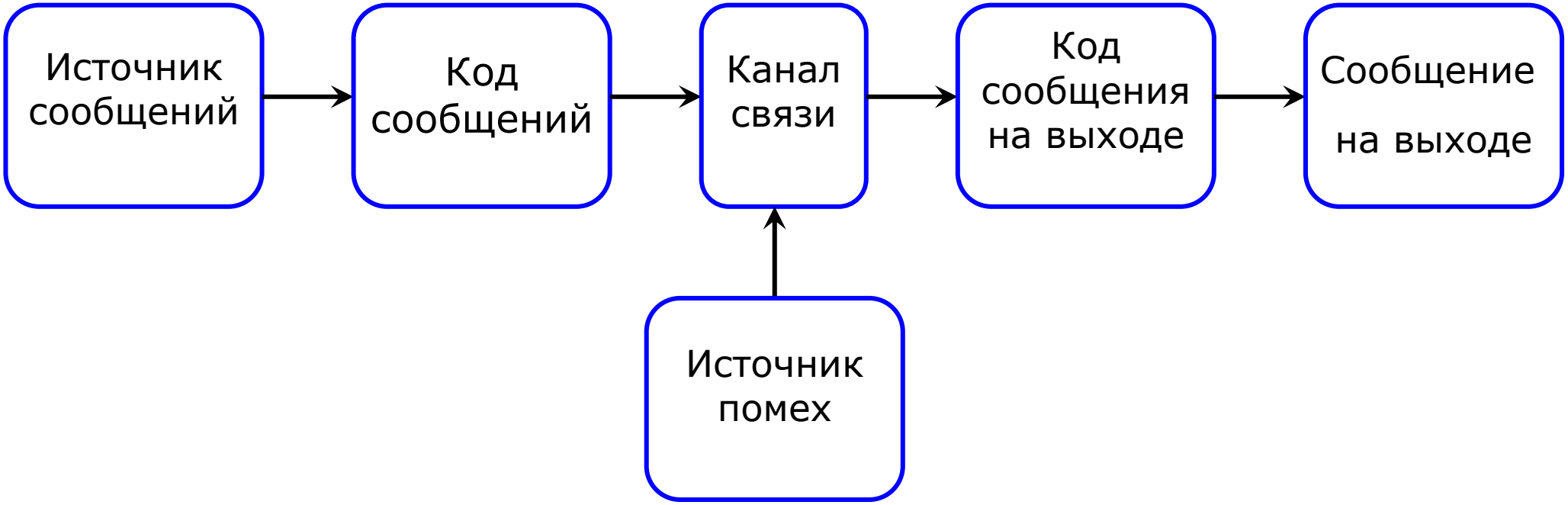
КОДИРОВАНИЕ



В этой схеме источник сообщений хочет передать по каналу связи некоторый набор *слов* — конечных последовательностей символов из заданного конечного алфавита $\mathcal{A} = \{a_1, \ldots, a_r\}$. Для передачи ему нужно (или он хочет) *закодировать* это сообщение — переписать его словами во вспомогательном алфавите $\mathcal{B} = \{b_1, \ldots, b_q\}$. После получения сообщения (возможно искаженного помехами) его нужно снова записать словами в алфавите $\mathcal{A}$ (возможно исправив возникшие ошибки).

Выбор кодов связан с различными обстоятельствами, а именно:

- с удобством передачи кодов,
- со стремлением увеличить пропускную способность канала,
- с удобством обработки кодов,
- с обеспечением помехоустойчивости,
- с удобством декодирования,
- с необходимостью однозначного декодирования,
- с другими возможными требованиями к кодам.

Ниже будут рассматриваться два вида кодирования:

(а) Алфавитное кодирование. Каждой букве a_i из $\mathcal{A} = \{a_1, \ldots, a_r\}$ ставится в соответствие некоторое слово B_i из алфавита $\mathcal{B} = \{b_1, \ldots, b_q\}$. Схема кодирования, сопоставляющая эти слова, будет обозначаться буквой Σ.

(б) Равномерное кодирование. Некоторое слово B_i из алфавита $\mathcal{B}$ ставится в соответствие не букве, а какому-то слову A_i фиксированной длины в алфавите $\mathcal{A}$.

Конечно, одно из первых требований к используемому коду — требование однозначности восстановления сообщения по его коду.

Проверка однозначности декодирования

Рассмотрим алфавитные коды.

Каждое из слов B_i, $i = 1,\ldots, r$, называется *элементарным кодом*.

Слово в алфавите $\mathcal{B}$ назовем кодовым, если его можно *расшифровать*, т.е. разбить на элементарные коды.

Одна из трудностей проверки однозначности декодирования состоит в том, что формально надо проверять бесконечное число кодовых слов.

Оказывается, этой бесконечности можно избежать.

Пусть дана схема кодирования Σ и l_i — длина слова B_i, $L = l_1 + \ldots + l_r$.

Назовем *нетривиальным разложением* слова B_i его представление в виде $B_i = \beta' B_{j_1} \ldots B_{j_w} \beta''$, где $B_{j_1} \neq B_i$, β'' является началом какого-нибудь элементарного кода, а β' является концом какого-нибудь элементарного кода. Слова β' и β'' могут быть пустыми.

Пример.

Σ: $A_1 = (1\ 0\ 0\ 1)$ $l_1 = 4$

$A_2 = (0)$ $l_2 = 1$

$A_3 = (0\ 1\ 0)$ $l_3 = 3$

Рассмотрим слово $B = 0\ 1\ 0\ 0\ 1\ 0 = A_2\, A_1\, A_2 = A_3\, A_3$

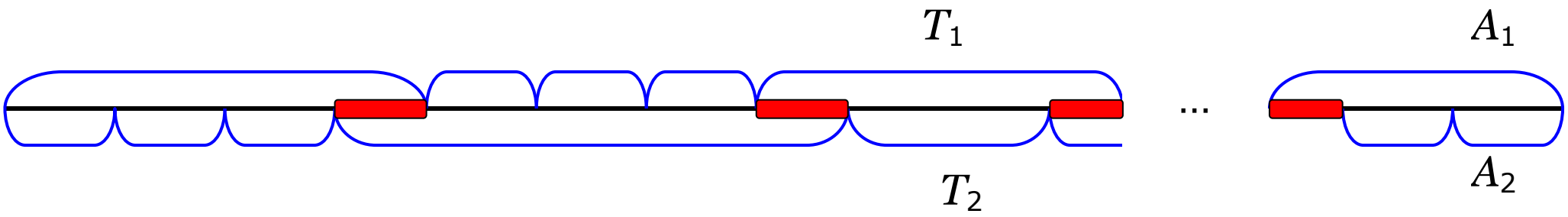
Нет однозначности декодирования!

Очевидно, что для каждого i число нетривиальных разложений слова B_i конечно. Обозначим через W максимум чисел w, взятый по всем нетривиальным разложениям всех слов B_i, $i = 1,\ldots, r$.

Теорема 1. Для любой схемы кодирования Σ найдется такое $N = N(\Sigma)$, что для проверки однозначности декодирования в Σ достаточно проверить коды слов из $\mathcal{A}$ длины не более N, и

$$N \le \lfloor (W + 1)(L - r + 2)/2 \rfloor.$$

Доказательство. Выберем самое короткое слово B в алфавите $\mathcal{B}$, допускающее две различные расшифровки A_1 и A_2. С ними связаны два разбиения слова B на элементарные коды T_1 и T_2 :



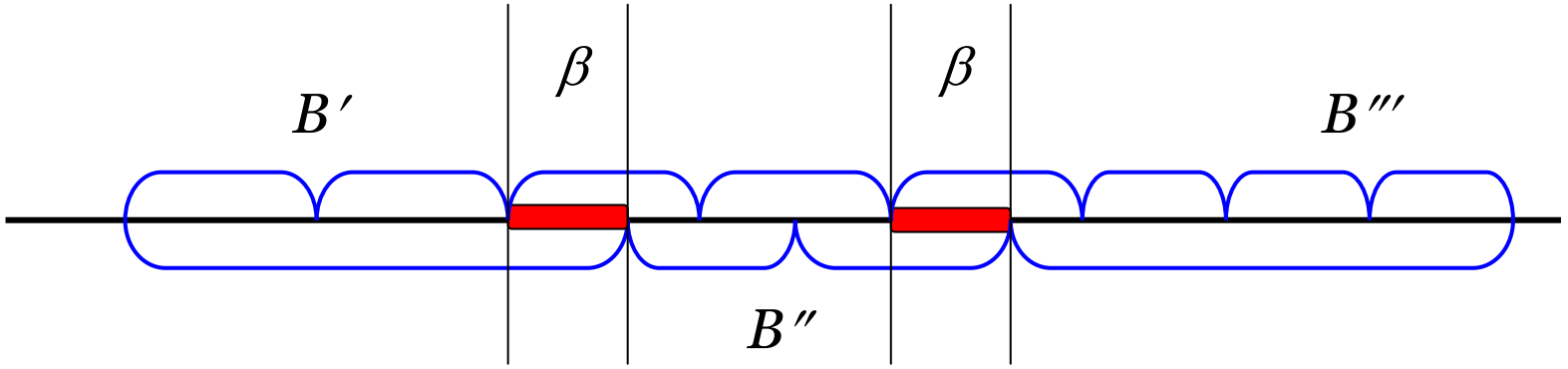
Обозначим через T разбиение, полученное после «разрезания» В там, где его «разрезало» хотя бы одно из разбиений T_1 и T_2. Части разбиения T разделим на два класса: к первому отнесем части, являющиеся элементарными кодами, ко второму — все остальные (огрызки).

Каждая часть β, принадлежащая второму классу, является концом одного из элементарных кодов и началом другого. Причем если β оканчивает некоторое элементарное кодовое слово в T_1, то оно начинает какое-то элементарное кодовое слово в T_2 и наоборот (см. рис.).

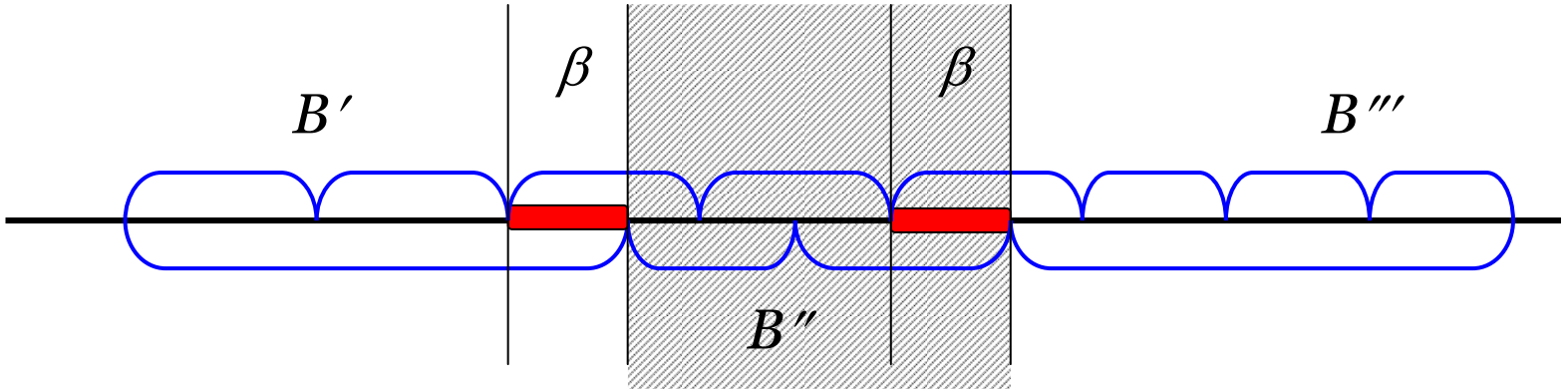
Более точно, если $B = B'\beta\ B''$, то либо $B'\beta$ и B'' являются кодовыми словами в T_1, а B' и $\beta\ B''$ являются кодовыми словами в T_2, либо наоборот.

Покажем, что все части из второго класса различны.

Допустим, что $B = B'\beta\, B''\beta\, B'''$.



Тогда слово $B'\beta\, B'''$ имеет две расшифровки в противоречие с выбором B. Чтобы убедиться в этом, заметим, что согласно вышесказанному, слова $B'\beta$, B', $\beta B'''$ и B''' являются кодовыми.

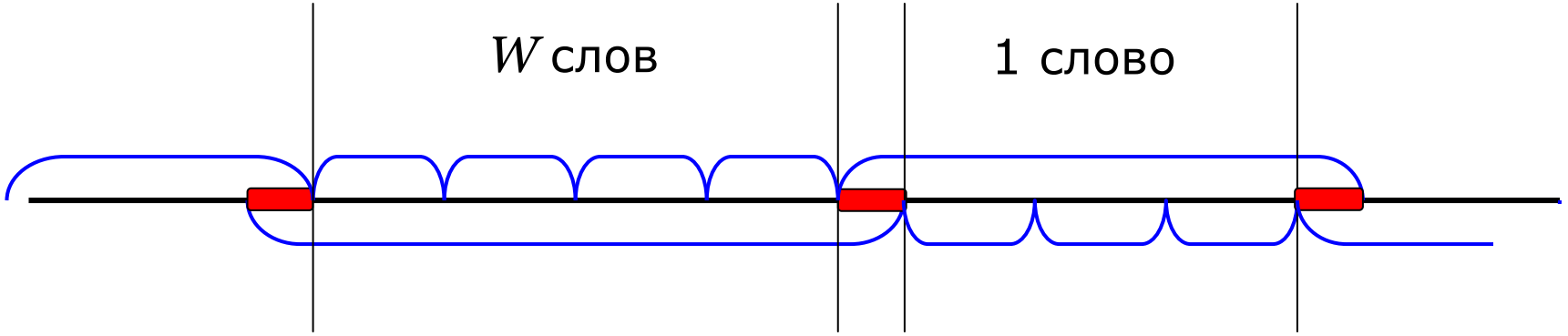


Число огрызков не превосходит числа непустых начал элементарных кодов, т.е. $(l_1 - 1) + \ldots + (l_r - 1) = L - r$.

Они дают не более $L - r + 1$ кусков.

Каждый из кусков, на которые разбивается B после выбрасывания всех огрызков, является кодовым словом, входящим в одно из разбиений T_i, и частью некоторого элементарного кода, входящего в T_{3-i}.

Соседние куски являются частями элементарных кодов, входящих в различные T_i.



Два соседних куска

Имеем не более $L - r + 1$ кусков. Рассматриваем их парами.

Всего пар $\lfloor (L - r + 1)/2 \rfloor$.

В каждой паре не более $W + 1$ слов.

Следовательно, длина каждого из A_i не превосходит

$$W \cdot \lceil (L - r + 1)/2 \rceil + 1 \cdot \lfloor (L - r + 1)/2 \rfloor \le \lfloor (W + 1)(L - r + 2)/2 \rfloor . \quad \blacksquare$$

Пример.

$r = 6$, $W = 3$, $L = 20$,

$\lfloor (W + 1)(L - r + 2)/2 \rfloor = \lfloor 4 \cdot 16/2 \rfloor = 32$,

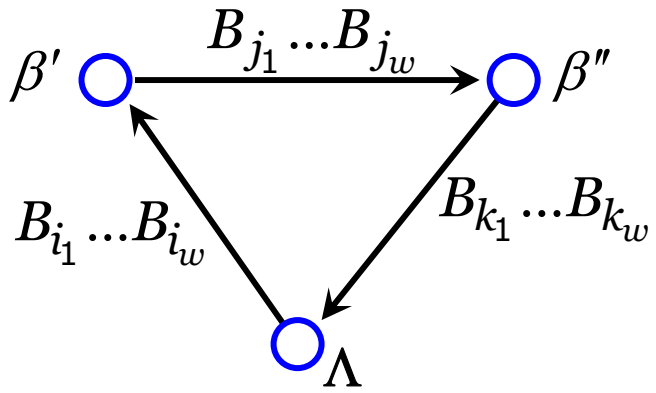
то есть требуется проверить 6^{32} слов.

Из доказательства теоремы можно извлечь существенно более эффективный алгоритм.

Пусть дана схема кодирования Σ. Для каждого элементарного кода B_i рассмотрим все его нетривиальные разложения

$$B_i = \beta' B_{j_1} \ldots B_{j_w} \beta''. \qquad (1)$$

Обозначим через $V = V(\Sigma)$ множество, содержащее пустое слово Λ и слова β, встречающиеся в разложениях вида (1) как в виде начал, так и в виде окончаний. Построим далее помеченный ориентированный граф $\Gamma = \Gamma(\Sigma)$ по следующим правилам. Множеством вершин графа Γ является $V = V(\Sigma)$. Проводим дугу из вершины $\beta' \in V$ в вершину $\beta'' \in V$, если и только если в некотором разложении вида (1) β' является началом, а β'' — концом. При этом дуга (β', β'') помечается словом $B_{j_1} \ldots B_{j_w}$.



$$B_1 = \beta' B_{j_1} \ldots B_{j_w} \beta''$$

$$B_2 = B_{i_1} \ldots B_{i_w} \beta'$$

$$B_3 = \beta'' B_{k_1} \ldots B_{k_w}$$

Теорема 2. Схема кодирования Σ не обладает свойством однозначности декодирования тогда и только тогда, когда граф $\Gamma(\Sigma)$ содержит контур, проходящий через вершину Λ.

Доказательство. Допустим, что Σ не обладает свойством однозначности декодирования. Тогда, как следует из доказательства теоремы 1, кратчайшее слово, имеющее две расшифровки в схеме Σ, имеет вид

$$B = B_{i_{1,1}} \ldots B_{i_{1,k(1)}} \beta_1 B_{i_{2,1}} \ldots B_{i_{2,k(2)}} \beta_2 \ldots \beta_{s-1} B_{i_{s,1}} \ldots B_{i_{s,k(s)}},$$

где все β_i различны и слова $B_{i_{1,1}} \ldots B_{i_{1,k(1)}}$, β_1, $\beta_1 B_{i_{2,1}} \ldots B_{i_{2,k(2)}} \beta_2, \ldots$ $\ldots, \beta_{s-1} B_{i_{s,1}} \ldots B_{i_{s,k(s)}}$ являются элементарными кодами. Это значит, что в $\Gamma(\Sigma)$ есть контур, проходящий через вершины Λ, $\beta_1, \ldots, \beta_{s-1}$.

Обратно, пусть в $\Gamma(\Sigma)$ существует контур, проходящий через вершины $\beta_0, \beta_1, \ldots, \beta_{s-1}$, где $\beta_0 = \Lambda$ и дуга (β_j, β_{j+1}), $j = 0, 1, \ldots, s-1$, $((s-1)+1=0)$, помечена словом $B_{i_{j+1,1}} \ldots B_{i_{j+1,k(j+1)}}$. Тогда слово

$$B = B_{i_{1,1}} \ldots B_{i_{1,k(1)}} \beta_1 B_{i_{2,1}} \ldots B_{i_{2,k(2)}} \beta_2 \ldots \beta_{s-1} B_{i_{s,1}} \ldots B_{i_{s,k(s)}},$$

имеет две различные расшифровки. ■

Пример. Σ: $a_1 — b_1\, b_2$

$a_2 — b_1\, b_3\, b_2$

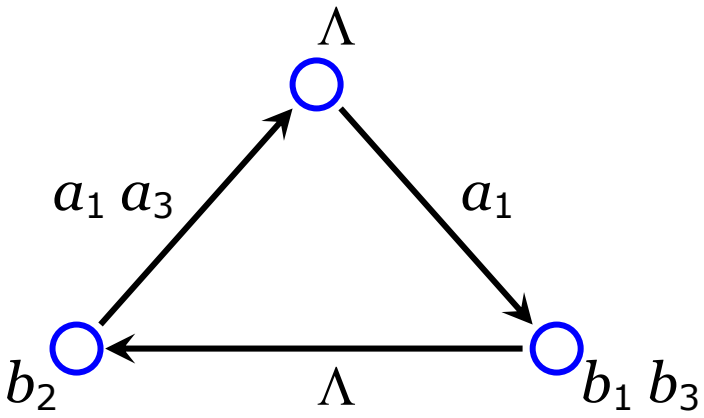
$a_3 — b_2\, b_3$

$a_4 — b_1\, b_2\, b_1\, b_3$

$a_5 — b_2\, b_1\, b_2\, b_2\, b_3$

Находим все префиксы, которые одновременно являются суффиксами и не являются кодовыми словами:

$\{\Lambda,\ b_2,\ b_1\, b_3\}$, то есть три вершины в графе



$a_1\, a_2\, a_1\, a_3 =$

$= b_1\, b_2\, b_1\, b_3\, b_2\, b_1\, b_2\, b_2\, b_3 =$

$= a_4\, a_5$

Пример.

Σ: $a_1 — b_1$

$a_2 — b_2\, b_1$

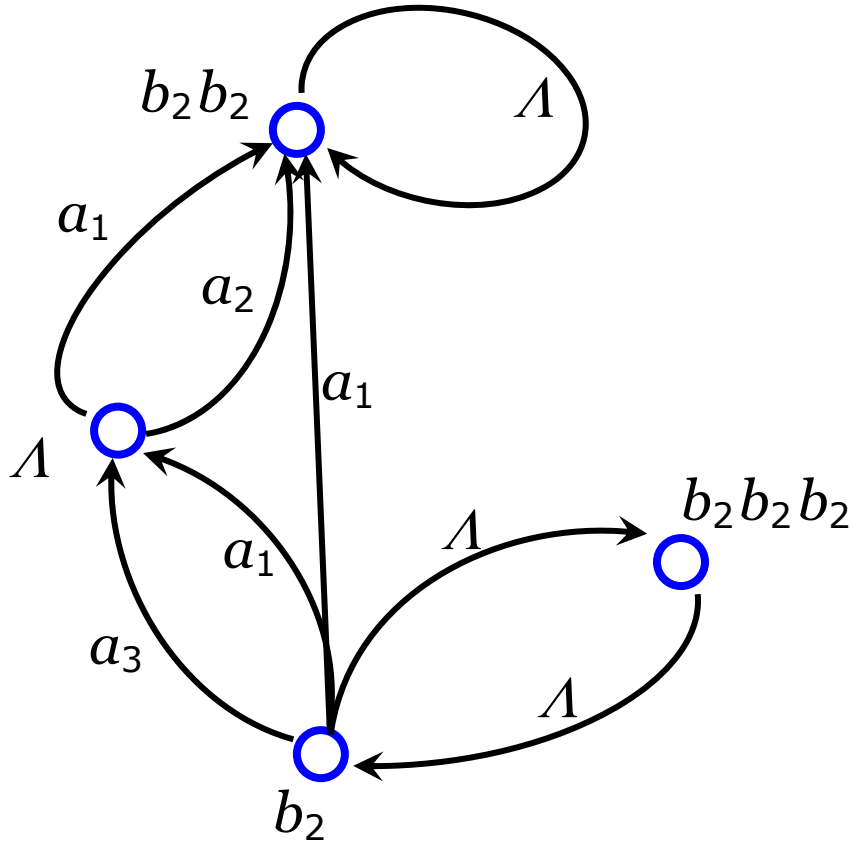
$a_3 — b_1 b_2\, b_2$

$a_4 — b_2\, b_1\, b_2\, b_2$

$a_5 — b_2\, b_2\, b_2\, b_2$

Находим все β: $\{\Lambda, b_2,\ b_2\, b_2,\ b_2\, b_2\, b_2\}$

Тогда получаем граф:



Нет цикла через вершину Λ.

Код однозначно декодируется.

Префиксные коды

Важным классом однозначно декодируемых кодов являются *префиксные коды* — такие алфавитные коды, где ни один элементарный код не является *префиксом* (т.е. началом) другого элементарного кода.

Упражнение. Доказать, что любой префиксный код является однозначно декодируемым.

Обозначим через q значность алфавита, например, $q = 2$, и $l_i = l(B_i)$, $i = 1, \ldots, r$.

Теорема 3. (Неравенство Макмиллана) Если схема кодирования Σ обладает свойством однозначности декодирования, то

$$\sum_{i=1}^{r} q^{-l_i} \le 1. \qquad (2)$$

Доказательство. Выберем произвольное n. Рассмотрим коды всех r^n слов длины n в алфавите $\mathcal{A}$, полученные с помощью Σ. Все они могут быть порождены выражением

$$(a_1 + \ldots + a_r)^n,$$

если рассматривать произведение $a_{i_1} a_{i_2} \ldots a_{i_n}$ как запись слова. Имеем

$$(a_1 + \ldots + a_r)^n = \sum_{(i_1 i_2 \ldots i_n)} a_{i_1} a_{i_2} \ldots a_{i_n} .$$

Соответствующие этим словам коды получаются заменой символов $a_1, \ldots, a_r$ на элементарные коды $B_1, \ldots, B_r$. Получаем

$$(B_1 + \ldots + B_r)^n = \sum_{(i_1 i_2 \ldots i_n)} B_{i_1} B_{i_2} \ldots B_{i_n} .$$

Этому тождеству соответствует

$$\left(\frac{1}{q^{l_1}} + \ldots + \frac{1}{q^{l_r}} \right)^n = \sum_{(i_1 \ldots i_n)} \frac{1}{q^{l_{i_1} + \ldots + l_{i_n}}} . \qquad (3)$$

Положим $t = l_{i_1} + \ldots + l_{i_n}$ и $\nu(n, t)$ — число кодовых слов $B_{i_1} B_{i_2} \ldots B_{i_n}$ длины t. Пусть $l = \max\limits_{1 \le i \le r} l_i$. Из взаимной однозначности алфавитного кодирования вытекает $\nu(n, t) \le q^t$ и длина каждого из наших кодовых слов не превосходит nl.

Следовательно,

$$\sum_{(i_1 \ldots i_n)} \frac{1}{q^{l_{i_1} + \ldots + l_{i_n}}} = \sum_{t=1}^{nl} \frac{\nu(n,t)}{q^t} \le nl.$$

Используя (3), получаем

$$\left(\frac{1}{q^{l_1}} + \ldots + \frac{1}{q^{l_r}} \right) \le \sqrt[n]{nl}$$

Это неравенство справедливо для любого n, а его правая часть стремится к 1 при $n \to \infty$. Поскольку его левая часть не зависит от n, необходимо, чтобы $q^{-l_1} + \ldots + q^{-l_r} \le 1$ ■

Следующий факт характеризует префиксные коды с положительной стороны.

Теорема 4. Если схема кодирования Σ обладает свойством однозначности декодирования, то существует такая префиксная схема кодирования Σ', что для каждого i, i=1,… , s длина l_i' элементарного кода B_i' в Σ' равна длине l_i элементарного кода B_i в Σ.

Доказательство. Можно считать, что элементарные коды B_i занумерованы в порядке неубывания их длин. Пусть длинами элементарных кодов в Σ являются числа λ_1,…, λ_s, $\lambda_1 < \lambda_2 < \ldots < \lambda_s$ и число элементарных кодов длины λ_i , i =1,…, s равно ν_i. Тогда неравенство Макмиллана можно переписать в виде

$$\sum_{t=1}^{s} \frac{\nu_t}{q^{\lambda_t}} \le 1. \qquad (4)$$

В частности, $\nu_1 / q^{\lambda_1} \le 1$, откуда $\nu_1 \le q^{\lambda_1}$. Выберем среди q^{λ_1} слов длины λ_1 в алфавите $\mathcal{B}$ произвольные ν_1 слов в качестве элементарных кодов $B'_1,\ldots,B'_{\nu_1}$. Перейдем к словам длины λ_2. Из (4) получаем

$$\frac{\nu_1}{q^{\lambda_1}} + \frac{\nu_2}{q^{\lambda_2}} \le 1,$$

$$\nu_2 \le q^{\lambda_2} - \nu_1 q^{\lambda_2 - \lambda_1}. \qquad (5)$$

Рассмотрим множество слов длины λ_2 в алфавите $\mathcal{B}$, не начинающихся с $B'_1,\ldots,B'_{\nu_1}$. В силу (5) из этого множества можно выбрать ν_2 каких-нибудь слов в качестве элементарных кодов $B'_{\nu_1+1},\ldots,B'_{\nu_1+\nu_2}$.

Далее из (4) получаем

$$\nu_3 \le q^{\lambda_3} - \nu_1 q^{\lambda_3-\lambda_1} - \nu_2 q^{\lambda_3-\lambda_2}$$

и строим ν_3 слов длины λ_3, не начинающихся с $B'_1,\ldots,B'_{\nu_1+\nu_2}$ и т.д. Через конечное число шагов построим нужное количество слов нужной длины. По построению новый код будет префиксным. ■