ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Май—июнь 2013. Том 20, № 3. С. 71–83

УДК 621.391.15

УНИВЕРСАЛЬНОЕ КОДИРОВАНИЕ МАРКОВСКИХ ИСТОЧНИКОВ НЕРАВНОЗНАЧНЫМИ СИМВОЛАМИ *)

В. К. Трофимов, Т. В. Храмова

Аннотация. Предложен алгоритм кодирования информации, порождённой неизвестным марковским источником, при неравнозначности символов кодового алфавита, и доказана его оптимальность.

Ключевые слова: энтропия, кодирование, стоимость кодирования, источник сообщений.

Введение

В настоящей работе исследуется вопрос о кодировании информации, порождённой неизвестным марковским источником с конечной памятью s, буквами выходного (кодового) алфавита, имеющими различные длительности. Задача кодирования неравнозначными буквами последовательности, порождённой известным источником, впервые рассмотрена в [11]. В дальнейшем вопросы кодирования сообщений, порождённых известным источником, неравнозначными символами рассматривались в [10, 16].

Кодирование сообщений, порождённых неизвестным источником, буквами алфавита с равнозначными длительностями символов впервые рассмотрено в [9]. Кодирование сообщений, порождённых неизвестным источником, получило название *универсального* кодирования. Точная постановка задачи универсального кодирования сделана в [2]. Универсальное кодирование равнозначными символами для различных множеств источников и для различных типов кодирований интенсивно изучалось как у нас в стране, так и за рубежом. Кроме упомянутых работ отметим [4–6, 12–14]. Наиболее полная библиография по данному вопросу содержится в [3, 16, 17]. В частности, в [2] предложен метод универсального кодирования бернуллиевских источников и доказана его асимптотическая оптимальность. В [12] предложен метод универсального кодирования для множества источников с конечной памятью, а в [6] доказана его

*) Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проект № 12–07–00188) и Совета по грантам Президента Российской федерации (проект НШ–2175.2012.9).

асимптотическая оптимальность. В [7] авторами предложен метод универсального кодирования множества бернуллиевских источников неравнозначными символами и доказана его асимптотическая оптимальность. Как и в случае равнозначности длительности букв кодового алфавита, платой за незнание источника является множитель, пропорциональный $\log N$, здесь и далее $\log x = \log_2 x$, $0 \log 0 = 0$, N — длина кодируемого блока. В настоящей работе результаты из [7] обобщаются на множество марковских источников с памятью s.

1. Постановка задачи

Обозначим через Ω_s множество марковских источников порядка s с входным алфавитом $X = \{x_1, x_2, \dots, x_k\}$, $2 \leqslant k < \infty$. Рассмотрим произвольный источник $\Theta \in \Omega_s$, генерирующий блоки u длины N над алфавитом X:

$$u = x_{i_1} x_{i_2} \cdots x_{i_j} \cdots x_{i_N}, \quad x_{i_j} \in X.$$

Вероятность появления очередной буквы x_{i_j} в блоке u определяется s предшествующими, т. е.

$$P(x_{i_j} \mid x_{i_1} x_{i_2} \dots x_{i_{j-1}}) = P(x_{i_j} \mid x_{i_{j-s}} x_{i_{j-s+1}} \dots x_{i_{j-1}}).$$

Таким образом, вероятность произвольного сообщения $u \in X^N$, порождённого марковским источником $\Theta \in \Omega_s$, определяется k^s-мерным вектором вероятностей начальных блоков слов $v \in X^s$:

$$P_v^0 = \left(p_{v_1}^0, p_{v_2}^0, \dots, p_{v_{k^s}}^0\right), \quad \sum_{v \in X^s} p_v^0 = 1,$$

и $(k^s \times k)$-матрицей условных вероятностей появления буквы $x_i \in X$ после блока $v \in X^s$:

$$P = (p_{vi})_{v \in X^s;\, i = \overline{1,k}}, \quad p_{vi} = p\left(x_i \middle| v\right).$$

Введём следующие обозначения:

$$\begin{aligned} &n_i(u) \text{ — число вхождений } x_i \in X \text{ в блок } u \in X^N, \\ &n_{vi}(u) \text{ — число вхождений } x_i \in X \text{ в } u \in X^N \text{ после блока } v \in X^s, \\ &n_v(u) = \sum_{x_i \in X} n_{vi}. \end{aligned} \tag{1}$$

Обозначим через $p(u)$ вероятность появления блока $u \in X^N$ на выходе источника $\Theta \in \Omega_s$. Имеет место формула

$$p(u) = p_0(u) \prod_{v \in X^s} \prod_{i=1}^{k} p_{vi}^{n_{vi}(u)}, \tag{2}$$

где $p_0(u)$ — вероятность появления начального s-блока слова u.

Как известно [1, 8, 10, 11, 16], энтропия марковского источника $\Theta \in \Omega_s$ вычисляется по формуле

$$H(\Theta) = -\sum_{v \in X^s} p_v^0 \sum_{i=1}^{k} p_{vi} \log p_{vi}, \tag{3}$$

где p_v^0 — стационарное распределение.

В случае, если Θ^N — марковский источник, все сообщения которого имеют длину N, имеет место равенство [10, 16]

$$H(\Theta) = \lim_{N \to \infty} \frac{H(\Theta^N)}{N}. \tag{4}$$

Рассмотрим задачу кодирования сообщений источника буквами кодового алфавита $Y = \{y_1, y_2, \ldots, y_m\}$, которые могут быть как равнозначны по длительности (стоимости передачи), так и нет. Через $t(y_i) = t_i$ обозначим длительность кодового символа $y_i \in Y$, $i = \overline{1, m}$. Рассмотрим определяемый кодовым алфавитом вектор длительностей кодовых символов $\bar{t}_Y$, компоненты которого равны длительностям кодовых символов:

$$\bar{t}_Y = (t_1, t_2, \ldots, t_m), \quad t(y_i) = t_i, \quad y_i \in Y,\ i = \overline{1, m}.$$

Положим $\bar{t}_Y = \bar{t}_1$ в случае равнозначных длительностей $t_1 = t_2 = \ldots = t_m = 1$.

Пропускная способность канала без шума $C(t_Y)$ полностью определяется кодовым алфавитом Y [10, 11, 15] и вычисляется по формуле

$$C(\bar{t}_Y) = \log \omega_o(\bar{t}_Y), \tag{5}$$

где $\omega_o(\bar{t}_Y)$ — наибольший положительный корень уравнения

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

В частности, $C(\bar{t}_1) = \log m$.

Кодирование φ источника Θ заключается в преобразовании каждого слова $u \in X^N$, порождённого Θ, в некоторую последовательность $\varphi(u)$ из букв кодового алфавита Y. Обозначим через $l\left(\varphi\left(u\right), \bar{t}_Y\right)$ длительность кодового слова $\varphi(u)$ при кодировании буквами алфавита Y:

$$l(\varphi(u), \bar{t}_Y) = \sum_{y \in Y} t(y).$$

Стоимостью кодирования φ блоков длины N, порождённых источником Θ, словами кодового алфавита Y назовём величину

$$\widetilde{L}(N,\Theta,\varphi,\bar{t}_Y)=\sum_{u\in X^N}p(u)l(\varphi(u),\bar{t}_Y). \tag{6}$$

Стоимость кодирования является его качественным показателем: чем она меньше, тем эффективнее код.

Среднюю длительность букв выходного алфавита, приходящихся на одну букву входного алфавита, обозначим через

$$L(N,\Theta,\varphi,\bar{t}_Y)=\frac{\widetilde{L}(N,\Theta,\varphi,\bar{t}_Y)}{N}.$$

Разность между стоимостью кодирования на букву сообщения источника и отношением энтропии источника к пропускной способности канала назовём *избыточностью кодирования* φ и обозначим через

$$R(N,\Theta,\varphi,\bar{t}_Y)=L(N,\Theta,\varphi,\bar{t}_Y)-\frac{H(\Theta)}{C(\bar{t}_Y)}. \tag{7}$$

Избыточностью универсального кодирования для множества источников Ω и выходного алфавита Y, следуя [2], назовём величину $R(N,\Omega,\bar{t}_Y)$, определяемую равенством

$$R(N,\Omega,\bar{t}_Y)=\inf_{\varphi}\sup_{\Theta\in\Omega}R(N,\Theta,\varphi,\bar{t}_Y). \tag{8}$$

При $\Omega=\Omega_s$ получаем избыточность универсального кодирования для множества марковских источников с памятью $s\geqslant 0$ и выходным алфавитом Y. Если же Ω содержит единственный источник Θ, то имеем дело с кодированием известного источника.

2. Кодирование неизвестного марковского источника

Рассмотрим произвольный марковский источник $\Theta\in\Omega_s$, генерирующий блоки длины N над алфавитом X. Используя идею, предложенную в [15], построим универсальное кодирование в алфавит с неравнозначными символами. Зададим на множестве X^N всех блоков $u\in X^N$ средние вероятности $\bar{p}^s(u)$, полученные с помощью КТ-распределения [16]:

$$\bar{p}^s(u)=\bar{p}^s=\left[\frac{\Gamma\left(\frac{k}{2}\right)}{k\pi^{\frac{k}{2}}}\right]^{k^s}\prod_{v\in X^s}\frac{\prod\limits_{i=1}^{k}\Gamma\left(n_{vi}(u)+\frac{1}{2}\right)}{\Gamma\left(n_v(u)+\frac{k}{2}\right)}, \tag{9}$$

где $\Gamma(z)$ — гамма-функция.

Упорядочим слова u, порождённые источником, по убыванию средних вероятностей $\overline{p}^s(u)$. Определим разбиение произвольного интервала $\Im_Y([a,b))$, которое будет играть важную роль в построении оптимального кодирования. Положим

$$\Im_Y([a,b)) = \{I_j([a,b))\}_{j=\overline{1,m}},$$

где каждый интервал $I_j([a,b))$ определяется следующим образом:

$$I_j([a,b)) = [a_{j-1}, a_j),$$

$$a_{j-1} = a + (b-a)\sum_{i=1}^{j-1} \omega_0^{-t_i}, \quad a_j = a + (b-a)\sum_{i=1}^{j} \omega_0^{-t_i}, \quad a_0 = a, \quad a_m = b.$$

Через $\Im_Y^n([a,b))$ обозначим разбиение, полученное n-кратным применением процедуры разбиения $\Im_Y([a,b))$ к интервалу $[a,b)$, т.е. разбиение разбиений:

$$\Im_Y^n([a,b)) = \{I_{j_1 j_2 \ldots j_n}([a,b))\}_{j_i=\overline{1,m},\ i=\overline{1,n}},$$
$$I_{j_1 j_2 \ldots j_n}([a,b)) = I_{j_n}(I_{j_{n-1}}(\ldots I_{j_2}(I_{j_1}([a,b)))\ldots)).$$

По определению $\Im^1 = \Im$.

В соответствие каждой последовательности $u_i = x_{i_1} x_{i_2} \cdots x_{i_N}$ поставим величину

$$\alpha(u_i) = \sum_{n=1}^{i-1} p(u_n), \quad i = \overline{2, k^N}, \quad \alpha(u_1) = 0.$$

Пусть $j_1 j_2 \ldots j_n$ — набор индексов интервала $I_{j_1 j_2 \ldots j_n}([0,1)) \subset \Im_Y^n([0,1))$, удовлетворяющего системе условий

$$\left\{\begin{array}{c} \alpha(u_i) \in I_{j_1 j_2 \ldots j_n}([0,1)), \\ \left\{\begin{array}{l} \alpha(u_{i-1}) \notin I_{j_1 j_2 \ldots j_n}([0,1)), \\ \alpha(u_{i+1}) \notin I_{j_1 j_2 \ldots j_n}([0,1)), \end{array}\right. \\ \left[\begin{array}{l} \alpha(u_{i-1}) \in I_{j_1 j_2 \ldots j_{n-1}}([0,1)), \\ \alpha(u_{i+1}) \in I_{j_1 j_2 \ldots j_{n-1}}([0,1)). \end{array}\right. \end{array}\right. \tag{10}$$

Определим кодирование $\varphi_{\omega_o,p}$ как функцию

$$\varphi_{\omega_o,p}(u_i) = y_{j_1} y_{j_2} \ldots y_{j_n}, \tag{11}$$

отображающую каждую последовательность $u_i \in X^N$ в слово $\varphi(u_i) = y_{j_1} y_{j_2} \ldots y_{j_n}$ из букв кодового алфавита Y, где $j_1 j_2 \ldots j_n$ — набор индексов интервала $I_{j_1 j_2 \ldots j_n}$, определяемого условиями (10)

Отметим, что условие строгой положительности значений $\overline{p}^s(u)$ обеспечивает конечность предложенной процедуры кодирования. Действительно, поскольку $\alpha(u_{i-1}) \neq \alpha(u_i)$, $i = \overline{1, n}$, каким бы малым ни было различие, всегда найдётся $n' \in \mathbb{N}$ такое, что для разбиения $\Im_Y^{n'}([0,1))$ значения $\alpha(u_{i-1})$ и $\alpha(u_i)$ окажутся в различных (не обязательно смежных) интервалах из $\Im_Y^{n'}([0,1))$.

Условия (10) гарантируют дешифруемость кодирования (11).

Для кодирования (11) имеют место неравенства

$$|I_{j_1 j_2 \ldots j_{n-1}}([0,1))| > p(u_i),$$
$$|I_{j_1 j_2 \ldots j_{n-1}}([0,1))| \leqslant \omega_0^{-l(\varphi_{\omega_o, p(u_i)}, t_Y) + t^{**}}, \qquad (12)$$

где $t^{**} = \max\limits_{1 \leqslant j \leqslant m} \{t_j\}$.

Оценить максимальное различие между неизвестным распределением источника и КТ-распределением позволяет

Лемма 1. *Для марковского источника $\Theta \in \Omega_s$ с алфавитом X, $|X| = k$, и законом распределения $p(u)$ с условными вероятностями $p_{vj}(u) = p_{vj}$, $u \in X^N$, $v \in X^s$, $x_j \in X$, имеет место неравенство*

$$\log \frac{p(u)}{\overline{p}^s(u)} \leqslant \frac{k^s(k-1)}{2} \log(N-s)(1+o(1)) + \log p_0(u), \qquad (13)$$

где $p(u)$ определяется равенством (2), *а $\overline{p}^s(u)$ — равенством* (9).

Доказательство. Рассмотрим произвольное слово (блок) $u \in X^N$. Если это не приводит к неоднозначности, то в дальнейших выкладках положим $n_{vj}(u) \equiv n_{vj}$, $n_v(u) \equiv n_v$ и $n_j(u) \equiv n_j$. Сравним значения $p(u)$ и $\overline{p}^s(u)$. Из (2) имеем $p(u) = p_0(u) \prod\limits_{v \in X^s} \prod\limits_{j=1}^{k} p_{vj}^{n_j}(u)$. Очевидно, что

$$\frac{\overline{p}^s(u)}{p(u)} \geqslant \frac{\overline{p}^s(u)}{\max\limits_{\Theta \in \Omega} p(u)}.$$

Максимальное значение функции $p(u)$ достигается при

$$p_{vj}(u) = \frac{1}{n_v(u)^{n_v(u)}} \prod_{j=1}^{k} n_{vj}^{n_{vj}}$$

и равно

$$\max_{\Theta\in\Omega} p(u) = p_0(u) \prod_{v\in X^s} \frac{\prod\limits_{j=1}^{k} n_{vj}^{n_{vj}}}{n_v^{n_v}}.$$

Используя (13) и последнее равенство, получим

$$\log\frac{\overline{p}^s(u)}{p(u)} \geqslant k^s \log\Gamma\left(\frac{k}{2}\right) - k^s\log k - \frac{k^s\cdot k}{2}\log\pi + \sum_{v\in X^s}\sum_{j=1}^{k}\log\Gamma\left(n_{vj}+\frac{1}{2}\right) - \sum_{v\in X^s}\log\Gamma\left(n_v+\frac{k}{2}\right) + \sum_{v\in X^s}\log n_v^{n_v} - \log p_0(u) - \sum_{v\in X^s}\sum_{j=1}^{k}\log n_{vj}^{n_{vj}}. \quad (14)$$

Применим к слагаемым в правой части формулу Стирлинга в виде

$$\log\Gamma(z) = \log\sqrt{2\pi} + \left(z-\frac{1}{2}\right)\log\left(z-\frac{1}{2}\right) - z\log e + \xi(z)\log e,$$

где

$$\frac{1}{2} - \frac{1}{2\log e} \leqslant \xi(z) \leqslant \frac{1}{2}. \quad (15)$$

Положим $\Xi = k^s\xi + \sum\limits_{v\in X^s}\sum\limits_{j=1}^{k}\xi_{vj} - \sum\limits_{v\in X^s}\xi_v$, где каждая из констант ξ, ξ_{vj}, ξ_v ($v\in X^s$, $j=\overline{1,k}$) удовлетворяет ограничениям (15). Упрощая (17) и учитывая (1), получаем соотношение

$$\log\frac{\overline{p}^s(u)}{p(u)} \geqslant \frac{k^s\cdot k}{2} - k^s\log k + \frac{k^s\log(k-1)^{(k-1)}}{2} - \frac{k^s\cdot k}{2} + \frac{k^s}{2} - \log\prod_{v\in X^s}\frac{\left(n_v+\frac{k-1}{2}\right)^{\left(n_v+\frac{k-1}{2}\right)}}{n_v^{n_v}} - \log p_0(u) - \frac{k^s\cdot k}{2}\log e + \Xi\log e.$$

Используя ограничения на величину $\Xi\log e$, преобразуем последнее нера-

венство:

$$\log \frac{p(u)}{\overline{p}^s(u)} \leqslant \frac{k^s}{2}(k + \log k^2 - \log(k-1)^{(k-1)}) + \log \prod_{v \in X^s} \frac{\left(n_v + \frac{k-1}{2}\right)^{\left(n_v + \frac{k-1}{2}\right)}}{n_v^{n_v}} + \log p_0(u).$$

Наибольшее значение функции $\log \prod_{v \in X^s} \frac{\left(n_v + \frac{k-1}{2}\right)^{\left(n_v + \frac{k-1}{2}\right)}}{n_v^{n_v}}$ на промежутке $[1, N-s]$ достигается на правой границе, следовательно,

$$\log \frac{p(u)}{\overline{p}^s(u)} \leqslant \frac{k^s(k + \log k^2 - \log(k-1)^{(k-1)})}{2} + \frac{k^s\,(k-1)}{2}\left(1 + \frac{k-1}{2\,(N-s)}\right)\log e + \frac{k^s\,(k-1)}{2}\log\,(N-s) + \log p_0(u).$$

Пусть $T^*(k) = \frac{k+\log k^2}{k-1} - \log\,(k-1) + \log e$ и $\alpha(N,k,s) = \frac{(k-1)\log e}{2(N-s)}$. Тогда предыдущее неравенство принимает вид

$$\log \frac{p\,(u)}{\overline{p}^s\,(u)} \leqslant \frac{k^s\,(k-1)}{2}\,(T^*(k) + \alpha(N,s,k) + \log\,(N-s)) + \log p_0(u).$$

Очевидно, что $T^*(k)$ постоянна относительно длины блока N, а $\alpha(N, k, s)$ — бесконечно малая величина при $N \to \infty$. Поэтому, вынося за скобки в первом слагаемом правой части $\log\,(N-s)$, получим (13). Лемма 1 доказана.

Из леммы 1 вытекает

Лемма 2. *Для источника $\Theta \in \Omega_s$ с алфавитом X, $|X| = k$, и законом распределения $p(u)$ с условными вероятностями $p^s_{vj}(u) = p_{vj}$, $u \in X^N$, имеет место неравенство*

$$-\sum_{u \in X^N} p\,(u)\log \overline{p}^s\,(u) \leqslant H(\Theta^N) + \frac{k^s\,(k-1)}{2}\log\,(N-s)\,(1+o(1))\,. \tag{16}$$

Доказательство. Используем результат леммы 1. Умножая правую и левую части неравенства (13) на $p(u)$ и суммируя по $u \in X^N$, после преобразований получим

$$-\sum_{u \in X^N} p\,(u)\log \overline{p}^s\,(u) \leqslant -\sum_{u \in X^N} p\,(u)\log \frac{p\,(u)}{p_0(u)} + \frac{k^s\,(k-1)}{2}\,(T^*(k) + \alpha(N,s,k) + \log\,(N-s))\,.$$

Утверждение леммы следует из определения и свойства условной энтропии (см., например, [15]):

$$-\sum_{u\in X^N} p(u)\log\frac{p(u)}{p_0(u)} = H(\Theta^N \big| \mathbf{P}^0) \leqslant H(\Theta^N).$$

Лемма 2 доказана.

3. Верхняя оценка избыточности универсального кодирования множества марковских источников

Теорема 1. *Для избыточности равномерного по входу кодирования* φ (11) *неизвестного марковского источника* $\Theta \in \Omega_s$ *в слова выходного алфавита с различными длительностями кодовых букв и определяемой этим алфавитом пропускной способностью канала передачи информации* $C(\bar{t}_Y)$ *имеет место неравенство*

$$R\left(N, \Omega_s, \varphi, \bar{t}_Y\right) \leqslant \frac{k^s\,(k-1)}{2C(\bar{t}_Y)}\cdot\frac{\log\,(N-s)}{N}\,(1+o(1))\,, \tag{17}$$

где t^{**} *— константа, зависящая только от канала.*

Доказательство. Рассмотрим средние вероятности, полученные с помощью КТ-распределения $\overline{p}^s = \overline{p}^s(u)$, $u \in X^N$, и кодирование $\varphi_{\omega_0,\overline{p}^s}$ источника Θ, определяемые равенствами (9) и (11) соответственно. Из свойств (12) и (13) имеем $\overline{p}^s(u) < \omega_0^{-l\left(\varphi_{\omega_0,\overline{p}^s}(u_i),t_Y\right)+t^{**}}$, где $t^{**} = \max\limits_{1\leqslant j\leqslant m}\{t_j\}$, $\varphi_{\omega_o,\overline{p}^s}(u_i) = y_{j_1}y_{j_2}\ldots y_{j_n}$. Логарифмируя и преобразовывая это неравенство, получим

$$\frac{\log\overline{p}^s}{\log\omega_0} < -(l(\varphi_{\omega_0,\overline{p}^s}(u_i),t_Y) - t^{**}).$$

Умножая это неравенство на $p\,(u)$ и суммируя по всем словам $u \in X^N$, имеем

$$\sum_{u\in X^N} p\,(u)\log\overline{p}^s(u) < -\log\omega_0\Big(\sum_{u\in X^N} p\,(u)\,l(\varphi_{\omega_0,\overline{p}^s}(u),t_Y) - \sum_{u\in X^N} p\,(u)t^{**}\Big).$$

С учётом (5) и (6) получаем

$$C\left(\bar{t}_Y\right)(\widetilde{L}_{\varphi_{\omega_0,\overline{p}^s}} - t^{**}) < -\sum_{u\in X^N} p\,(u)\log\overline{p}^s\,(u).$$

Преобразовывая правую часть и используя оценку (16), имеем

$$C\left(\bar{t}_Y\right)(\widetilde{L}_{\varphi_{\omega_0},\overline{p}^s} - t^{**}) < H(\Theta^N) + \frac{k^s\,(k-1)}{2}\log\,(N-s) \\ + \frac{k^s\,(k-1)}{2}\left(T^*(k) + \alpha(N,s,k)\right).$$

Проводя преобразования этого неравенства с использованием определения избыточности (7), получаем (17). Теорема 1 доказана.

4. Нижняя оценка избыточности универсального кодирования множества марковских источников

Пусть $R\left(N,\Theta,\varphi,\bar{t}_Y\right)$ — избыточность кодирования φ источника Θ, определяемая равенством (7). Обозначим через $f\,(\Theta)$ распределение вероятностей, заданное на множестве марковских источников Ω_s и определяемое равенством [16]:

$$f\,(\Theta) = \left[\frac{\Gamma\left(\frac{k}{2}\right)}{\pi^{\frac{k}{2}}}\right]^{k^s} \cdot \frac{1}{\prod\limits_{v\in X^s}\prod\limits_{i=1}^{k}\sqrt{p_{vi}^{\Theta}}}. \tag{18}$$

Величину $\overline{R}(N,\Omega,\bar{t})$, определяемую равенством

$$\overline{R}(N,\Omega_s,\bar{t}_Y) = \inf_{\varphi}\int\limits_{\Omega_s} f(\Theta)R(N,\Theta,\varphi,\bar{t}_Y)\,\mathrm{d}\Theta, \tag{19}$$

где $\mathrm{d}\Theta = \prod\limits_{v\in X^s}\prod\limits_{i=1}^{k}\mathrm{d}p_{vi}^{\Theta}$, $\int\limits_{\Omega_s}$ — $k^s(k-1)$-мерный интеграл по множеству Ω_s, назовём *средней избыточностью универсального кодирования* множества источников Ω_s.

Теорема 2. *Для средней избыточности* $\overline{R}(N,\Omega,\bar{t})$ *универсального кодирования множества марковских источников* Ω_s *справедливо асимптотическое неравенство*

$$\overline{R}(N,\Omega_s,\bar{t}_Y) \gtrsim \frac{k^s(k-1)}{2C(\bar{t}_Y)}\cdot\frac{\log N}{N}. \tag{20}$$

Доказательство. Как следует из теоремы кодирования для каналов без шума [15], правая часть (19) достигает своего минимума при кодировании $\varphi_{\frac{1}{2}}$ таком, что

$$l(\varphi_{\frac{1}{2}}(u),t_Y) = \log_{\omega_0}\overline{p}^s(u) = \frac{1}{C(\bar{t}_Y)}\cdot\log\overline{p}^s(u).$$

Отсюда и из (19) следует, что

$$\overline{R}(N,\Omega_s,\bar{t}_Y) \geqslant \frac{1}{C(\bar{t}_Y)} \int\limits_{\Omega_s} f(\Theta) R(N,\Theta,\varphi_{\frac{1}{2}},\bar{t}_1)\,\mathrm{d}\Theta. \qquad (21)$$

В силу результатов из [16] кодирование $\varphi_{\frac{1}{2}}$ оптимально для множества всех источников Ω_s и имеет место неравенство

$$\left| R(N,\Theta,\varphi_{\frac{1}{2}},\bar{t}_1) - \frac{k^s(k-1)}{2}\cdot\frac{\log N}{N} \right| < \frac{C(\bar{t}_Y)}{N}.$$

Отсюда и из (21) получаем

$$\overline{R}(N,\Omega_s,\bar{t}_Y) \gtrsim \frac{k^s(k-1)}{2C(\bar{t}_Y)}\cdot\frac{\log N}{N}.$$

Теорема 2 доказана.

5. Асимптотика $R(N,\Omega_s,\bar{t}_Y)$

Теорема 3. *Для избыточности* $(N,\Omega_s,\bar{t}_Y)$ *универсального кодирования множества марковских источников* Ω_s *буквами неравнозначного выходного алфавита справедливо асимптотическое равенство*

$$R(N,\Omega_s,\bar{t}_Y) \cong \frac{k^s(k-1)}{2C(\bar{t}_Y)}\cdot\frac{\log N}{N}. \qquad (22)$$

Доказательство. Оценка сверху следует из (17). В самом деле, из (17), (7) и (8) для любого источника $\Theta \in \Omega_s$ имеем

$$R(N,\Theta,\varphi_{\omega_0,p},\bar{t}_Y) \leqslant \frac{k^s(k-1)}{2C(\bar{t}_Y)}\cdot\frac{\log(N-s)+T^*(k)+\alpha(N,s,k)}{N} + \frac{t^{**}}{N}. \qquad (23)$$

Неравенство (23) справедливо при любом законе распределения вероятностей p, величины $T^*(k)$, $\alpha(N,s,k)$ и t^{**} ограничены. Правая часть не зависит от закона распределения p , следовательно, из соотношения (23) получаем

$$\sup_{\Theta\in\Omega_s} R(N,\Theta,\varphi_{\omega_0,p},\bar{t}_Y) \leqslant \frac{k^s(k-1)}{2C(\bar{t}_Y)}\cdot\frac{\log(N-s)+T^*(k)+\alpha(N,s,k)}{N} + \frac{t^{**}}{N}. \qquad (24)$$

Так как согласно (8)

$$R(N,\Omega_s,\bar{t}_Y) \leqslant \sup_{\Theta\in\Omega_s} R(N,\Theta,\varphi_{\omega_0,p},\bar{t}_Y),$$

из (24) следует верхняя оценка для $R(N,\Omega_s,\bar{t}_Y)$.

Справедливость нижней оценки вытекает из определения (19), неравенств (20) и

$$R(N,\Omega_s,\bar{t}_Y) \geqslant \overline{R}(N,\Omega_s,\bar{t}_Y).$$

Теорема 3 доказана.

ЛИТЕРАТУРА

1. **Галлагер Р. Г.** Теория информации и надежная связь. — М.: Сов. радио, 1974. — 720 с.
2. **Кричевский Р. Е.** Связь между избыточностью кодирования и достоверностью сведений об источнике // Пробл. передачи информ. — 1968. — Т. 4, № 3. — С. 48–57.
3. **Потапов В. Н.** Обзор методов неискажающего кодирования дискретных источников // Дискрет. анализ и исслед. операций. Сер. 1. — 1999. — Т. 6, № 4. — С. 49–91.
4. **Рябко Б. Я.** Кодирование источника с неизвестными, но упорядоченными вероятностями // Пробл. передачи информ. — 1979. — Т. 15, № 2. — С. 71–77.
5. **Рябко Б. Я.** Дважды универсальное кодирование // Пробл. передачи информ. — 1984. — Т. 20, № 3. — С. 24–28.
6. **Трофимов В. К.** Избыточность универсального кодирования произвольных марковских источников // Пробл. передачи информ. — 1974. — Т. 10, № 4. — С. 16–24.
7. **Трофимов В. К., Храмова Т. В.** Сжатие неравнозначными символами информации, порождённой неизвестным источником без памяти // Автометрия. — 2012. — Т. 48, № 1. — С. 30–44.
8. **Фано Р.** Передача информации. Статистическая теория связи. — М.: Мир, 1965. — 440 с.
9. **Фитингоф Б. М.** Оптимальное кодирование при неизвестной и меняющейся статистике сообщений // Пробл. передачи информ. — 1966. — Т. 2, № 2. — С. 3–11.
10. **Чисар И.** О каналах без шума // Пробл. передачи информ. — 1970. — Т. 6, № 4. — С. 3–15.
11. **Шеннон К.** Математическая теория связи. Работы по теории информации и кибернетике. — М.: Изд-во иностр. лит., 1969. — С. 243–332.
12. **Штарьков Ю. М.** Кодирование сообщений конечной длины на выходе источника с неизвестной статистикой // Мат. V конф. по теории кодирования и передачи инф. Ч. 1. — Москва; Горький, 1972. — С. 147–152.

13. **Davisson L. D.** Universal noiseless coding // IEEE Trans. Inform. Theory. — 1973. — Vol. 19, N 6. — P. 783–795.
14. **Elias P.** Universal codeword sets and representations of the integers // IEEE Trans. Inform. Theory. — 1975. — Vol. 21, N 2. — P. 194–203.
15. **Katona G.** General theory of noiseless channels // UDINE 1970. Courses and lectures. — N 31. — 69 p.
16. **Krichevsky R. E., Trofimov V. K.** The performance of universal encoding // IEEE Trans. Inform. Theory. — 1981. — Vol. 27, N 2. — P. 199–207.
17. **Verdu S.** Fifty years of Shannon theory // IEEE Trans. Inform. Theory. — 1998. — Vol. 44, N 6. — P. 2057–2078.

Трофимов Виктор Куприянович,
e-mail: trofimov@sibsutis.ru
Храмова Татьяна Викторовна,
e-mail: tvkhramova@gmail.com

Статья поступила
9 августа 2012 г.

Переработанный вариант —
10 октября 2012 г.