Школа лингвистики, 2018-19 уч. год
Дискретная математика для лингвистов
Примерное содержание лекций по кодированию (черновой вариант!!!)
16 декабря 2018 г. (Просьба о замеченных опечатках сообщать авторам по электронной почте на адрес vvkoch@yandex.ru и avmikhailovich@gmail.com)
В. В. Кочергин, А. В. Михайлович

1 Кодирование

Вопросы кодирования в математике возникали давно. Изначально они имели важное, но вспомогательное значение — например, изображение чисел в десятичной системе счисления, введение координат как алгебраических образов геометрических объектов. Толчком к формированию теории кодирования как самостоятельного раздела дискретной математики стало стремительное развитие таких направлений как связь и криптология, а также управляющих систем (простейшие примеры — диагностика двигателя или программирование сигнализации).

Процессы, так или иначе связанные с кодированием, схематично изображены на рис. 1.

Рис. 1.

Введем некоторые обозначения.

Пусть $A = \{a_1, \ldots, a_r\}$ — исходный алфавит, $r \geq 2$. Тогда $\alpha = a_{i_1} a_{i_2} \ldots a_{i_k}$, где $a_{i_j} \in A$, — *слово* над алфавитом A, *длина* $l(\alpha)$ слова α равна k.

Обозначим через Λ пустое слово. Отметим, что для произвольного слова α справедливы равенства $\Lambda\alpha = \alpha\Lambda = \alpha$; кроме того, $l(\Lambda) = 0$, $\{\Lambda\} \neq \varnothing$.

Множество всех слов длины k над алфавитом A обозначим через A^k. Положим

$$A^+ = \bigcup_{k \geq 1} A^k, \quad A^* = \bigcup_{k \geq 0} A^k.$$

Пусть $B = \{b_1, \ldots, b_q\}$ — второй алфавит, $S \subset A^*$. Отображение

$$F: \ S \to B^*$$

называется *кодированием*, множество S — *множеством сообщений*, произвольное слово α из множества S — *сообщением*, слово $F(\alpha)$ из множества B^* — *кодом сообщения* α, а множество $F(S)$ — *кодом*.

Кодирование обладает свойством *взаимной однозначности*, если для любых двух различных сообщений α_1 и α_2 из множества сообщений S выполняется соотношение $F(\alpha_1) \neq F(\alpha_2)$.

Пример 1. Равномерное кодирование.

Пусть $\{\alpha_1, \ldots, \alpha_m\} \subset A^*$. Определим множество сообщений S, $S \subset A^*$, как множество всех слов $\{\alpha_1, \ldots, \alpha_m\}^*$, рассматриваемое как подмножество множества A^*, т. е как множество слов над алфавитом A. Зададим *схему кодирования* Σ следующим образом:

$$\alpha_i \to \beta_i, \quad l(\beta_i) = n; \qquad i = 1, \ldots, m.$$

Тогда схема Σ определяет кодирование F, заданное на множестве сообщений S:

$$\alpha_{i_1} \ldots \alpha_{i_k} \xrightarrow{F} \beta_{i_1} \ldots \beta_{i_k}.$$

Другим примером кодирования является алфавитное (побуквенное) кодирование, к более подробному изучению которого и переходим.

1.1 Алфавитное кодирование

Пусть $A = \{a_1, \ldots, a_r\}$, $B = \{b_1, \ldots, b_q\}$. Схема Σ, определяемая следующим образом:

$$\begin{aligned} a_1 &\to v_1 = b_{11}b_{12}\ldots b_{1l_1}, \\ &\vdots \\ a_r &\to v_r = b_{r1}b_{r2}\ldots b_{rl_r}, \end{aligned} \tag{1}$$

порождает *алфавитное* (*побуквенное*) *кодирование* F:

$$a_{i_1}\ldots a_{i_k} \xrightarrow{F} v_{i_1}\ldots v_{i_k}.$$

Слова $v_1, \ldots, v_r$ называются *кодовыми словами*, множество кодовых слов $V = \{v_1, \ldots, v_r\}$ — *кодом алфавита* A. Для множества слов V будем допускать также название *алфавитный код* и даже просто код, хотя, формально говоря, в данной ситуации кодом будет множество слов $F(A^*)$.

Алфавитный код V называется *разделимым*, если из справедливости равенства

$$v_{i_1}v_{i_2}\ldots v_{i_s} = v_{j_1}v_{j_2}\ldots v_{j_t}$$

двух слов из множества V^* следуют равенства:

1) $s = t$;

2) $i_1 = j_1,\ i_2 = j_2,\ \ldots,\ i_s = j_s$.

Свойства разделимых кодов:

1. $v_i \neq v_j$ при $i \neq j$.

2. $v_i \neq \Lambda$, $i = 1, \ldots, r$.

3. Кодирование, осуществляемое с помощью разделимого кода, обладает свойством взаимной однозначности.

Примеры разделимых кодов:

1. Азбука Морзе.

2. Кодирование, задаваемое схемой:

$$\begin{aligned} a_1 &\to 010, \\ a_2 &\to 0. \end{aligned}$$

3. Кодирование, задаваемое схемой:

$$\begin{aligned} a_1 &\to 00, \\ a_2 &\to 01, \\ a_3 &\to 10, \\ a_4 &\to 11. \end{aligned}$$

Если слово $\alpha \in A^*$ представимо в виде $\alpha = \alpha_1\alpha_2$, где $\alpha_1, \alpha_2 \in A^*$, то говорят, что α_1 — *префикс* слова α, а α_2 — *суффикс* слова α.

Отметим, что слова Λ и α являются и префиксами, и суффиксами слова α.

Алфавитный код V называется *префиксным*, если никакое кодовое слово v_i не является префиксом другого кодового слова v_j $(i \neq j)$.

Алфавитный код V называется *суффиксным*, если никакое кодовое слово v_i не является суффиксом другого кодового слова v_j $(i \neq j)$.

Алфавитное кодирование из примера 3 является и префиксным, и суффиксным, а кодирование из примера 2 не является ни префиксным, ни суффиксным.

Очевидно, что и префиксность, и суффиксность алфавитного кодирования являются достаточными условиями разделимости, но как показывает тот же пример 2, не являются необходимыми условиями разделимости.

Префиксный код удобно представлять в виде корневого помеченного дерева. Пусть $A = \{a_1, \ldots, a_r\}$, $B = \{b_1, \ldots, b_q\}$ и префиксный код из алфавита A в алфавит B задается схемой

$$a_i \to b_{i1}b_{i2}\ldots b_{il_i}, \quad i = 1, \ldots, r.$$

Такому коду естественным образом сопоставляется дерево с корнем: из каждой вершины исходит не более q ребер, которым приписываются буквы из алфавита B, концевым вершинам (т. е. вершинам степени 1) соответствуют буквы алфавита A, причем если концевой вершине приписана буква a_i, то ребрам на пути от корня до этой концевой вершины последовательно приписаны буквы $b_{i1}, b_{i2}, \ldots, b_{il_i}$ (ребру, инцидентному корню, приписана буква b_{i1}).

Построенное таким образом корневое помеченное дерево будем называть *кодовым деревом*.

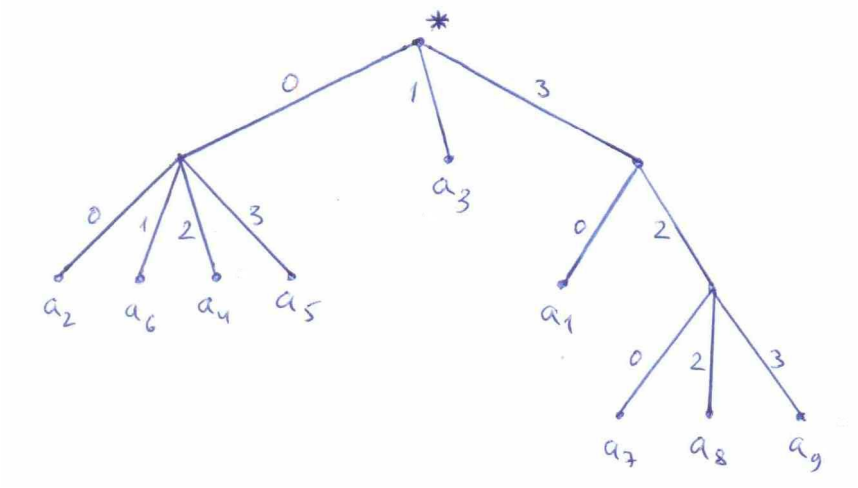

Рис. 1:

Пример 2. На рис. 1 представлено кодовое дерево для префиксного кода из алфавита $A = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8$ в алфавит $B = \{0, 1, 2, 3\}$, задаваемого схемой

$$a_1 \to 30,$$
$$a_2 \to 00,$$
$$a_3 \to 1,$$
$$a_4 \to 02,$$
$$a_5 \to 03,$$
$$a_6 \to 01,$$
$$a_7 \to 320,$$
$$a_8 \to 322,$$
$$a_9 \to 323.$$

Корень дерева помечен звездочкой.

Теорема 1 (неравенство Крафта — МакМиллана)**.** *Пусть разделимый код задается схемой* $a_i \to v_i$, $v_i \in B^*$, $i = 1, \ldots, r$. *Тогда*

$$\sum_{i=1}^{r} \frac{1}{q^{l(v_i)}} \leq 1,$$

где $q = |B|$.

Доказательство. Зафиксируем $n \in \mathbb{N}$. Обозначим через l максимальную длину кодового слова, т. е. $l = \max l(v_i)$. Положим

$$U = \{u \in B^* \mid u = v_{i_1} v_{i_2} \dots v_{i_n}\}, \quad U_k = \{u \in U \mid l(u) = k\}.$$

Рассмотрим производящую функцию последовательности $\{|U_k|\}$. С одной строны,

$$\sum_{k=0}^{\infty} |U_k| x^k = \sum_{k=0}^{nl} |U_k| x^k,$$

с другой стороны,

$$\sum_{k=0}^{\infty} |U_k| x^k = \left(x^{l(v_1)} + x^{l(v_2)} + \ldots + x^{l(v_r)} \right)^n.$$

Поэтому, полагая $x = \frac{1}{q}$, имеем:

$$\left(\sum_{i=1}^{r} \frac{1}{q^{l(v_i)}} \right)^n = \sum_{k=0}^{nl} \frac{|U_k|}{q^k} \leq nl.$$

Таким образом для любого натурального n выполняется неравенство

$$\sum_{i=1}^{r} \frac{1}{q^{l(v_i)}} \leq \sqrt[n]{nl}.$$

Для завершения доказательства достаточно в последнем неравенстве перейти к пределу при $n \to \infty$. □

Теорема 2 (о построении префиксного кода с заданным набором длин)**.** *Пусть натуральные* $l_1, l_2, \dots, l_r$ *удовлетворяют неравенству*

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

Тогда существует префиксный код, задаваемый схемой $a_i \to v_i,\ v_i \in \{0, 1, \dots, q-1\}^*,\ i = 1, \dots, r$, *такой что* $l(v_i) = l_i,\ i = 1, \dots, r$.

Доказательство. Без ограничения общности будем считать, что

$$l_1 \leq l_2 \leq \ldots \leq l_r.$$

Положим

$$n_1 = 0, \quad n_k = \sum_{i=1}^{k-1} q^{-l_i},\ k = 2, \dots, r.$$

Тогда $0 = n_1 < n_2 < \ldots < n_r < 1$. При этом каждое число n_k $(k = 1, 2, \dots, r)$ является суммой чисел, у которых в записи по основанию q не более l_k разрядов после запятой. Кроме того все числа n_k меньше единицы. Поэтому число n_k в q-ичной системе счисления может быть записано следующим образом:

$$n_k = 0, b_{1k} b_{2k} \dots b_{l_k k}, \quad k = 1, 2, \dots, r,$$

где $b_{ij} \in \{0, 1, \dots, q-1\}$.

Теперь рассмотрим код из алфавита $\{a_1, \dots, a_r\}$ в алфавит $\{0, 1, \dots, q-1\}$, задаваемый схемой:

$$a_k \to b_{1k} b_{2k} \dots b_{l_k k} = v_k, \quad k = 1, 2, \dots, r.$$

Условие $l(v_k) = l_k$ выполняется. Покажем что указанная схема задает префиксный код. Пусть это не так, т. е. существуют такие s и k, что

$$v_s = v_k b_{l_k+1,s} \dots b_{l_s s}.$$

Но тогда $s > k$ (так как $l_s > l_k$) и справедливы соотношения

$$n_s = n_k + 0, \underbrace{00\dots00}_{l_k \text{ разрядов}} b_{l_k+1,s} \dots b_{l_s s} \le n_k + 0, \underbrace{00\dots00}_{l_k \text{ разрядов}} \underbrace{(q-1)\dots(q-1)}_{l_s - l_k \text{ разрядов}} <$$
$$< n_k + 0, \underbrace{00\dots01}_{l_k \text{ разрядов}} = n_k + q^{-l_k} = n_{k+1} \le n_s.$$

Получили противоречие ($n_s < n_s$), которое завершает доказательство. □

Следствие 1. Для любого разделимого кода существует префиксный код с таким же набором кодовых слов.

Следствие 2. Для существования разделимого кода с заданным набором $l_1, l_2, \dots, l_r$ длин кодовых слов необходимо и достаточно выполнение неравенства

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

1.1.1 Полные коды

Вспомним два примера разделимых кодов со стр. 2, задаваемых двумя разными схемами. В одном случае есть последовательности букв второго алфавита, которые никогда не могут встретиться (в примере 2 такой последовательностью, в частности, является последовательность из двух единиц), а в другом случае (пример 3) таких последовательностей нет.

Разделимый алфавитный код, задаваемый схемой

$$a_i \to v_i,\ v_i \in B^*,\ i = 1, \dots, r,$$

называется *полным*, если любое слово $v \in B^*$ представимо в виде

$$v = v_{i_1} \dots v_{i_s} \beta,$$

где v_{i_j}, $j = 1, \dots, s$, — кодовые слова, β — префикс некоторого кодового слова.

Свойство полных кодов:

Для любого слова $v \in B^*$ либо v — префикс некоторого кодового слова (соответствует случаю $s = 0$ в определении полного кода), либо некоторое кодовое слово — префикс слова v (соответствует случаю $s > 0$).

Теорема 3 (критерий полноты разделимого кода)**.** *Пусть алфавитный код V, задаваемый схемой*

$$a_i \to v_i,\ v_i \in B^*,\ i = 1, \dots, r,$$

является разделимым. Тогда для того, чтобы код V был полным, необходимо и достаточно, чтобы выполнялось следующее условие: для любого слова $u \in B^*$, удовлетворяющего неравенству $l(u) \ge \max l(v_i)$, найдется кодовое слово v_j, являющееся префиксом слова u.

Доказательство. Необходимость следует непосредственно из определения полного кода.

Достаточность. Установим существование требуемого определением полного кода представления произвольного слова $u \in B^*$.

Будем «отрезать» от слова u слева кодовое слово, пока это сделать можно (это делать точно можно пока длина слова не менее чем $\max l(v_i)$). В какой-то момент получаем слово $u' \in B^*$, никакой префикс которого не является кодовым словом. Очевидно, что $l(u') < \max l(v_i)$.

Если $u' = \Lambda$, то искомое представление получено.

Если $u' \neq \Lambda$, то к слову u' допишем справа произвольным образом $\max l(v_i) - l(u')$ символов из множества B. Полученное слово u'' имеет длину $\max l(v_i)$ и, следовательно, по условию имеет в качестве префикса некоторое кодовое слово v_j. С другой стороны, по построению слово u'' имеет префикс u'. Таким образом слова v_j и u' являются префиксами одного и того же слова и следовательно одно из них является префиксом другого. Но слово u' по построению не имеет префикса в виде кодового слова. Поэтому u' — префикс кодового слова v_j. Следовательно искомое представление получено. □

Теорема 4 (критерий полноты кода)**.** *Пусть алфавитный код V, задан схемой*

$$a_i \to v_i,\ v_i \in B^*,\ i = 1, \ldots, r.$$

Тогда для того, чтобы код V был полным, необходимо и достаточно, чтобы код V был префиксным и выполнялось условие

$$\sum_{i=1}^{r} \frac{1}{q^{l(v_i)}} = 1,$$

где $q = |B|$.

Доказательство. Необходимость. Код V полный, и следовательно разделимый. Поэтому выполняется неравенство Крафта—МакМиллана

$$\sum_{i=1}^{r} \frac{1}{q^{l(v_i)}} \leq 1.$$

Пусть $l = \max l(v_i)$. Обозначим через U_i множество слов из B^l (длины l) с префиксом v_i:

$$U_i = \{u \in B^l \mid u = v_i v,\ v \in B^*\}, \quad i = 1, \ldots, r.$$

В силу полноты кода выполняется равенство $B^l = \bigcup\limits_{i=1}^{r} U_i$. Поэтому

$$q^l = |B_l| = \left|\bigcup_{i=1}^{r} U_i\right| \leq \sum_{i=1}^{r} |U_i| = \sum_{i=1}^{r} q^{l-l(v_i)} = q^l \sum_{i=1}^{r} q^{-l(v_i)} \leq q^l.$$

Следовательно, содержащиеся в этой цепочке два нестрогих неравенства на самом деле являются равенствами. Второе из них устанавливает равенство в неравенстве Крафта—МакМиллана. А равенство

$$\left|\bigcup_{i=1}^{r} U_i\right| = \sum_{i=1}^{r} |U_i|$$

означает, что множества U_i, $i = 1, \ldots, r$, не пересекаются. Следовательно, алфавитный код V — префиксный.

Достаточность. Код V префиксный и поэтому разделимый. Покажем, что для любого слова $u \in B^*$, удовлетворяющего неравенству $l(u) \geq l$, найдется кодовое слово v_j, являющееся префиксом слова u.

Пусть $l(u) = n$. Положим

$$U_i = \{u \in B^n \mid u = v_i v,\ v \in B^*\}, \quad i = 1, \ldots, r.$$

Код V — префиксный, поэтому $U_i \cap U_j =$ при $i \neq j$. Тогда

$$\left|\bigcup_{i=1}^{r} U_i\right| = \sum_{i=1}^{r} |U_i| = \sum_{i=1}^{r} q^{n-l(v_i)} = q^n \sum_{i=1}^{r} q^{-l(v_i)} = q^n = |B^n|.$$

Следовательно существует такое i, что $u \in U_i$. По предыдущей теореме код V — полный. □

Теорема 5. *Пусть натуральные $l_1, \ldots, l_r$ удовлетворяют условию*

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

Тогда существует полный код со схемой

$$a_i \to v_i,\ v_i \in \{0,1\}^*,\ i = 1, \ldots, r,$$

такой что выполняются неравенства $l(v_i) \leq l_i$, $i = 1, \ldots, r$.

Доказательство. Строим префиксный двоичный код с длинами кодовых слов $l_1, \ldots, l_r$ — согласно теореме 2 такой код существует. Теперь будем переделывать этот код сохраняя свойство префиксности.

Если $\sum\limits_{i=1}^{r} 2^{-l_i} = 1$, то по теореме 4 код полный.

Пусть $\sum\limits_{i=1}^{r} 2^{-l_i} < 1$. Без ограничения общности будем считать, что $l_1 \leq l_2 \leq \ldots \leq l_r$. Тогда

$$\sum_{i=1}^{r} 2^{-l_i} + 2^{-l_r} \leq 1.$$

Теперь, полагая $l'_1 = l_1, \ldots, l'_{r-1} = l_{r-1}$, $l'_r = l_r - 1$, получаем:

$$\sum_{i=1}^{r} 2^{-l'_i} = \sum_{i=1}^{r-1} 2^{-l_i} + 2^{-l'_r} = \sum_{i=1}^{r-1} 2^{-l_i} + 2^{-l_r} + 2^{-l_r} = \sum_{i=1}^{r} 2^{-l_i} + 2^{-l_r} \leq 1.$$

Последовательно уменьшая в наборе длин наибольшее значение на единицу, за конечное число шагов получим набор, для которого соответствующая сумма обращается в единицу. □

1.1.2 Оптимальное кодирование

Пусть $A = \{a_1, \ldots, a_r\}$, $B = \{b_1, \ldots, b_q\}$ — алфавиты, алфавитное кодирование задано схемой Σ:

$$a_1 \to v_1, \ldots, a_r \to v_r; \quad v_i \in B^*,\ i = 1, \ldots, r.$$

Пусть $V = \{v_1, \ldots, v_r\}$ — алфавитный код, заданный схемой Σ.

Будем считать, что заданы частоты появления символов из алфавита A и как следствие — их вероятности $p_i = p(a_i)$, удовлетворяющие условиям $p_i > 0,\ i = 1, \ldots, r$, $p_1 + \ldots p_r = 1$.

Набор $P = (p_1, \ldots, p_r)$ будем называть *распределением*.

Далее везде считаем, что алфавитный код V — разделимый.

Определим *стоимость* $L_V(P)$ алфавитного кода V при распределении P равенством

$$L_V(P) = \sum_{i=1}^{r} p_i l(v_i).$$

Положим

$$L(P) = \inf L_V(P),$$

где инфинум берется по всем разделимым кодам V. Если справедливо равенство $L_P(V) = L(P)$, то код V называется *оптимальным кодом* (или *кодом с минимальной избыточностью*).

Свойства оптимальных кодов:

1. Для любого распределения существует оптимальный код.

▷ Построим равномерный код $V = \{v_1, \ldots, v_r\}$. Тогда $l(v_i) = \lceil \log_q r \rceil$, $i = 1, \ldots, r$. Обозначим $p_{\min} = \min\{p_1, \ldots, p_r\}$. Любой код $V' = \{v'_1, \ldots, v'_r\}$, в котором найдется кодовое слово v'_i, удовлетворяющее неравенству

$$l(v'_i) > \frac{\lceil \log_q r \rceil}{p_{\min}},$$

не является оптимальным:

$$L_{V'}(P) > p_i \frac{\lceil \log_q r \rceil}{p_{\min}} \geq \lceil \log_q r \rceil = \sum_{i=1}^{r} p_i l(v_i) = L_V(P).$$

Таким образом, достаточно взять инфинум по конечному множеству кодов. □

2. Для любого распределения существует оптимальный префиксный код.

3. В оптимальном коде:

а) если $p_i < p_j$, то $l(v_i) \geq l(v_j)$;

б) если $l(v_i) < l(v_j)$, то $p_i \geq p_j$.

▷ Пусть это не так, т. е. $p_i < p_j$ и $l(v_i) < l(v_j)$. Тогда ввиду неравенства

$$(p_j - p_i)(l(v_j) - l(v_i)) > 0$$

выполняется соотношение

$$p_i l(v_i) + p_j l(v_j) > p_i l(v_j) + p_j l(v_i).$$

Поэтому, поменяв местами кодовые слова v_i и v_j, получим код меньшей стоимости — противоречие. □

4. Справедлива следующая нижняя оценка стоимости оптимального кода:

$$L(P) \geq \sum_{i=1}^{r} p_i \log_q \frac{1}{p_i}.$$

▷ Действительно,

$$\begin{aligned} -L(P) + \sum_{i=1}^{r} p_i \log_q \frac{1}{p_i} &= \sum_{i=1}^{r} \left(-l(v_i) p_i + p_i \log_q \frac{1}{p_i} \right) = \\ &= \sum_{i=1}^{r} p_i \log_q \frac{q^{-l(v_i)}}{p_i} = \sum_{i=1}^{r} p_i \frac{1}{\ln q} \ln \frac{q^{-l(v_i)}}{p_i} \leq \\ &\leq \sum_{i=1}^{r} p_i \frac{1}{\ln q} \left(\frac{q^{-l(v_i)}}{p_i} - 1 \right) = \frac{1}{\ln q} \left(\sum_{i=1}^{r} q^{-l(v_i)} - \sum_{i=1}^{r} p_i \right) \leq 0. \quad \square \end{aligned}$$

5. Построение кода, близкого к оптимальному, методом Шеннона.

Пусть $P = (p_1, \ldots, p_r)$, $p_i > 0$, $p_1 + \ldots + p_r = 1$, $q \geq 2$. Полагаем $l_i = \lceil \log_q \frac{1}{p_i} \rceil$. Тогда $-l_i \leq \log_q p_i$ и следовательно $q^{-l_i} \leq p_i$. Значит,

$$\sum_{i=1}^{r} q^{-l_i} \leq \sum_{i=1}^{r} p_i = 1.$$

При доказательстве теоремы 2 предъявлен алгоритм построения префиксного кода V с длинами кодовых слов, удовлетворяющих неравенству Крафта—МакМиллана. При этом

$$L_V(P) = \sum_{i=1}^{r} p_i \left\lceil \log_q \frac{1}{p_i} \right\rceil \leq \sum_{i=1}^{r} p_i \log_q \frac{1}{p_i} + \sum_{i=1}^{r} p_i = \sum_{i=1}^{r} p_i \log_q \frac{1}{p_i} + 1.$$

6. Величину $L(P)$ можно установить с точностью до единицы:

$$\sum_{i=1}^{r} p_i \log_q \frac{1}{p_i} \le L(P) \le \sum_{i=1}^{r} p_i \log_q \frac{1}{p_i} + 1.$$

7. Если все числа $\log_q \frac{1}{p_i}$ — натуральные, то код, построенный методом Шеннона, — оптимальный.

8. Пусть $V = \{v_1, \dots, v_r\}$ — полный код. Тогда найдется распределение $P = \{p_1, \dots, p_r\}$, для которого выполняется равенство $L_V(P) = L(P)$, т. е. при этом распределении код V — оптимальный.

▷ Положим $p_i = q^{-l(v_i)}$, $i = 1, \dots, r$. Тогда $p_1 + \dots p_r = 1$. Кроме того,

$$L_V(P) = \sum_{i=1}^{r} p_i l(v_i) = \sum_{i=1}^{r} p_i \log_q q^{l(v_i)} = \sum_{i=1}^{r} p_i \log_q \frac{1}{p_i} \le_{\text{(по свойству 4)}} L(P).$$

Следовательно, код V оптимальный. □

Задание 1. Доказать, что двоичный ($q = 2$) префиксный оптимальный (для некоторого распределения) код является полным.

Указание. Если в неравенстве Крафта—МакМиллана строгое неравенство, то можно построить код с меньшим набором длин — противоречие с оптимальностью.

1.1.3 Построение оптимального кода

1. Можно искать оптимальный код в классе префиксных кодов.

2. Префиксный код задается кодовым деревом — корневым помеченным деревом, ребрам которого приписаны буквы из алфавита B, а концевым вершинам соответствуют буквы алфавита A. Припишем концевым вершинам вероятности соответствующих им букв алфавита A.

3. Остальным вершинам также припишем вероятности как суммы вероятностей всех концевых вершин поддерева с корнем в этой вершине.

4. Все вершины разобъем на ярусы, в зависимости от расстояния до корня. Корень — вершина нулевого уровня.

Свойства кодовых деревьев оптимальных префиксных кодов:

1. Если вероятность, приписанная вершине w_1, меньше вероятности, приписанной вершине w_2, то номер яруса вершины w_1 не меньше номера яруса вершины w_2.

2. Все «пучки» ребер, исходящие не из предпоследнего яруса, либо «насыщенные» (т. е. состоят из q ребер), либо пустые.

Без ограничения общности можно рассматривать приведенное кодовое дерево — такое дерево, в котором все «пучки» ребер, кроме, быть может, одного (исходящего из некоторой вершины предпоследнего яруса), — насыщенные, при этом число q_0 ребер в этом особом пучке определяется равенством

$$q_0 = \begin{cases} 2, & \text{если } q = 2; \\ q-1, & \text{если } r \equiv 0 \pmod{q-1}; \\ q, & \text{если } r \equiv 1 \pmod{q-1}; \\ r - (q-1)\lfloor \frac{r}{q-1} \rfloor, & \text{иначе,} \end{cases}$$

и эти ребра ведут в концевые вершины, которым приписаны q_0 самых маленьких вероятностей из исходного распределения.

Теорема 6 (о редукции). *Пусть есть два кодовых дерева, причем второе дерево получается из первого путем замены концевой вершины на пучок из s ребер (при этом распределение P_2 получается из распределения P_1 соответствующей заменой p на $p_{i_1}, \dots, p_{i_s}$). Тогда:*

1. *Если второй код — оптимальный, то первый — тоже оптимальный.*

2. *Если выполняются следующие условия: а) первый код — оптимальный; б) вероятности* $p_{i_1}, \ldots, p_{i_s}$ *— это* s *самых маленьких вероятностей в распределении* P_2*; в)* $s = q_0$*; то второй код — тоже оптимальный.*

Доказательство. Очевидно, что справедливо равенство $L_{V_1}(P_1) + p = L_{V_2}(P_2)$.

1. Предположим, что первый код не будет оптимальным. В этом случае найдется такой код V_1', что $L_{V_1'}(P_1) < L_{V_1}(P_1)$. Делая аналогичную замену вершины на пучок ребер в кодовом дереве кода V_1', получаем кодовое дерево некоторого кода V_2', для которого

$$L_{V_2'}(P_2) = L_{V_1'}(P_1) + p < L_{V_1}(P_1) + p = L_{V_2}(P_2),$$

что противоречит оптимальности кода V_2.

2. Предположим, что второй код не будет оптимальным. Тогда существует оптимальный код V_1'' с приведенным кодовым деревом, в котором есть соответствующий пучок из s ребер. Заменяя этот пучок на концевую вершину, получаем кодовое дерево некоторого кода V_1'', для которого

$$L_{V_1''}(P_1) = L_{V_2''}(P_2) - p < L_{V_2}(P_2) - p = L_{V_1}(P_1),$$

что противоречит оптимальности кода V_1. □

На использовании теоремы о редукции основан **метод Хаффмана** построения оптимального префиксного кода, заключающегося в последовательной замене q (на первом шаге q_0) букв, имеющих наименьшие вероятности, на одну новую, имеющую вероятность, равную сумме вероятностей заменяемых букв.

Пример 3. Построение оптимального кода методом Хаффмана в случае, когда $|A| = 11$, $|B| = 4$,

$$P = (0,25, 0, 24, 0, 15, 0, 08, 0, 07, 0, 06, 0, 05, 0, 04, 0, 03, 0, 02, 0, 01).$$

Сначала определяем, что $q_0 = 2$. Затем четырежды проделываем следующие процедуры: выписываем все вероятности в порядке убывания, заменяем 4 (в первый раз 2) самые маленькие вероятности на их сумму и переходим к новому распределению. Проделав это, последовательно в обратном порядке строим корневое дерево, заменяя вершину на пучок из 4 ребер (в последний раз — на пучок из 2 ребер), как показано на рис. 2.

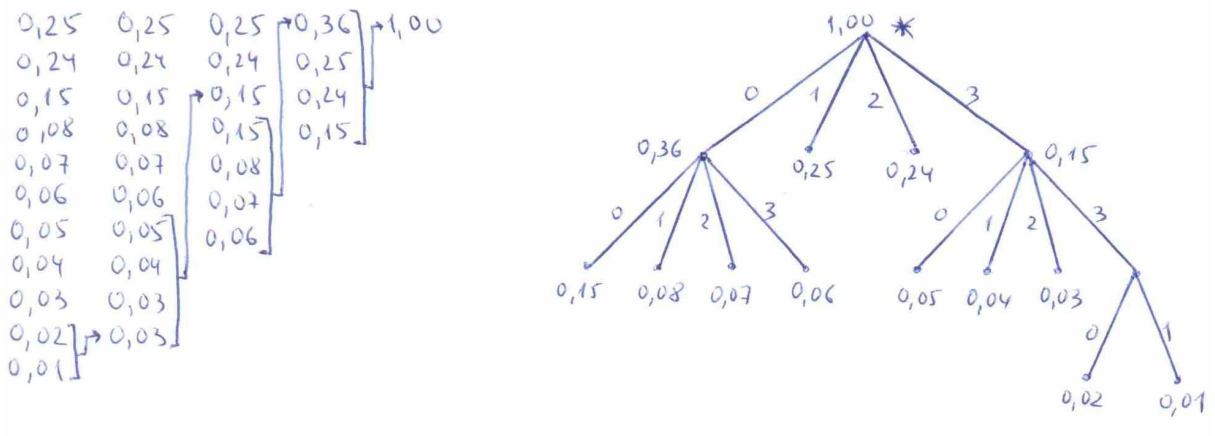

Рис. 2:

Пусть $p(a_1) = 0,25$, $p(a_2) = 0,24$, $p(a_3) = 0,15$, $p(a_4) = 0,08$, $p(a_5) = 0,07$, $p(a_6) = 0,06$, $p(a_7) = 0,05$, $p(a_8) = 0,04$, $p(a_9) = 0,03$, $p(a_{10}) = 0,02$, $p(a_{11}) = 0,01$. Тогда построенное кодовое дерево

определяет такую схему кодирования:

$$a_1 \to 1, \quad a_2 \to 2, \quad a_3 \to 00, \quad a_4 \to 01,$$
$$a_5 \to 02, \quad a_6 \to 03, \quad a_7 \to 30, \quad a_8 \to 31,$$
$$a_9 \to 32, \quad a_{10} \to 330, \quad a_{11} \to 331.$$

Отметим, что если при третьем упорядочивании вероятностей поменять местами две одинаковые вероятности, равные 0,15, то мы построим оптимальный код, в котором будет три кодовых слова длины 1 и два кодовых слова длины 4.

Задание 2. При алфавитном кодировании четырехбуквенного алфавита с распределением $P = (p_1, p_2, p_3, p_4)$, $p_1 \geq p_2 \geq p_3 \geq p_4 > 0$, в двухбуквенный существует оптимальный код, содержащий кодовое слово длины 3, и существует оптимальный код, в котором нет кодовых слов длины 3. Найти все возможные значения, которые может принимать величина

а) p_1,
б) p_2,
в) p_3,
г) p_4.

Для каждого такого значения привести пример набора вероятностей, на котором это значение достигается.

1.2 Коды, исправляющие ошибки

Пусть требуется по зашумленному каналу связи передать некоторое сообщение, т. е. конечный набор символов фиксированного алфавита (на самом деле, как правило, передается не само сообщение, а код этого сообщения, что соответствует схеме на рис. 1). Зашумленность подразумевает возможность искажения передаваемой информации:

$$x_1 \dots x_n \ \to \ \boxed{\text{КАНАЛ СВЯЗИ}} \ \to \ y_1 \dots y_m.$$

Обозначим через $R(\tilde{\alpha})$ множество слов, в которое может перейти сообщение $\tilde{\alpha}$ под воздействием «шума».

Если для любых сообщений $\tilde{\alpha}$ и $\tilde{\beta}$ справедливо соотношение $R(\tilde{\alpha}) \cap R(\tilde{\beta}) = \varnothing$, то код называется *самокорректирующимся* относительно заданного источника ошибок.

Далее будем предполагать, что возможны только ошибки типа замещения, а ошибки типа вставки или выпадения символа отсутсвуют. Отметим, что ошибки типа замещения являются наиболее характерными ошибками, особенно при хранении («передаче во времени») информации.

Теперь под кодом будем понимать множество слов одинаковой длины.

Пусть: p — простое, $B = \mathbb{Z}_p$ — поле вычетов по модулю p, $B^n = \{(\alpha_1, \dots, \alpha_n) \mid \alpha_i \in B\}$.

Введем обозначения: $\rho(\tilde{\alpha}, \tilde{\beta})$ — *расстояние Хэмминга между наборами* $\tilde{\alpha}$ и $\tilde{\beta}$, т. е. число несовпадающих разрядов; $S_n^t(\tilde{\alpha}) = \{\tilde{\beta} \in B^n \mid \rho(\tilde{\alpha}, \tilde{\beta}) \leq t\}$ — n-мерный шар радиуса t с центром в $\tilde{\alpha}$.

Если в канале происходит не более t ошибок, то $R(\tilde{\alpha}) = S_n^t(\tilde{\alpha})$.

Код $V = \{\tilde{\alpha}^1, \dots, \tilde{\alpha}^N\}$ *исправляет* t ошибок типа замещения, если для любых различных i и j выполняется неравенство $\rho(\tilde{\alpha}^i, \tilde{\alpha}^j) \geq 2t + 1$.

Код $V = \{\tilde{\alpha}^1, \dots, \tilde{\alpha}^N\}$ *обнаруживает* t ошибок типа замещения, если для любых различных i и j выполняется неравенство $\rho(\tilde{\alpha}^i, \tilde{\alpha}^j) \geq t + 1$.

Положим $d(V) = \min \rho(\tilde{\alpha}, \tilde{\beta})$, где минимум берется по всем парам $(\tilde{\alpha}, \tilde{\beta})$ различных наборов из кода V. Величина $d(V)$ называется *кодовым расстоянием кода* V.

Код V обнаруживает $d(V) - 1$ ошибку и исправляет $\lfloor (d(V) - 1)/2 \rfloor$ ошибок.

Положим $N_n^t = |S_n^t(\tilde{\alpha})|$.

Теорема 7 (Граница Хэмминга или граница сферической упаковки)**.** *Пусть* V_n^t *— код из* B^n, *исправляющий* t *ошибок* (*типа замещения*). *Тогда*

$$|V_n^t| \leq \frac{p^n}{N_n^t}.$$

Доказательство. Из неравенства

$$|V_n^t|\ N_n^t \le p^n$$

непосредственно следует утверждение теоремы. □

Если в границе Хэмминга равенство, то такой код называется *совершенным*.

Пример 4. $p = 2$, $n = 2k + 1$, $t = k$, $V = \{(0, \dots, 0), (1, \dots, 1)\}$.

Метод построения кода, исправляющего t ошибок

В качестве набора $\tilde{\alpha}^1$ берем произвольный набор, например, $\tilde{0}$.
Если наборы $\tilde{\alpha}^1, \dots, \tilde{\alpha}^{k-1}$ уже построены, то набор $\tilde{\alpha}^k$ определяем из условия

$$\tilde{\alpha}^k \notin \bigcup_{i=1}^{k-1} S_n^{2t}(\tilde{\alpha}^i)$$

пока это можно сделать.

Получаем код V, исправляющий t ошибок. По построению

$$B^n \subseteq \bigcup_{i=1}^{|V|} S_n^{2t}(\tilde{\alpha}^i).$$

Поэтому

$$|V| \ge \frac{p^n}{N_n^{2t}}.$$

Положим $M_n^t(p) = \max |V_n^t|$, где максимум берется по всем кодам V_n^t, $V_n^t \subset B^n$, исправляющим t ошибок. Тогда

$$\frac{p^n}{N_n^{2t}} \le M_n^t(p) \le \frac{p^n}{N_n^t}.$$

Оценим число наборов в шаре радиуса t:

$$N_n^t = \sum_{k=0}^{t} C_n^k (p-1)^k,$$

$$n^k \frac{1}{k^k} \le C_n^k = \frac{n(n-1)\dots(n-k+1)}{k(k-1)\dots(k-k+1)} \le \frac{n^k}{k!} \le n^k \frac{3^k}{k^k}.$$

Таким образом,

$$c'(t) n^t \le N_n^t \le c''(t) n^t,$$

и следовательно

$$c_1(t) \frac{p^n}{n^{2t}} \le M_n^t(p) \le c_2(t) \frac{p^n}{n^t}.$$

Содержательно это означает, что на исправление одной ошибки требуется «заложить» порядка $\log n$ разрядов.

Отметим, что

$$M_n^1(p) \le \frac{p^n}{1 + (p-1)n}, \quad M_n^1(2) \le \frac{2^n}{1+n}.$$

1.3 Метод Хэмминга построения кодов, исправляющих одну ошибку

Закодируем слово $\widetilde{\alpha} = (\alpha_1, \ldots, \alpha_m)$ словом $\widetilde{\beta} = (\beta_1, \ldots, \beta_n)$. Длина слова $\widetilde{\beta}$ определяется как минимальное число l, удовлетворяющее неравенству $2^m \leq \frac{2^l}{l+1}$, то есть $n = \min\left\{l : 2^m \leq \frac{2^l}{l+1}\right\}$. В слове $\widetilde{\beta}$ есть все разряды слова $\widetilde{\alpha}$ и ещё на местах 1, 2, 4, 8, ... (т.е. на местах с номерами 2^i) проверочные разряды p_0, p_1, ..., p_{k-1}, $k = n - m$.

Определим проверочные разряды p_i следующим образом. $p_i = \beta_{2^i+1} \oplus \beta_{2^i+2} \oplus \ldots$, где сумма берётся по всем β_j $(2^i < j \leq n)$, у которых в разложении индекса j коэффициент при 2^i равен 1.

Пример 5. $\widetilde{\alpha} = (1011)$, $m = 4$, $n = \min\{l : 2^m \leq \frac{2^l}{l+1}\} = 7$. Построим кодовое слово $\widetilde{\beta} = \beta_1\beta_2\beta_3\beta_4\beta_5\beta_6\beta_7 = p_0p_11p_2001$.

0	**0**	1	**1**	0	0	1	
x		x		x		x	$p_0 = \beta_1 = 1 \oplus 0 \oplus 1 = 0$
	x	x			x	x	$p_1 = \beta_2 = 1 \oplus 0 \oplus 1 = 0$
			x	x	x	x	$p_2 = \beta_4 = 0 \oplus 0 \oplus 1 = 1$

Декодирование.

Пусть $m = \left\lfloor \log_2 \left(\frac{2^m}{(n+1)} \right) \right\rfloor$, $k = n - m$.

По полученному вектору $\widetilde{\beta} = (\beta_1, \ldots, \beta_n)$ ищем k сумм вида $v_i = \beta_{2^i} \oplus \beta_{2^i+1} \oplus \ldots$, $i = 0, \ldots, k-1$, где в $i-$ю сумму входят все β_j, $2^i \leq j \leq n$, у которых двоичное разложение индекса j коэффициент при 2^i равный 1. Номер разряда ошибки

$$\sum_{0 \leq i < k} v_i 2^i.$$

Пример 6. Декодировать $\widetilde{\beta} = (1001110)$, $n - 7$, $m = 4$, $k = 3$.

$$v_0 = \beta_1 \oplus \beta_3 \oplus \beta_5 \oplus \beta_7 = 1 \oplus 0 \oplus 1 \oplus 0 = 0,$$
$$v_1 = \beta_2 \oplus \beta_3 \oplus \beta_6 \oplus \beta_7 = 0 \oplus 0 \oplus 1 \oplus 0 = 1,$$
$$v_2 = \beta_4 \oplus \beta_5 \oplus \beta_6 \oplus \beta_7 = 1 \oplus 1 \oplus 1 \oplus 0 = 1.$$

Номер разряда ошибки $110_2 = 6_{10}$. Следовательно, $\widetilde{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (\beta_3, \beta_5, \overline{\beta_6}, \beta_7) = (0100)$.