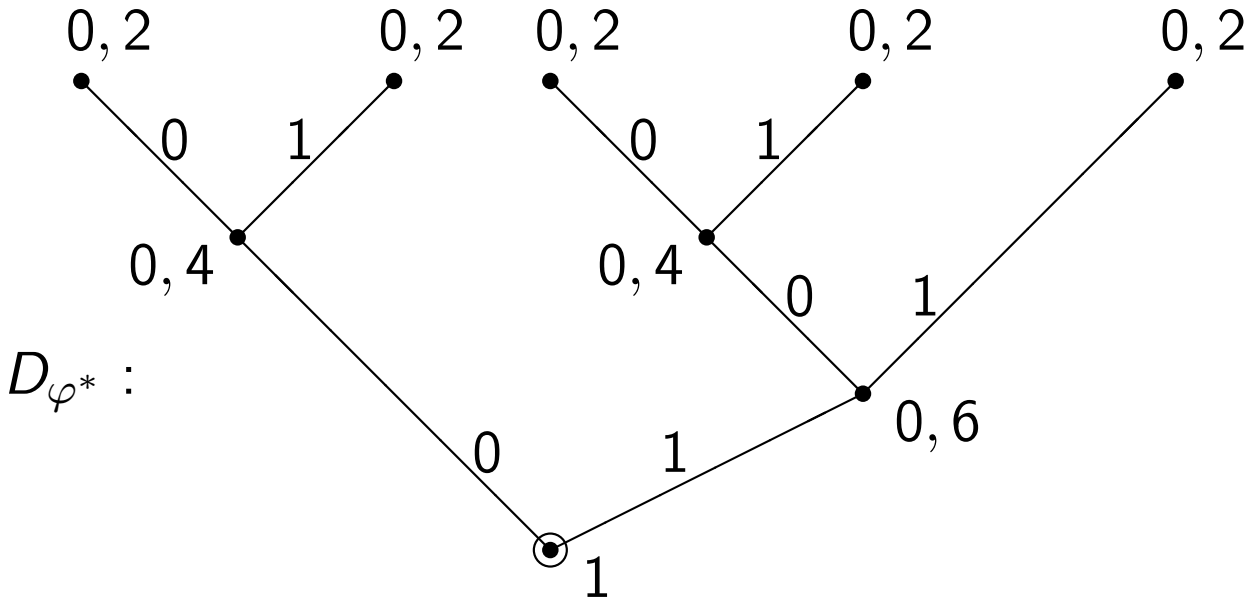


Получаем: $C_{\varphi^*} = \{00, 01, 100, 101, 11\}$.

Гл. 7, 2.1

Гл. 7, 2.1(3). Построить оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для набора частот $P = (0,2; 0,2; 0,2; 0,2; 0,2)$.

Решение.



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

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

Для самостоятельного разбра: гл. 7, 2.1

Гл. 7, 2.1(1, 2, 4). Построить оптимальный префиксный код в кодирующем алфавите $B = \{0, 1\}$ для заданного набора частот P.

1) $P = (0,4; 0,2; 0,2; 0,2)$;

2) $P = (0,7; 0,1; 0,1; 0,1)$;

4) $P = (0,5; 0,2; 0,1; 0,09; 0,08; 0,03)$.

Гл. 7, 2.10(1, 3, 4). Проверить, найдется ли **оптимальный** алфавитный код в кодирующем алфавите из $q = 2$ букв с заданным набором длин кодовых слов:

1) $L = (2, 3, 3, 3)$;

3) $L = (1, 3, 3, 3, 3)$;

4) $L = (1, 2, 3, 4)$.

Для решения задач

Домашнее задание

По задачнику: Гаврилов Г. П., Сапоженко А. А. Задачи и упражнения по дискретной математике. М.: Физматлит, 2004.

Гл. 7: 1.2(4, 5, 6), 1.6(2, 5, 6), 1.7(4, 6), 1.8, 2.1(5, 7, 8), 2.10(2, 5, 6).