УДК 519.723.6

# КОДИРОВАНИЕ НЕРАВНОЗНАЧНЫМИ СИМВОЛАМИ ИСТОЧНИКОВ МУРА И МИЛИ ПРИ НЕИЗВЕСТНОЙ СТАТИСТИКЕ СООБЩЕНИЙ

© **В. К. Трофимов**[1,2], **Т. В. Храмова**[1]

[1] *Сибирский государственный университет телекоммуникаций и информатики, 630102, г. Новосибирск, ул. Кирова, 86*
[2] *Институт систем информатики им. А. П. Ершова СО РАН, 630090, г. Новосибирск, просп. Академика Лаврентьева, 6*
*E-mail: trofimov@sibguti.ru*
*tvkhramova@gmail.com*

Найдена избыточность универсального кодирования неравнозначными символами марковских источников, задаваемых матрицами переходных вероятностей, имеющих фиксированное число различных строк. В качестве следствия получены оценки избыточности для марковских источников с памятью $s$ и марковских источников Мили, заданных графом. Установлена скорость убывания избыточности в зависимости от характеристик графа, длины кодируемого блока и пропускной способности канала.

*Ключевые слова:* энтропия, кодирование, избыточность кодирования, источник сообщений, пропускная способность.

DOI: 10.15372/AUT20210207

**Введение.** В основополагающей работе Шеннона [1] были заложены два фундаментальных раздела, позволяющих либо уменьшить объём передаваемой информации, либо с помощью увеличения объёма передаваемой информации исключить появление возможных ошибок.

В представленной работе сосредоточено внимание на сокращении объёма информации с использованием знания о том, какому множеству принадлежит источник, порождающий информацию.

Рассмотрим источник $\theta$, порождающий полубесконечную последовательность букв конечного алфавита $A = \{a_1, \ldots, a_k\}$. Далее алфавит $A$ будем называть входным. Как обычно, тип вероятностной меры, заданной на последовательности порождаемых букв, определяет тип источника. Основными являются бернуллиевские, пуассоновские и марковские источники. Известно, по крайней мере, два типа марковских источников. В одном из них вероятность появления очередной буквы зависит от конечного фиксированного числа предыдущих букв. Эти источники называются источниками с памятью. В других случаях слова порождаются стохастическим автоматом [2]. Таким образом, появление каждой порождаемой источником буквы может однозначно определяться либо состоянием (автомат Мура), либо переходом из одного состояния в другое (автомат Мили) [3].

Целью данной работы является построение оптимальных универсальных кодов для специального класса источников Мура и с их помощью образование оптимальных универсальных кодов для множества источников, заданных автоматом Мили. В той или другой степени неизвестные источники изучаются при оптимальном поиске импульсно-точечных источников [4], при округлении распределения статистик [5].

**1. Основные определения и обозначения.** Полубесконечная последовательность букв входного алфавита $A$ разбивается на блоки $w$, каждый из которых содержит $N$ букв входного алфавита, т. е. $w \in A^N$. Число букв в блоке (слове) $w$ будем называть его длиной.

Процедура кодирования последовательности букв, порождённой источником $\theta$, состоит в том, что каждому блоку $w$, $w \in A^N$, ставится в соответствие слово $\varphi(w)$ в алфавите $B = \{b_1, b_2, \ldots, b_m\}$.

При этом рассмотрим только дешифруемые кодирования, которые позволяют по принятой последовательности однозначно восстановить переданную. В данной работе будем использовать выходной алфавит $B = \{b_1, \ldots, b_m\}$, в котором буквы неравнозначны, т. е. каждая из букв алфавита $B$ имеет свою длительность $t_j = t(b_j)$, $j = \overline{1, m}$. Например, при кодировании букв кодом Морзе каждой букве ставится в соответствие последовательность из точек и тире, причём длительность тире в три раза больше, чем длительность точки. Таким образом, код Морзе является примером кодирования букв символами различной длительности.

На основании вышесказанного можно сделать вывод, что каждому выходному алфавиту $B = \{b_1, b_2, \ldots, b_m\}$ соответствует целочисленный вектор $\mathbf{t}(B) = (t_1, t_2, \ldots, t_m)$, $t_j = t(b_j)$, $j = \overline{1, m}$. В случае если буквы алфавита $B$ равнозначны, будем считать, что $\mathbf{t}(B) = \mathbf{t}_1 = (1, 1, \ldots, 1)$.

Как показано в [1], по каждому выходному алфавиту $B$ вычисляется пропускная способность канала $c(B) = \log \omega_0$, где $\omega_0$ — наибольший корень уравнения

$$\omega^{-t_1} + \omega^{-t_2} + \ldots + \omega^{-t_m} = 1. \tag{1}$$

Как обычно, $\log x = \log_2 x$.

В частности, как видно из (1), для выходного алфавита $B$, у которого $\mathbf{t}(B) = \mathbf{t}_1$, пропускная способность $c(B) = \log m$. Для двоичного алфавита $B$, у которого $\mathbf{t}(B) = (1, 2)$, несложно вычислить, что $c(B) = \log\left((1 + \sqrt{5}\,)/2\right)$.

Длительность кодового слова $\varphi(w)$ при кодировании буквами алфавита $B$ обозначим $l(\varphi(w), B)$. По определению

$$l(\varphi(w), B) = \sum_{b \in \varphi(w)} t(b). \tag{2}$$

При $\mathbf{t}(B) = \mathbf{t}_1$ величина $l(\varphi(w), B)$ равна числу букв в слове $\varphi(w)$, т. е. $l(\varphi(w), B) = |\varphi(w)|$.

Пусть $P_\theta(w)$ — вероятность порождения блока $w$ источником $\theta$.

Среднюю длительность букв выходного алфавита, приходящуюся на одну букву входного алфавита, назовём стоимостью кодирования и обозначим $L(N, \theta, \varphi, B)$. Учитывая (2), имеем

$$L(N, \theta, \varphi, B) = \frac{1}{N} \sum_{w \in A^N} P_\theta(w) l(\varphi(w), B). \tag{3}$$

Пусть $H(\theta)$ — энтропия источника $\theta$ [1, 2]. По определению

$$H(\theta) = -\lim_{N \to \infty} \frac{1}{N} \sum_{w \in A^N} P_\theta(w) \log P_\theta(w). \tag{4}$$

Если источник бернуллиевский и задаётся вектором $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, то

$$H(\theta) = -\sum_{i=1}^{k} P(\theta_i) \log P(\theta_i)$$

(здесь и далее $0 \log 0 = 0$).

Для марковского источника с памятью $s$, определяемого переходными вероятностями $P(a_j \mid a_\alpha) = P_{j\alpha}$, где $\alpha = (j_1, \ldots, j_s)$, а $a_\alpha = (a_{j_1}, \ldots, a_{j_s})$, и финальными вероятностями $P_{0\alpha}$,

$$H(\theta) = -\sum_{\alpha \in A^s} P_{0\alpha} \sum_{j=1}^{m} P_{j\alpha} \log P_{j\alpha}.$$

Разность между стоимостью кодирования $L(N, \theta, \varphi, B)$, определяемой равенством (3), и величиной $H(\theta)/c(B)$, где $H(\theta)$ задано соотношением (4), обозначим $R(N, \theta, \varphi, B)$ и назовём избыточностью кодирования блоков длины $N$ источником $\theta$ при кодировании $\varphi$ с применением выходного алфавита $B$:

$$R(N, \theta, \varphi, B) = L(N, \theta, \varphi, B) - \frac{1}{c(B)} H(\theta). \tag{5}$$

**2. Универсальное кодирование марковских источников.** Для формулировки и доказательства основных утверждений предлагаемой работы введём ещё некоторые понятия и сформулируем полученные ранее результаты.

Как обычно, обозначим через $\Omega_s$ множество марковских источников с памятью $s$, т. е. если источник $\theta \in \Omega_s$, то вероятность порождения им очередной буквы входного алфавита зависит от $s$ предшествующих. Избыточностью кода $\varphi$ на множестве источников $\Omega \subseteq \Omega_s$ назовём величину

$$R(N, \Omega, \varphi, B) = \sup_{\theta \in \Omega} R(N, \theta, \varphi, B). \tag{6}$$

Нижнюю грань величины $R(N, \Omega, \varphi, B)$ (6) по всевозможным дешифруемым кодированиям $\varphi$ назовём избыточностью универсального кодирования на множестве $\Omega$ при заданном выходном алфавите $B$ и обозначим $R(N, \Omega, B)$. Таким образом,

$$R(N, \Omega, B) = \inf_{\varphi} R(N, \Omega, \varphi, B). \tag{7}$$

В случае если символы выходного алфавита равнозначны, избыточность $R(N, \Omega, B)$ фактически не зависит от $B$. Как и прежде, будем использовать для избыточности универсального кодирования множества источников $\Omega$ обозначение $R(N, \Omega)$. Первые результаты по оценке $R(N, \Omega_s)$ были получены в [6], для бернуллиевских источников асимптотическое поведение $R(N, \Omega_0)$ полностью изучено в [7]. Асимптотическое поведение $R(N, \Omega_s)$ установлено в [8]. Следует отметить, что верхняя оценка для $R(N, \Omega_s)$ получена в [9]. В [10–13] приведён краткий обзор по универсальному кодированию. Кодирование источников Мили и Мура для равнозначных букв выходного алфавита изучалось в [14]. Поведению величины $R(N, \Omega_s, B)$, $s \geqslant 0$, посвящены работы [15–17].

С использованием ранее предложенных методов кодирования в данных обстоятельствах получены совпадающие по порядку убывания верхние и нижние оценки этих величин.

**3. Источники Мура.** Как уже отмечалось, для определения марковского источника достаточно задать вероятностную меру на последовательности, порождаемой источником. Таким образом, для задания источника достаточно найти матрицу вероятностей переходов и финальный вектор распределения вероятностей. Для того чтобы описать источник Мура, представим эргодическую стационарную марковскую цепь, заданную матрицей вероятностей переходов $\theta = \|\theta_{ij}\|$, $i, j = \overline{1, k}$, и начальным вектором распределения вероятностей $\boldsymbol{\theta}_0 = (\theta_{01}, \ldots, \theta_{0k})$.

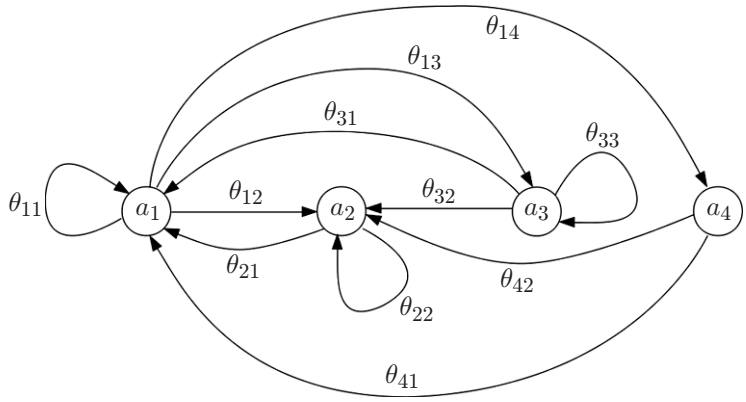


*Рис. 1*

Рассмотрим автомат Мура, имеющий $k$, $k > 0$, состояний. В состоянии $i$ этот автомат порождает букву $a_i$ входного алфавита $A$ и переходит в состояние $j$ с вероятностью $\theta_{ij}$, $i, j = \overline{1,k}$. Работа автомата начинается из состояния $i$, $i = \overline{1,k}$, в соответствии с начальным распределением вероятностей $\boldsymbol{\theta}_0$. Порождаемую автоматом последовательность букв разобьём на слова (блоки) длины $N$. Обозначим $P_\theta(u)$ — вероятность слова $u$, порождённого автоматом Мура с матрицей переходных вероятностей $\theta$.

**Пример.** Пусть

$$A = \{a_1, a_2, a_3, a_4\}, \qquad \theta = \left\| \begin{array}{cccc} \theta_{11} & \theta_{12} & \theta_{13} & \theta_{14} \\ \theta_{21} & \theta_{22} & 0 & 0 \\ \theta_{31} & \theta_{32} & \theta_{33} & 0 \\ \theta_{41} & \theta_{42} & 0 & 0 \end{array} \right\|.$$

Соответствующий этому источнику автомат Мура изображён на рис. 1.

Зафиксируем неотрицательные целые числа $l$, $m_1, m_2, \ldots, m_l$ и $t_1, t_2, \ldots, t_l$ такие, что

$$m_1 + m_2 + \ldots + m_l = k, \qquad t_i \leqslant k, \quad i = \overline{1,l}, \quad l \leqslant k.$$

Кроме того, пусть заданы подмножества $A_i \subset A$ и $D_i \subset A$, причём $|A_i| = m_i$, $|D_i| = t_i$, $i = \overline{1,l}$. Рассмотрим множество $\Omega(l, t_1, \ldots, t_l) = \Omega$ источников Мура, у которых матрица вероятностей переходов $\theta$ удовлетворяет следующим условиям.

1. Строки матрицы $\theta$ можно разделить на $l$ классов, при этом $i$-класс содержит все строки, первый индекс которых принадлежит $B_i$, внутри $i$-го класса строки матрицы $\theta$ равны между собой, $i = \overline{1,l}$.

2. Элементы $\theta_{ij}$, $j \in A \backslash D_i$, $i = \overline{1,l}$, матрицы $\theta$ равны 0.

Не уменьшая общности, можно считать, что если $\theta \in \Omega$, то

$$\theta = \left\| \begin{array}{cccc} \theta_{11} & \theta_{12} & \ldots & \theta_{1k} \\ \ldots & \ldots & \ldots & \ldots \\ \theta_{11} & \theta_{12} & \ldots & \theta_{1k} \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ \theta_{l1} & \theta_{l2} & \ldots & \theta_{lk} \\ \ldots & \ldots & \ldots & \ldots \\ \theta_{l1} & \theta_{l2} & \ldots & \theta_{lk} \end{array} \right\|.$$

Возьмём произвольное слово $u$, $|u| = N$, которое порождено источником $\theta$, $\theta \in \Omega$, и начинается с буквы $a_{j_0}$. Обозначим через $r_{ij}(u)$ число появлений буквы $a_j$ после $a_i$ в слове $u$, тогда

$$P_\theta(u) = \theta_{j_0} \prod_{f=1}^{l} \prod_{j=1}^{k} \theta_{fj}^{\sum_{i=m_{f-1}+1}^{m_f} r_{ij}(u)}. \tag{8}$$

Здесь и далее $m_0 = 0$, $0^0 = 1$.

Далее нам потребуется равенство

$$\sum_{u \in A^N} P_\theta(u) r_{ij}(u) = (N-1)\theta_{0i}\theta_{ij}. \tag{9}$$

Справедливо следующее утверждение.

**Лемма 1.** При заданном выходном алфавите $B$ для избыточности $R(N, \Omega, B)$ универсального кодирования множества источников $\Omega = \Omega(l, t_1, \ldots, t_l)$ справедливо следующее неравенство:

$$R(N, \Omega, B) \leqslant \frac{\sum_{i=1}^{l} t_i - l}{2c(B)} \frac{\log N}{N} + \frac{\lambda}{N}, \tag{10}$$

где $\lambda$ — постоянная, не зависящая от $N$ и $\theta$.

**Доказательство**. Воспользуемся значением интеграла Дирихле

$$\int_{\theta_1 + \ldots + \theta_t = 1} \prod_{i=1}^{t} \theta_i^{r_i} \, d\theta = \frac{\prod_{i=1}^{t} \Gamma(r_i + 1)}{\Gamma(\sum_{i=1}^{t} r_i + k + 1)}. \tag{11}$$

Здесь и далее $d\theta = d\theta_1 \ldots d\theta_t$.

Можно вычислить среднюю вероятность слова $u$, $u \in A^N$. Обозначим через $\omega(\theta)$ плотность распределения Дирихле на заданном множестве $\Omega$:

$$\omega(\theta) = \Big(\alpha \prod_{i=1}^{l} \prod_{j \in A_i} \theta_{ij}^{1/2}\Big)^{-1}, \tag{12}$$

где $\alpha$ — нормирующий множитель. Среднюю вероятность слова $u$, $u \in A^N$, по множеству $\Omega$, на котором равенством (12) задана плотность $\omega(\theta)$, обозначим $\bar{P}(u)$. По определению

$$\bar{P}(u) = \int_\Omega P_\theta(u)\omega(\theta) \, d\theta. \tag{13}$$

Очевидно, что

$$\sum_{u \in A^N} \bar{P}(u) = 1. \tag{14}$$

В [17] строится дешифруемый код $\varphi$ для всех слов, удовлетворяющих равенству (14). Если выходной алфавит $B$ имеет пропускную способность $c(B)$, то длительность кодового слова $\varphi$ имеет вид

$$|\varphi(u)| = \frac{1}{c(B)} \log \bar{P}(u) + T(u). \tag{15}$$

Постоянная $T(u)$ монотонно убывает и не превосходит величины $4 + \log e$, $k \geqslant 2$.

Используя (13), значение интеграла Дирихле (11) и формулу Стирлинга в виде

$$|\log\Gamma(z) - z\log e - (z-1/2)\log z| < c, \qquad z \geqslant 1/2,$$

из (15) получаем

$$|\varphi(u)| = \frac{1}{c(B)} \sum_{f=1}^{l} \sum_{j\in D_i} \sum_{i=m_{f-1}+1}^{m_f} r_{ij} \log \frac{r_{ij}(u)}{r_i(u)} +$$

$$+ \frac{1}{c(B)} \sum_{i=1}^{l} \frac{t_i - 1}{2} \log\Big(r_i(u) + \frac{k}{2}\Big) + T(u),$$

где $r_i(u) = \sum_{j=1}^{k} r_{ij}$.

Отсюда из определения избыточности $R(N,\theta,\varphi,B)$ (6) следует:

$$R(N,\theta,\varphi,B) = \frac{1}{c(B)}\Big[\sum_{u\in A^N} P_\theta(u)\log\bar{P}(u) - H(\theta)\Big] \leqslant$$

$$\leqslant \frac{1}{c(B)}\Big(-\sum_{u\in A^N} P_\theta(u) \sum_{f=1}^{l} \sum_{j\in D_f} \sum_{i=m_{f-1}+1}^{m_f} r_{ij}(u)\log\frac{r_{ij}(u)}{r_i(u)} - H(\theta)\Big) +$$

$$+ \frac{1}{c(B)}\Big(\sum_{u\in A^N} P_\theta(u) \sum_{f=1}^{l} \frac{t_i-1}{2}\log\Big(r_f(u) + \frac{k}{2}\Big) + T\Big).$$

Из (9) и неравенства Йенсена для функций $-x\log x$ и $\log x$ из предыдущего неравенства вытекает

$$R(N,\theta,\varphi,B) \leqslant \frac{\sum_{i=1}^{l} t_i - l}{2c(B)} \frac{\log N}{N} + \frac{\lambda}{N},$$

где $\lambda$ не зависит от $\theta$.

Из последнего неравенства и определения $R(N,\Omega,B)$ (7) получаем справедливость оценки (10). Лемма доказана.

Найдём нижнюю границу избыточности универсального кодирования множества источников $\Omega$ при заданном выходном алфавите $B$, т. е. нижнюю оценку для $R(N,\Omega,B)$, которую будем искать методом усреднения, предложенным в [5].

Для этого потребуется определить среднюю избыточность $\bar{R}(N,\Omega,B)$ универсального кодирования на множестве источников $\Omega$ с заданным выходным алфавитом $B$:

$$\bar{R}(N,\Omega,B) = \inf_{\varphi} \int_{\Omega} \omega(\theta) R(N,\theta,\varphi,B)\, d\theta.$$

Пусть $\bar{\Omega}$ — подмножество источников из $\Omega(l,t_1,\ldots,t_l)$, у которых все ненулевые элементы $\theta_{ij}$ не меньше $\delta$. Аналогично лемме 1 из [6] доказывается лемма 2.

**Лемма 2.** Для произвольного источника $\theta \in \bar{\Omega}$ выполняются неравенства

$$-\frac{\lambda}{N} \leqslant \frac{1}{N} \sum_{u \in A^N} P_\theta(u) r_{ij} \log \frac{r_{ij}}{r_j} + \theta_{ji}\theta_{ij} \log \theta_{ij} \leqslant 0,$$

где $\lambda > 0$ не зависит от $\theta$.

Установим нижнюю оценку для избыточности $R(N, \Omega, B)$.

**Лемма 3.** Для избыточности $R(N, \Omega, B)$, $\Omega = \Omega(l, t_1, \ldots, t_l)$, универсального кодирования множества источников $\Omega$ буквами выходного алфавита $B$ имеет место неравенство

$$R(N, \Omega, B) \geqslant \frac{\sum_{i=1}^{l} t_i - l}{2c(B)} \frac{\log N}{N} + \frac{\lambda}{N},$$

где $\lambda$ не зависит от $\theta \in \Omega$.

**Доказательство**. В силу очевидных неравенств

$$R(N, \Omega, B) \geqslant R(N, \bar{\Omega}, B) \geqslant \bar{R}(N, \bar{\Omega}, B) \tag{16}$$

достаточно установить оценку

$$\bar{R}(N, \bar{\Omega}, B) \geqslant \frac{\sum_{i=1}^{l} t_i - l}{2c(B)} \frac{\log N}{N} + \frac{\lambda}{N}. \tag{17}$$

Так как $\bar{\Omega} \subset \Omega$, то для любого $u \in A^N$ выполнено неравенство

$$\int_{\Omega} \frac{P_\theta(u)\, d\theta}{\prod_{i=1}^{l} \prod_{j \in D_i} \theta_{ij}^{1/2}} \geqslant \int_{\bar{\Omega}} \frac{P_\theta(u)\, d\theta}{\prod_{i=1}^{l} \prod_{j \in D_i} \theta_{ij}^{1/2}}.$$

При неравнозначности букв выходного алфавита из определения $\bar{R}(N, \bar{\Omega}, B)$ и теоремы Шеннона [1] для каналов без шума следует

$$\bar{R}(N, \bar{\Omega}, B) = \inf_{\varphi} \int_{\bar{\Omega}} \frac{1}{\alpha \prod_{i=1}^{l} \prod_{j \in D_i} \theta_{ij}^{1/2}} R(N, \theta, \varphi, B)\, d\theta \geqslant$$

$$\geqslant \frac{1}{N} \int_{\bar{\Omega}} \frac{1}{c(B) \prod_{i=1}^{l} \prod_{j \in D_i} \theta_{ij}^{1/2}} \sum_{u \in A^N} P_\theta(u) (\log P_\theta(u) - \log \bar{P}(u))\, d\theta. \tag{18}$$

Из (18) получим неравенство

$$\bar{R}(\bar{\Omega}) \geqslant \int_{\Omega} \frac{1}{\alpha \prod_{i=1}^{l} \prod_{j \in D_i} \theta_{ij}^{1/2} c(B)} \Big[ \frac{1}{N} \sum_{u \in A^N} P_\theta(u) \log P_\theta(u) -$$

$$- \frac{1}{N} \sum_{i=1}^{l} \sum_{j \in D_f} \sum_{m_f - 1}^{m_f} r_{fj}(u) \log \frac{r_{fj}(u)}{r_f(u)} + \frac{1}{N} \sum_{i=1}^{l} \frac{t_i - 1}{2} \log r_i(u) + \frac{\lambda}{N} \Big] d\theta. \tag{19}$$

Здесь $\alpha$ — нормирующий множитель, а $\lambda$ не зависит от $\theta$.

Используя теорему о больших уклонениях [18], аналогично лемме 2 из [8] можно доказать, что при $\theta \in \bar{\Omega}$ имеет место равенство

$$\frac{1}{N} \sum_{u \in A^N} P_\theta(u) \log r_j(u) = \frac{\log N}{N} + \frac{\lambda}{N}, \qquad j = \overline{1,l}. \tag{20}$$

Из леммы 2, соотношений (19) и (20) вытекает справедливость (17). Тем самым в силу (16) лемма 3 доказана.

**Теорема 1.** Для избыточности $R(N, \Omega, B)$ универсального кодирования множества источников Мура $\Omega = \Omega(l, t_1, \ldots, t_l)$ буквами алфавита $B$ имеет место асимптотическое равенство

$$R(N, \Omega, B) \sim \frac{\sum_{i=1}^{l} t_i - l}{2c(B)} \frac{\log N}{N}.$$

**Доказательство**. Оценка сверху вытекает из леммы 1, а оценка снизу — из леммы 3. Теорема доказана.

Из этой теоремы легко получить асимптотику избыточности универсального кодирования множества всех марковских источников $\Omega_s$ при кодировании буквами алфавита $B$. Эта оценка получена в [15–17].

**Следствие.** Для избыточности $R(N, \Omega_s, B)$ универсального кодирования всех марковских источников с памятью $s$ буквами алфавита $B$ имеет место асимптотическое равенство

$$R(N, \Omega_s, B) \sim \frac{k^s(k-1)}{2c(B)} \frac{\log N}{N}.$$

**Доказательство**. Известно [19], что марковскую цепь с памятью $s$ можно рассматривать как обычную марковскую цепь, если входным алфавитом считать $A^s$. При этом длина кодируемого блока для нового алфавита будет, естественно, $N - s + 1$, а марковская цепь, определяющая источник с алфавитом $A^s$, имеет матрицу вероятностей переходов размера $k^s \times k^s$, причём в каждой строке этой матрицы ровно $k$ отличных от нуля элементов. В этом случае $l = k^s$, $t_1 = t_2 = \ldots = t_{k^s} = k$, поэтому по теореме 1

$$R(N - k + 1, \Omega_s, B) \sim \frac{k^{s+1} - k^s}{2c(B)} \frac{\log N - k + 1}{N - s + 1}.$$

Так как

$$\frac{\log N - k + 1}{N - k + 1} \sim \frac{\log N}{N},$$

из предыдущего асимптотического равенства вытекает справедливость доказываемого следствия.

**4. Кодирование источников Мили.** Пусть буквы входного алфавита $A = \{a_1, \ldots, a_k\}$ порождаются некоторым автоматом с $m$ состояниями $\{s_1, s_2, \ldots, s_m\}$. Вероятность порождения очередной буквы зависит только от состояния, в котором в данный момент находится автомат. После порождения очередной буквы автомат переходит в новое состояние. При этом вероятность перехода $P(s_i \mid s_j)$ из состояния $s_j$ в состояние $s_i$ вычисляется суммированием вероятностей букв, приводящих к переходу из $s_j$ в $s_i$, $i, j = \overline{1,m}$. Таким образом, работа вышеописанного источника определяется матрицей вероятностей переходов

$$\mu = \|P(a_i, s_j)\|, \quad i = \overline{1,k}, \quad j = \overline{1,m}$$

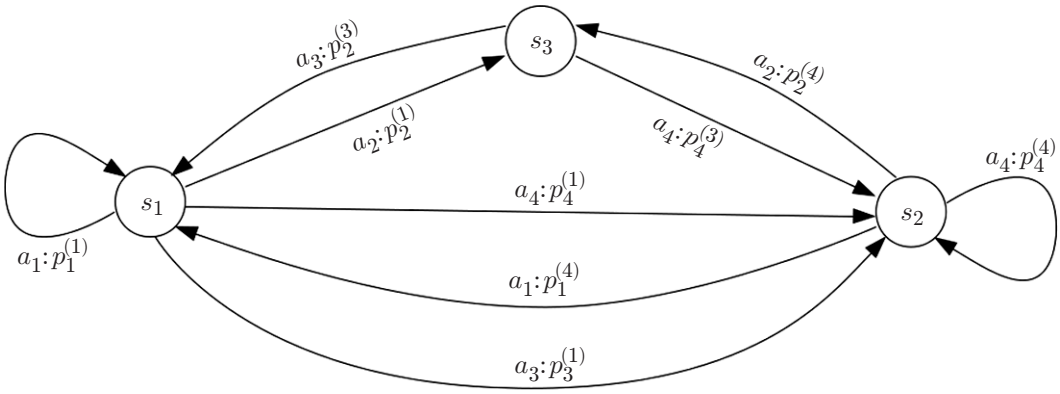


*Рис. 2*

и начальным вектором распределения для состояний автомата.

В дальнейшем рассматриваются только эргодические источники, которые будем называть марковскими источниками Мили или просто источниками Мили.

На рис. 2 изображён источник Мили с тремя состояниями автомата $s_1$, $s_2$, $s_3$, порождающего четыре буквы $a_1$, $a_2$, $a_3$, $a_4$. Узлы соответствуют состояниям автомата, а направленные рёбра — буквам источника и переходам между состояниями. Рёбра, выходящие из конкретного узла, должны отвечать различным буквам. Из этого следует, что новое состояние источника однозначно определяется его предыдущим состоянием и соответствующей буквой. Отсюда марковский источник Мили однозначно задаётся матрицей переходных вероятностей $\mu$ и автоматом Мили $G$. Этот граф обозначим той же буквой $G$.

Последовательность букв, порождаемую автоматом $G$, разобьём на слова длины $N$. Через $P_\mu(u)$ обозначим вероятность слова $u$, порождённого источником $G$ с матрицей вероятностей переходов $\mu$. В этом случае совокупность $\{P_\mu(u),\ u \in A^N\}$ задаёт источник Мили, т. е. пара $(\mu, G) = \theta_\mu$ определяет источник Мили. Множество всех пар $(\mu, G) = \theta$ обозначим $W(G)$. Для формулировки следующего утверждения нам потребуется ввести несколько дополнительных обозначений.

Для заданного автомата Мили $G$ через $f(s_j)$, $j = \overline{1,m}$, обозначим число рёбер, выходящих из состояния $s_j$ графа $G$. Тогда $V(G) = \sum_{j=1}^{m} f(s_j)$ — число всех рёбер в графе $G$. Для единообразия терминологии число состояний в графе $G$ будем обозначать $Q(G) = m$. Используя результаты разд. 3, докажем следующее утверждение.

**Теорема 2.** Для избыточности $R(N, w(G), B)$ универсального кодирования множества источников Мили $w(G)$ имеет место асимптотическое равенство

$$R(W(G)) \sim \frac{V(G) - Q(G)}{2c(B)} \frac{\log N}{N}.$$

**Доказательство**. Возьмём произвольный источник Мили $\theta_\mu$, $\theta_\mu \in W(G)$. Как указано в [3, с. 53], «для всякого автомата Мили существует эквивалентный ему (индуцирующий то же самое отображение) автомат Мура. Существует единственный конструктивный способ построения по автомату Мили эквивалентного ему автомата Мура».

Рассмотрим стационарный марковский источник Мура $\theta(\mu)$, состоящий из пар $s_i a_j$, $i = \overline{1,m}$, $j = \overline{1,k}$, с вероятностью перехода из состояния $s_l a_t$ в состояние $s_i a_j$:

$$P(s_i a_j \mid s_l a_t) = P(a_j \mid s_l). \tag{21}$$

Существует взаимооднозначное соответствие между словами, порождёнными источниками $\theta_\mu$ и $\theta(\mu)$. Если $u$ — слово, порождённое источником Мили $\theta_\mu$, то через $\tilde{u}$ обозначим соответствующее слово, порождённое источником Мура $\theta(\mu)$. Из определения источников Мура и Мили и равенства (21) следует, что

$$P_{\theta_\mu}(u) = P_{\theta(\mu)}(\tilde{u}). \tag{22}$$

Множество полученных источников обозначим $\Omega(G)$:

$$-\sum_{u\in A^N} P_{\theta_\mu}(u)\log P_{\theta_\mu}(u) = -\sum_{\tilde{u}\in A^N} P_{\theta(\mu)}(\tilde{u})\log P_{\theta(\mu)}(\tilde{u}). \tag{23}$$

Из определения избыточности универсального кодирования и равенств (22), (23) имеем

$$R(N, W(G), B) = R(N, \Omega(G), B). \tag{24}$$

Покажем, что

$$R(N, \Omega(G), B) \sim \frac{\sum_{j=1}^m f(s_j) - m}{2c(B)}\,\frac{\log N}{N}.$$

Так как $W(G)$ — эргодические источники, $\Omega(G)$ — также эргодические. Кроме того, из (21) следует, что множество источников из $\Omega(G)$ удовлетворяет условиям теоремы из разд. 3, если положить $l = m = Q(G)$, $t_i = f(s_i)$, $i = \overline{1,m}$. Поэтому согласно теореме 1 получаем

$$R(\Omega(G)) \sim \frac{\sum_{j=1}^m f(s_j) - m}{2c(B)}\,\frac{\log N}{N}.$$

Отсюда и из соотношения (24), а также с учётом $\sum_{j=1}^m f(s_j) = W(G)$ и $m = Q(G)$ вытекает справедливость теоремы 2.

**Заключение.** В данной работе рассматривается универсальное кодирование неравнозначными символами марковских источников, у которых некоторые строки матриц перехода совпадают. Для класса источников, когда число строк в таких матрицах фиксировано, предложено асимптотически оптимальное кодирование. В качестве следствия получены оценки [17]. Найдена оценка избыточности в случае, если источник задаётся графом с $v$ вершинами и $m$ рёбрами. Для таких источников избыточность асимптотически равна $\dfrac{v-m}{2c}\,\dfrac{\log N}{N}$.

## СПИСОК ЛИТЕРАТУРЫ


1. **Шеннон К.** Математическая теория связи. Работы по теории информации и кибернетике. М.: ИЛ, 1963. С. 243–332.
2. **Галлагер Р.** Теория информации и надёжная связь. М.: Сов. радио, 1974. 720 с.
3. **Глушков В. М.** Синтез цифровых автоматов. М.: Изд-во физ.-мат. лит., 1962. 432 с.
4. **Резник А. Л., Соловьев А. А., Торгов А. В.** Локализация случайных импульсно-точечных источников с применением физически реализуемых поисковых алгоритмов // Автометрия. 2020. **56**, № 6. С. 49–60. DOI: 10.15372/AUT20200606.

5. **Лемешко Б. Ю., Лемешко С. Б.** Влияние округления на свойства критериев проверки статистических гипотез // Автометрия. 2020. **56**, № 3. С. 35–45. DOI: 10.15372/AUT20200305.
6. **Фитингоф Б. М.** Оптимальное кодирование при неизвестной и меняющейся статистике сообщений // Проблемы передачи информации. 1966. **2**, № 2. С. 3–11.
7. **Кричевский Р. Е.** Связь между избыточностью кодирования и достоверностью сведений об источнике // Проблемы передачи информации. 1968. **4**, № 3. С. 48–57.
8. **Трофимов В. К.** Избыточность универсального кодирования произвольных марковских источников // Проблемы передачи информации. 1974. **10**, № 4. С. 16–24.
9. **Штарьков Ю. М.** Универсальное кодирование. Теория и алгоритмы. М.: Изд-во мат. лит., 2013. 288 с.
10. **Krichevsky R. E., Trofimov V. K.** The performance of universal encoding // IEEE Trans. Inform. Theory. 1981. **27**, N 2. P. 199–207.
11. **Davisson L. D.** Universal noiseless coding // IEEE Trans. Inform. Theory. 1973. **19**, N 6. P. 783–795.
12. **Штарьков Ю. М.** Обобщённые коды Шеннона // Проблемы передачи информации. 1984. **20**, № 3. С. 3–16.
13. **Штарьков Ю. М., Чокенс Ч. Дж., Виллемс Ф. М. Дж.** Оптимальное универсальное кодирование по критерию максимальной индивидуальной относительной избыточности // Проблемы передачи информации. 1997. **33**, № 1. С. 21–34.
14. **Кричевский Р. Е., Трофимов В. К.** Избыточность универсального кодирования. Новосибирск, 1981. 40 с. (Препр./ИМ СО АН СССР).
15. **Храмова Т. В., Трофимов В. К.** Сжатие неравнозначными символами информации, порождённой неизвестным источником без памяти // Автометрия. 2012. **48**, № 1. С. 30–44.
16. **Храмова Т. В., Трофимов В. К.** Сжатие информации, порождённой неизвестным источником // Электросвязь. 2012. № 4. С. 41–44.
17. **Трофимов В. К., Храмова Т. В.** Универсальное кодирование марковских источников неравнозначными символами // Дискретный анализ и исследование операций. 2013. **20**, № 3. С. 71–83.
18. **Нагаев С. В.** Уточнение предельных теорем для однородных цепей Маркова // Теория вероятностей и её приложения. 1961. **4**, № 1. С. 67–86.
19. **Романовский В. И.** Дискретные цепи Маркова. М.: Гостехиздат, 1949. 435 с.

*Поступила в редакцию 17.11.2020*
*После доработки 15.12.2020*
*Принята к публикации 21.12.2020*