ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Июль—сентябрь 2001. Серия 1. Том 8, № 3, 26–45

УДК 519.713

О НИЖНЕЙ ОЦЕНКЕ СТОИМОСТИ КОДИРОВАНИЯ И АСИМПТОТИЧЕСКИ ОПТИМАЛЬНОМ КОДИРОВАНИИ СТОХАСТИЧЕСКИХ КОНТЕКСТНО-СВОБОДНЫХ ЯЗЫКОВ*)

Л. П. Жильцова

Рассматривается язык, порожденный стохастической контекстно-свободной грамматикой с однозначным выводом, для которой матрица первых моментов неразложима, непериодична и ее перронов корень строго меньше единицы. Для такого языка получена неулучшаемая нижняя оценка стоимости двоичного кодирования. Построен также алгоритм асимптотически оптимального кодирования.

Введение

К. Шеннон в [10] рассматривал задачу кодирования сообщений, генерируемых эргодическим источником с конечным числом состояний. Он показал, что

1) все сообщения достаточно большой длины N можно разбить на две группы: маловероятные и высоковероятные;

2) стоимость любого кодирования, имеющего однозначное декодирование, ограничена снизу величиной энтропии $H(I)$ источника I (в качестве стоимости кодирования рассматривается среднее число двоичных символов, используемых на кодирование одной буквы сообщения);

3) для любого $\varepsilon > 0$ существует равномерное (блочное) кодирование f такое, что его стоимость $C_f(I)$ удовлетворяет неравенству $C_f(I) \leqslant H(I) + \varepsilon$.

В настоящей статье рассматриваются вопросы, относящиеся к кодированию сообщений, являющихся словами стохастического контекстно-свободного языка (стохастического КС-языка) при некоторых ограничениях на порождающую грамматику. А именно предполагается, что

*) Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проект 01–01–00464) и Федеральной целевой программы «Интеграция» (проект АО–110).

каждое слово языка, порождаемого грамматикой, имеет единственный левый вывод и матрица первых моментов грамматики неразложима, непериодична и ее максимальный по модулю собственный корень строго меньше единицы.

Автором в [4] для слов произвольного стохастического КС-языка, порождаемого грамматикой из рассматриваемого класса, установлены свойства, аналогичные свойствам слов, генерируемых произвольным эргодическим конечным источником. Для стохастического КС-языка в качестве слов большой длины рассматривается множество слов, каждому из которых соответствует дерево вывода высоты t. Установлено, что при $t \to \infty$ почти все такие слова (с суммарной вероятностью, стремящейся к единице) имеют приблизительно одинаковый состав правил в выводе и, как следствие, приблизительно одинаковые вероятности и буквенный состав.

В настоящей работе на основе полученных в [4] закономерностей применения правил грамматики получена неулучшаемая нижняя оценка стоимости кодирования и построен алгоритм асимптотически оптимального кодирования, сравнимый по сложности с алгоритмом кодирования Шеннона из [10]. В основе построенного алгоритма лежит «блочное» кодирование деревьев вывода, и блоком является поддерево дерева вывода, имеющее фиксированную высоту.

Полученные в работе результаты можно рассматривать как обобщение результатов К. Шеннона на класс рассматриваемых КС-языков.

1. Основные определения

Пусть $B = \{b_1, b_2, \dots, b_n\}$ — алфавит. Через B^* обозначим множество всех конечных последовательностей в алфавите B, включая пустое слово, через B^+ — множество всех непустых конечных последовательностей в алфавите B. Произвольное подмножество $L \subseteq B^*$ называется *языком* в алфавите B, а $\alpha \in L$ — *словом* в языке L. Будем рассматривать бесконечные языки.

Пусть на множестве слов языка L задано распределение вероятностей P. Через $p(\alpha)$ обозначим вероятность слова α. Будем рассматривать только такое распределение вероятностей, когда $p(\alpha) > 0$ для любого $\alpha \in L$. Множество $\mathscr{L} = \{(\alpha, p(\alpha)) \mid \alpha \in L\}$ будем называть *стохастическим языком.* Для стохастического языка будем применять запись $\mathscr{L} = (L, P)$.

Для изложения результатов о контекстно-свободных языках будем использовать определения контекстно-свободного языка и стохастического КС-языка из [1, 9].

Стохастической КС-грамматикой называется система $G = \langle V_T, V_N, R, s\rangle$, где V_T и V_N — конечные множества терминальных и нетерминальных символов (терминалов и нетерминалов) соответственно; $s \in V_N$ — аксиома, $R = \bigcup_{i=1}^{k} R_i$, где k — мощность алфавита V_N и $R_i = \{r_{i1}, \dots, r_{i,n_i}\}$ — множество правил с одинаковой левой частью A_i. Каждое правило r_{ij} из R_i имеет вид

$$r_{ij} : A_i \xrightarrow{p_{ij}} \beta_{ij}, j = 1, \dots, n_i,$$

где $A_i \in V_N$, $\beta_{ij} \in (V_T \cup V_N)^*$ и p_{ij} — вероятность применения правила r_{ij} (вероятность правила r_{ij}), которая удовлетворяет следующим условиям:

$$0 < p_{ij} \leqslant 1 \quad \text{и} \quad \sum_{j=1}^{n_i} p_{ij} = 1.$$

Для слов α и β из $(V_T \cup V_N)^*$ будем говорить, что β *непосредственно выводимо* из α (и записывать $\alpha \Rightarrow \beta$), если $\alpha = \alpha_1 A_i \alpha_2$, $\beta = \alpha_1 \beta_{ij} \alpha_2$ для некоторых $\alpha_1, \alpha_2 \in (V_T \cup V_N)^*$ и в грамматике G имеется правило $A_i \xrightarrow{p_{ij}} \beta_{ij}$.

Обозначим через $\Rightarrow_*$ рефлексивное транзитивное замыкание отношения $\Rightarrow$. Через L_G будем обозначать множество слов $\{\alpha \mid s \Rightarrow_* \alpha, \alpha \in V_T^*\}$.

Пусть $s \Rightarrow_* \alpha$. *Левым выводом* слова α будем называть вывод, при котором каждое правило в процессе вывода слова α из аксиомы s применяется к самому левому нетерминалу в слове. Последовательность правил в левом выводе будем обозначать через $\omega(\alpha)$. КС-грамматику будем называть *грамматикой с однозначным выводом,* если каждое слово языка имеет единственный левый вывод. В дальнейшем будем рассматривать грамматики с однозначным выводом.

Важное значение для нас имеет понятие дерева вывода [1]. Дерево строится следующим образом.

Корень дерева помечается аксиомой s. Пусть при выводе слова α на очередном шаге в процессе левого вывода применяется правило $A \xrightarrow{p_{ij}} b_{i_1} b_{i_2} \dots b_{i_m}$, где $b_{i_l} \in V_N \cup V_T$ $(l = 1, \dots, m)$. Тогда из самой левой вершины-листа дерева, помеченной символом A (при обходе листьев дерева слева направо), проводится m дуг в вершины следующего яруса, которые помечаются слева направо символами $b_{i_1}, b_{i_2}, \dots, b_{i_m}$ соответственно. После построения дуг и вершин для всех правил грамматики в выводе слова языка все листья дерева помечены терминальными символами и само слово получается при обходе листьев дерева слева направо.

Высотой дерева вывода будем называть максимальную длину пути от корня к листу.

Пример. Рассмотрим грамматику $G_0 = \langle\{x, \bar{x}\}, \{N\}, R, N\rangle$, в которой множество R состоит из двух правил:

$$r_1 : N \xrightarrow{p} xN\bar{x}N,$$
$$r_2 : N \xrightarrow{1-p} \lambda \ (\lambda \text{ – пустое слово}).$$

Грамматика G_0 порождает хорошо известный язык Дика. Если символ x интерпретировать как открывающую скобку «(», а символ $\bar{x}$ — как закрывающую скобку «)», то язык Дика — это множество «правильных» последовательностей скобок, обладающих следующими свойствами:

а) для любой начальной подпоследовательности число вхождений «(» не меньше числа вхождений «)»;

б) для всей последовательности число вхождений «(» равно числу вхождений «)».

Дерево вывода в грамматике G_0 изображено на рис. 1. Ему соответствуют левый вывод $r_1r_1r_2r_1r_2r_2r_1r_2r_2$ и слово $\alpha = xx\bar{x}x\bar{x}\bar{x}x\bar{x}$. Высота дерева вывода равна 4.

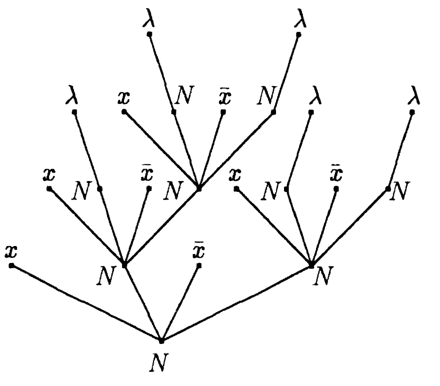

Рис. 1

Пусть $\omega(\alpha) = r_{i_1j_1}r_{i_2j_2}\ldots r_{i_nj_n}$ — левый вывод слова $\alpha \in L$. Для грамматики с однозначным выводом определим $p(\alpha)$ как произведение вероятностей правил, образующих левый вывод слова α: $p(\alpha) = p_{i_1j_1}p_{i_2j_2}\ldots p_{i_nj_n}$.

Грамматика G называется *согласованной*, если $\sum\limits_{\alpha\in L_G} p(\alpha) = 1$. В дальнейшем будем рассматривать согласованные КС-грамматики. Согласованная КС-грамматика G индуцирует распределение вероятностей P_G на множестве слов L_G.

Стохастический КС-язык, порожденный согласованной стохастической КС-грамматикой G, есть $\mathscr{L}_G = (L_G, P_G)$.

Стохастический язык $\mathscr{L}$ называется *стохастическим* КС-*языком*, если существует стохастическая КС-грамматика такая, что $\mathscr{L} = \mathscr{L}_G$.

В дальнейшем важное значение будет иметь матрица первых моментов, которая определяется следующим образом. Рассмотрим многомерные производящие функции

$$F_i(s_1, s_2, \dots, s_k),\ i = 1, \dots, k,$$

где переменная s_i соответствует нетерминальному символу A_i [7]. Функция $F_i(s_1, s_2, \dots, s_k)$ строится по множеству правил R_i с одинаковой левой частью A_i следующим образом.

Для каждого правила $A_i \xrightarrow{p_{ij}} \beta_{ij}$ выписывается слагаемое

$$q_{ij} = p_{ij} \cdot s_1^{l_1} \cdot s_2^{l_2} \cdot \ldots \cdot s_k^{l_k},$$

где l_m — число вхождений нетерминального символа A_m в правую часть правила ($m = 1, \dots, k$). Тогда

$$F_i(s_1, s_2, \dots, s_k) = \sum_{j=1}^{n_i} q_{ij}.$$

Пусть

$$a_{ij} = \frac{\partial F_i(s_1, \dots, s_k)}{\partial s_j} \Big|_{s_1 = s_2 = \ldots = s_k = 1}.$$

Квадратная матрица A порядка k, состоящая из элементов a_{ij}, называется *матрицей первых моментов* грамматики G. Так как матрица A неотрицательна, то существует максимальный по модулю действительный неотрицательный собственный корень (перронов корень) [2]. Обозначим этот корень через r.

В дальнейшем будем рассматривать такие грамматики с однозначным выводом, что матрица первых моментов каждой грамматики неразложима и непериодична [2] и $r < 1$. Неразложимость и непериодичность матрицы A означают, что существует натуральное n такое, что $A^n > 0$.

С помощью производящих функций определим также вторые моменты. Вторым моментом будем называть величину

$$b_{ijm} = \frac{\partial^2 F_i(s_1, \dots, s_k)}{\partial s_j \partial s_m} \Big|_{s_1 = s_2 = \ldots = s_k = 1} \quad (i, j, m \in \{1, 2, \dots, k\}).$$

Через D^t обозначим множество деревьев вывода высоты t для слов из $\mathscr{L}$ и через $\mathscr{L}^t$ — множество слов из $\mathscr{L}$, деревья вывода которых имеют высоту t. Для $\alpha \in \mathscr{L}^t$ через $p_t(\alpha)$ будем обозначать условную вероятность слова α: $p_t(\alpha) = \frac{p(\alpha)}{P(\mathscr{L}^t)}$, где $P(\mathscr{L}^t)$ — суммарная вероятность слов из $\mathscr{L}^t$.

Пусть $\mathscr{L} = (L, P)$ — стохастический КС-язык. *Кодированием* языка $\mathscr{L}$ будем называть инъективное отображение $f : \mathscr{L} \to \{0,1\}^+$. *Стоимостью* кодирования f назовем величину

$$C(\mathscr{L}, f) = \lim_{t\to\infty} \frac{\sum\limits_{\alpha\in\mathscr{L}^t} p_t(\alpha)\cdot|f(\alpha)|}{\sum\limits_{\alpha\in\mathscr{L}^t} p_t(\alpha)\cdot|\alpha|} \tag{1}$$

(здесь $|x|$ — длина последовательности x). Корректность определения величины $C(\mathscr{L}, f)$ нуждается в обосновании, так как предел может не существовать. Величина $C(\mathscr{L}, f)$ равна среднему числу двоичных символов, приходящихся на кодирование одного символа слова языка.

Через $F(\mathscr{L})$ обозначим класс всех инъективных отображений f из $\mathscr{L}$ в $\{0,1\}^+$, для которых существует $C(\mathscr{L}, f)$. Стоимостью оптимального кодирования языка $\mathscr{L}$ назовем величину

$$C_0(\mathscr{L}) = \inf_{f\in F(\mathscr{L})} C(\mathscr{L}, f).$$

2. Предварительные сведения о закономерностях применения правил стохастической КС-грамматики

Приведем результаты из [4], которые используются в дальнейшем для получения нижней оценки стоимости кодирования.

Пусть G — стохастическая КС-грамматика с однозначным выводом, матрица первых моментов которой неразложима, непериодична и перронов корень r строго меньше единицы. Пусть r_{ij} — правило грамматики G. Через $S_{ij}(t)$ обозначим число правил r_{ij} в левом выводе слова, которому соответствует дерево вывода из D^t ($S_{ij}(t)$ — случайная величина).

Рассмотрим величину $\frac{S_{ij}(t)}{t}$ — среднее число правил r_{ij}, приходящееся на один ярус дерева вывода из D^t. В [4] получена следующая оценка для математического ожидания случайной величины $\frac{S_{ij}(t)}{t}$:

$$M\left(\frac{S_{ij}(t)}{t}\right) = w_{ij} + O\left(\frac{\log^c t}{t}\right),$$

где $\log$ означает логарифм по основанию 2; c — некоторая константа; константа w_{ij} определяется равенством

$$w_{ij} = p_{ij}\left(\frac{v_i \sum\limits_{l=1}^{k} u_l s_l^{(ij)}}{r} + B_i\right), \tag{2}$$

в котором p_{ij} — вероятность правила r_{ij}; $U = (u_1, \dots, u_k)$ и $V = (v_1, \dots, v_k)$ — соответственно правый и левый неотрицательные собственные векторы для перронова корня при нормировке $\sum\limits_{i=1}^{k} u_i v_i = 1$;

$s_l^{(ij)}$ — число нетерминалов A_l в правой части правила r_{ij}; B_i — константа, определяемая формулой

$$B_i = \frac{1}{r} \sum_{l,m,n} v_l u_n b_{lmn} \sum_{\tau=1}^{\infty} a_{mi}(\tau - 1), \tag{3}$$

в которой $a_{mi}(\tau - 1)$ — элемент матрицы $A^{\tau-1}$.

Обозначим через $X_i(t)$ число вершин дерева вывода, помеченных нетерминальным символом A_i. Очевидно, что

$$X_i(t) = \sum_{j=1}^{n_i} S_{ij}(t).$$

В [4] установлено, что

$$M\Big(\frac{X_i(t)}{t}\Big) = w_i + O\Big(\frac{\log^c t}{t}\Big),$$

где константа w_i задается формулой $w_i = u_i v_i + B_i$, в которой величины u_i, v_i и B_i имеют те же значения, что и в определении величины w_{ij}.

Для дисперсии случайной величины $\frac{S_{ij}(t)}{t}$ в [4] получена следующая оценка:

$$D\Big(\frac{S_{ij}(t)}{t}\Big) = O\Big(\frac{\log^c t}{t}\Big) \quad \text{при} \quad t \to \infty.$$

Отсюда и из неравенства Чебышёва [8] следует, что для любого $\varepsilon > 0$ выполняется соотношение

$$P\Big(\Big|\frac{S_{ij}(t)}{t} - w_{ij}\Big| > \varepsilon\Big) = O\Big(\frac{\log^c t}{\varepsilon^2 t}\Big). \tag{4}$$

Так как мы рассматриваем грамматику с однозначным выводом, каждое слово языка имеет единственное дерево вывода. Из приведенных результатов следует, что множество всех слов языка с высотой дерева вывода t при $t \to \infty$ разбивается на две части: к первой относятся слова с суммарной вероятностью, стремящейся к нулю, ко второй — слова, имеющие приблизительно одинаковый состав правил грамматики в левом выводе и суммарная вероятность которых стремится к единице.

3. Нижняя оценка стоимости кодирования

Пусть $\mathscr{L}$ — стохастический КС-язык, порождаемый грамматикой G с однозначным выводом, матрица первых моментов которой неразложима, непериодична и ее перронов корень строго меньше единицы.

Для $\varepsilon > 0$ выделим множество таких слов из $\mathscr{L}^t$, что для любого правила r_{ij} грамматики G выполняется неравенство

$$|S_{ij}(t) - w_{ij} t| \leqslant \varepsilon t.$$

Множество таких слов обозначим через $M^t(\varepsilon)$.

Произведение $p_{11}^{w_{11}} \dots p_{1n_1}^{w_{1n_1}} \dots p_{k1}^{w_{k1}} \dots p_{kn_k}^{w_{kn_k}}$ назовем *типичной вероятностью* слова и обозначим через p_0, произведение $p_{11} \dots p_{1n_1} \dots p_{k1} \dots p_{kn_k}$ обозначим через p_1.

Для $\alpha \in M^t(\varepsilon)$ вероятность $p(\alpha)$ удовлетворяет следующему соотношению:

$$p_0^t \cdot p_1^{\varepsilon t} \leqslant p(\alpha) \leqslant p_0^t \cdot p_1^{-\varepsilon t}.$$

Ввиду (4) для суммарной вероятности $P(M^t(\varepsilon))$ справедлива оценка

$$P\left(M^t(\varepsilon)\right) = P\left(\mathscr{L}^t\right) + O\left(\frac{\log^c t}{\varepsilon^2 t}\right).$$

Из результатов в [7] может быть получена следующая оценка для $P\left(\mathscr{L}^t\right)$ в предположении, что нетерминальный символ A_i является аксиомой грамматики:

$$P\left(\mathscr{L}^t\right) = k_0 u_i (1-r) r^{t-1} + o\left(r^t\right),$$

где k_0 — некоторая константа и u_i — i-я компонента правого собственного вектора для перронова корня r.

Ясно, что число слов N в множестве $M^t(\varepsilon)$ удовлетворяет неравенствам

$$\frac{P\left(\mathscr{L}^t\right)\left(1 + O\left(\frac{\log^c t}{\varepsilon^2 t}\right)\right)}{p_0^t \cdot p_1^{-\varepsilon t}} \leqslant N \leqslant \frac{P\left(\mathscr{L}^t\right)\left(1 + O\left(\frac{\log^c t}{\varepsilon^2 t}\right)\right)}{p_0^t \cdot p_1^{\varepsilon t}}.$$

Рассмотрим способ кодирования слов из множества $\mathscr{L}^t$, состоящий в упорядочении слов в порядке невозрастания их вероятностей и кодировании слов по порядку сначала двоичными словами длины 1, затем двоичными словами длины 2 и т. д. Такое кодирование обозначим через f^*. Очевидно, что сумма $\sum\limits_{\alpha \in \mathscr{L}^t} p_t(\alpha) \cdot |f^*(\alpha)|$ является минимальной среди всех возможных кодирований множества слов $\mathscr{L}^t$. Поэтому для любого кодирования f на множестве всех слов языка $\mathscr{L}$, включающем множество $\mathscr{L}^t$, выполняется неравенство

$$\sum_{\alpha \in \mathscr{L}^t} p_t(\alpha) \cdot |f(\alpha)| \geqslant \sum_{\alpha \in \mathscr{L}^t} p_t(\alpha) \cdot |f^*(\alpha)|. \tag{5}$$

В [5] доказана нижняя оценка для стоимости кодирования f^* на конечном множестве элементов с заданным на нем распределением вероятностей. Применяя эту оценку к множеству $M^t(\varepsilon)$, можно записать следующее неравенство:

$$\sum_{\alpha \in M^t(\varepsilon)} p_\varepsilon(\alpha) \cdot |f^*(\alpha)| \geqslant H\left(M^t(\varepsilon)\right) - \log\log N - C, \tag{6}$$

где $p_\varepsilon(\alpha)$ — условная вероятность слова α в множестве $M^t(\varepsilon)$, т. е.

$$p_\varepsilon(\alpha) = \frac{p(\alpha)}{P(M^t(\varepsilon))} = p_t(\alpha) \cdot \frac{P(\mathscr{L}^t)}{P(M^t(\varepsilon))},$$

$H(M^t(\varepsilon)) = -\sum\limits_{\alpha \in M^t(\varepsilon)} p_\varepsilon(\alpha) \cdot \log p_\varepsilon(\alpha)$, N — число слов в множестве $M^t(\varepsilon)$ и C — некоторая константа.

Используя неравенство (6), из (5) получаем

$$\begin{aligned}
&\sum_{\alpha \in \mathscr{L}^t} p_t(\alpha) \cdot |f(\alpha)| \geqslant \sum_{\alpha \in M^t(\varepsilon)} p_t(\alpha) \cdot |f^*(\alpha)| \\
&= \sum_{\alpha \in M^t(\varepsilon)} \frac{p_\varepsilon(\alpha) \cdot P(M^t(\varepsilon))}{P(\mathscr{L}^t)} \cdot |f^*(\alpha)| = \frac{P(M^t(\varepsilon))}{P(\mathscr{L}^t)} \cdot \sum_{\alpha \in M^t(\varepsilon)} p_\varepsilon(\alpha) \cdot |f^*(\alpha)| \\
&\geqslant \frac{P(M^t(\varepsilon))}{P(\mathscr{L}^t)} \cdot \left(H\left(M^t(\varepsilon)\right) - \log\log N - C\right) \\
&\geqslant \frac{P(M^t(\varepsilon))}{P(\mathscr{L}^t)} \cdot \left\{-\log\left(\frac{p_0^t p_1^{-\varepsilon t}}{P(M^t(\varepsilon))}\right) - \log\log \frac{P(M^t(\varepsilon))}{p_0^t p_1^{\varepsilon t}} - C\right\} \\
&= \left(1 + O\left(\frac{\log^c t}{\varepsilon^2 t}\right)\right) \cdot \left\{t \cdot (-\log p_0 + \varepsilon \log p_1) + (t-1)\log r + O(\log t)\right\} \\
&= \left(1 + O\left(\frac{\log^c t}{\varepsilon^2 t}\right)\right) \cdot \left\{t \cdot (\log r - \log p_0 + \varepsilon \log p_1) + O(\log t)\right\} \\
&= t(\log r - \log p_0 + \varepsilon \log p_1) + O\left(\frac{\log^c t}{\varepsilon^2}\right).
\end{aligned}$$

Подсчитаем величину $\sum\limits_{\alpha \in \mathscr{L}^t} p_t(\alpha) \cdot |\alpha|$. Пусть правило r_{ij} содержит l_{ij} терминальных символов. Тогда

$$|\alpha| = w_{11}(\alpha) \cdot l_{11} + \ldots + w_{1n_1}(\alpha) \cdot l_{1n_1} \ldots + w_{k1}(\alpha) \cdot l_{k1} + \ldots + w_{kn_k}(\alpha) \cdot l_{kn_k},$$

где $w_{ij}(\alpha)$ — число применений правила r_{ij} в выводе слова α ($i = 1, \ldots, k$; $j = 1, \ldots, n_i$). Поэтому

$$\sum_{\alpha \in \mathscr{L}^t} p_t(\alpha) \cdot |\alpha| = \sum_{i,j} l_{ij} M(S_{ij}(t)) = t \cdot \sum_{i,j} l_{ij} w_{ij} + t \cdot O\left(\frac{\log^c t}{t}\right).$$

Пусть $h = \sum\limits_{i,j} l_{ij} w_{ij}$. Величина h характеризует среднее число терминальных символов на одном ярусе дерева вывода. Тогда имеем

$$\sum_{\alpha \in \mathscr{L}^t} p_t(\alpha) \cdot |\alpha| = th + O\left(\log^c t\right).$$

Суммируя полученные оценки, получаем

$$
\begin{aligned}
C(\mathscr{L}, f) &= \lim_{t\to\infty} \frac{\sum\limits_{\alpha\in\mathscr{L}^t} p_t(\alpha)\cdot|f(\alpha)|}{th + O\left(\log^c t\right)} \\
&\geqslant \lim_{t\to\infty} \left(\frac{\log r - \log p_0 + \varepsilon \log p_1}{h} + O\Big(\frac{\log^c t}{t}\Big)\right) \\
&= \frac{\log r - \log p_0 + \varepsilon \log p_1}{h} = \frac{\log r - \log p_0}{h} + \varepsilon \cdot \frac{\log p_1}{h}.
\end{aligned}
$$

Так как полученное неравенство справедливо при любом $\varepsilon > 0$, то

$$
C(\mathscr{L}, f) \geqslant \frac{\log r - \log p_0}{h}. \tag{7}
$$

В свою очередь,

$$
\begin{aligned}
\log p_0 &= \log\left(p_{11}^{w_{11}} \dots p_{1n_1}^{w_{1n_1}} \dots p_{k1}^{w_{k1}} \dots p_{kn_k}^{w_{kn_k}}\right) \\
&= \sum_{i,j} w_{ij} \log p_{ij} = \sum_{i,j} p_{ij} \log p_{ij} \cdot \left(\frac{v_i \sum\limits_l u_l s_l^{(ij)}}{r} + B_i\right) \\
&= \sum_{i=1}^{k} B_i \cdot \sum_j p_{ij} \log p_{ij} + \frac{1}{r} \sum_{i=1}^{k} v_i \sum_{j=1}^{n_i} p_{ij} \log p_{ij} \sum_{l=1}^{n_i} u_l s_l^{(ij)}.
\end{aligned}
$$

Подставив полученное выражение в (7) и используя обозначение $H(p_{i1}, \dots, p_{in_i}) = -\sum\limits_{ij} p_{ij} \log p_{ij}$, получаем

$$
C(\mathscr{L}, f) \geqslant \frac{\log r}{h} + \frac{1}{h} \sum_{i=1}^{k} B_i \cdot H(p_{i1}, \dots, p_{in_i}) - \frac{1}{rh} \cdot \sum_{i=1}^{k} v_i \sum_{j=1}^{n_i} p_{ij} \log p_{ij} \sum_{l=1}^{k} u_l s_l^{(ij)}.
$$

Таким образом, мы установили нижнюю оценку для стоимости произвольного кодирования $f \in F(\mathscr{L})$. Сформулируем полученный результат в виде следующего утверждения.

Теорема 1. *Пусть $\mathscr{L}$ — язык, порожденный стохастической КС-грамматикой с однозначным выводом, матрица первых моментов которой неразложима, непериодична и перронов корень строго меньше единицы. Тогда для любого кодирования $f \in F(\mathscr{L})$ стоимость кодирования $C(\mathscr{L}, f)$ удовлетворяет неравенству*

$$
C(\mathscr{L}, f) \geqslant \frac{\log r}{h} + \frac{1}{h} \sum_{i=1}^{k} B_i \cdot H(p_{i1}, \dots, p_{in_i}) - \frac{1}{rh} \cdot \sum_{i=1}^{k} v_i \sum_{j=1}^{n_i} p_{ij} \log p_{ij} \sum_{l=1}^{k} u_l s_l^{(ij)},
$$

где p_{ij} — вероятность правила r_{ij}; $U = (u_1, \dots, u_k)$ и $V = (v_1, \dots, v_k)$ — соответственно правый и левый неотрицательные собственные векторы

для перронова корня r *при нормировке* $\sum\limits_{i=1}^{k} u_i v_i = 1$; $s_l^{(ij)}$ *— число нетерминалов* A_l *в правой части правила* r_{ij}; B_i *— константа, определяемая формулой (3);* h *— предел математического ожидания среднего числа терминальных символов на одном ярусе дерева вывода при* $t \to \infty$.

Полученную нижнюю оценку стоимости кодирования обозначим через $C^*(\mathscr{L})$. В дальнейшем величину $C^*(\mathscr{L})$ будем также представлять в следующем виде:

$$C^*(\mathscr{L}) = \frac{\log r}{h} - \frac{1}{h}\sum_{i,j} w_{ij} \log p_{ij}.$$

4. Неулучшаемость нижней оценки стоимости кодирования и асимптотически оптимальное кодирование

Докажем, что нижняя оценка стоимости кодирования по теореме 1 является неулучшаемой, т. е. справедливо равенство $C^*(\mathscr{L}) = C_0(\mathscr{L})$.

Определим частоту p'_{ij} применения правила r_{ij} среди правил грамматики с одинаковой левой частью A_i:

$$p'_{ij} = \frac{w_{ij}}{w_i}, \quad \text{где} \quad w_i = \sum_{j=1}^{n_j} w_{ij}.$$

Для каждого множества правил R_i с одинаковой левой частью A_i построим схему двоичного префиксного кодирования по алгоритму Шеннона [10]. При этом правилу r_{ij} будет соответствовать элементарный код v_{ij} длины $\lceil -\log p'_{ij} \rceil$.

Пусть $\alpha \in \mathscr{L}^t$ имеет левый вывод $w(\alpha) = r_{i_1 j_1} r_{i_2 j_2} \ldots r_{i_m j_m}$. Тогда в качестве кода слова α будем рассматривать двоичную последовательность $v_{i_1 j_1} v_{i_2 j_2} \ldots v_{i_m j_m}$, полученную конкатенацией элементарных кодов правил, образующих левый вывод.

Предложенный алгоритм кодирования строит локально-префиксный код [6] на множестве левых выводов слов из $\mathscr{L}$, когда в качестве алфавита рассматривается множество правил R исходной грамматики G.

Построенное кодирование обозначим через f_{sh}. Определим стоимость кодирования для f_{sh} :

$$C(\mathscr{L}, f_{sh}) = \lim_{t\to\infty} \frac{\sum_{\alpha \in \mathscr{L}^t} p_t(\alpha) \cdot |f_{sh}(\alpha)|}{th + O(\log^c t)}.$$

Выражение в числителе дроби представим в виде

$$\sum_{\alpha \in \mathscr{L}^t} p_t(\alpha) \cdot |f_{sh}(\alpha)| = M\Big(\sum_{i,j} w_{ij}(\alpha) \cdot \lceil -\log p'_{ij} \rceil\Big)$$
$$= \sum_{i,j} \lceil -\log p'_{ij} \rceil M\left(w_{ij}(\alpha)\right) = t\sum_{i,j} w_{ij} \lceil -\log p'_{ij} \rceil + O\left(\log^c t\right)$$

(здесь $w_{ij}(\alpha)$ — число вхождений правила r_{ij} в $w(\alpha)$). Поэтому

$$C(\mathscr{L}, f_{sh}) = \lim_{t\to\infty} \frac{t\sum_{i,j} w_{ij}\lceil -\log p'_{ij}\rceil + O(\log^c t)}{th + O(\log^c t)}$$
$$= \frac{1}{h}\sum_{i,j} w_{ij}\lceil -\log p'_{ij}\rceil = \frac{1}{h}\sum_{i,j} w_{ij}\left\lceil -\log \frac{w_{ij}}{w_i}\right\rceil.$$

Оценим сверху разность $\Delta = C(\mathscr{L}, f_{sh}) - C^*(\mathscr{L})$:

$$\Delta = \frac{1}{h}\sum_{i,j} w_{ij}\left\lceil -\log\frac{w_{ij}}{w_i}\right\rceil - \frac{\log r}{h} + \frac{1}{h}\sum_{i,j} w_{ij}\log p_{ij}$$
$$\leqslant \frac{1}{h}\sum_{i,j} w_{ij}\left(-\log\frac{w_{ij}}{w_i} + 1\right) - \frac{\log r}{h} + \frac{1}{h}\sum_{i,j} w_{ij}\log p_{ij}$$
$$= \frac{\sum\limits_{i=1}^{k} w_i}{h} + \frac{1}{h}\sum_{i,j} w_{ij}\log\frac{w_i}{w_{ij}} - \frac{\log r}{h} + \frac{1}{h}\sum_{i,j} w_{ij}\log p_{ij}$$
$$= \frac{\sum\limits_{i=1}^{k} w_i}{h} - \frac{\log r}{h} + \frac{1}{h}\sum_{i,j} w_{ij}\log\frac{w_i\cdot p_{ij}}{w_{ij}}.$$

Далее $\sum\limits_{i=1}^{k} w_i$ будем обозначать через w.

Множество правил R_i с одинаковой левой частью A_i разобьем на два подмножества: множество $R_i^{\text{н}}$ незаключительных правил (т. е. содержащих в правой части нетерминальные символы) и множество $R_i^{\text{з}}$ заключительных правил (не содержащих в правой части нетерминальных символов). Тогда

$$\Delta \leqslant \frac{w}{h} - \frac{\log r}{h} + \frac{1}{h}\sum_i\sum_{R_i^{\text{н}}} w_{ij}\log\frac{w_i\cdot p_{ij}}{w_{ij}} + \frac{1}{h}\sum_i\sum_{R_i^{\text{з}}} w_{ij}\log\frac{w_i\cdot p_{ij}}{w_{ij}}.$$

Раскроем w_{ij}, используя (2). Предварительно заметим, что $s_l^{(ij)} = 0$ для любого заключительного правила r_{ij} и любого l; поэтому $\sum\limits_{l=1}^{k} u_l s_l^{(ij)} = 0$ и $w_{ij} = p_{ij}B_i$. Будем применять обозначение $\tilde{s}_{ij} = \sum\limits_{l=1}^{k} u_l s_l^{(ij)}$. Раскрывая

w_{ij}, получаем

$$\Delta \leqslant \frac{w}{h} - \frac{\log r}{h} + \frac{1}{h}\sum_{i=1}^{k}\sum_{R_i^{\text{н}}} w_{ij}\log w_i$$

$$+ \frac{1}{h}\sum_{i=1}^{k}\sum_{R_i^{\text{н}}} w_{ij}\log\frac{r}{v_i\tilde{s}_{ij} + B_i r} + \frac{1}{h}\sum_{i=1}^{k}\sum_{R_i^{\text{з}}} p_{ij}B_i\log\frac{w_i}{B_i}$$

$$\leqslant \frac{w}{h} - \frac{\log r}{h} + \frac{1}{h}\sum_{i=1}^{k}\sum_{R_i^{\text{н}}} w_{ij}\log w_i + \frac{\log r}{h}\sum_{i=1}^{k}\sum_{R_i^{\text{н}}} w_{ij}$$

$$+ \frac{1}{h}\sum_{i=1}^{k}\sum_{R_i^{\text{н}}} w_{ij}\log\frac{1}{v_i\tilde{s}_{ij} + B_i r} + \frac{1}{h}\sum_{i=1}^{k}\sum_{R_i^{\text{з}}} p_{ij}B_i\log\frac{w_i}{B_i}.$$

Используем (2) и неравенство $\log x \leqslant \log e \cdot x$. Тогда

$$\Delta \leqslant \frac{w}{h} - \frac{\log r}{h} + \frac{1}{h}\sum_{i=1}^{k} w_i\log w_i + \frac{\log r}{h}\sum_{i=1}^{k}\sum_{R_i^{\text{н}}} w_{ij}$$

$$+ \frac{\log e}{h}\sum_{i=1}^{k}\sum_{R_i^{\text{н}}} w_{ij}\left(\frac{1}{v_i\tilde{s}_{ij} + B_i r}\right) + \frac{\log e}{h}\sum_{i=1}^{k}\sum_{R_i^{\text{з}}} p_{ij}$$

$$= \frac{w}{h} - \frac{\log r}{h} + \frac{1}{h}\sum_{i=1}^{k} w_i\log w_i + \frac{\log r}{h}\sum_{i=1}^{k}\sum_{R_i^{\text{н}}} w_{ij}$$

$$+ \frac{\log e}{h}\sum_{i=1}^{k}\sum_{R_i^{\text{н}}} \frac{p_{ij}}{r} + \frac{\log e}{h}\sum_{i=1}^{k}\sum_{R_i^{\text{з}}} p_{ij}.$$

Опишем используемый в дальнейшем способ перехода от исходной грамматики G к грамматике $G(n)$, состоящий в укрупнении правил грамматики. Пусть $A_i \Rightarrow_* \alpha$. Через $d(\alpha)$ обозначим высоту дерева вывода слова α. Через M_i^n обозначим множество слов в алфавите $\{V_N \cup V_T\}$, выводимых из A_i, для которых высота дерева вывода не превосходит n, и нетерминалами могут быть помечены листья только n-го (последнего) яруса дерева. В [3] доказано, что

$$\sum_{\alpha\in M_i^n} p(\alpha) = 1.$$

Используя множества M_i^n, можно перейти от грамматики G к грамматике $G(n)$ с множеством правил $R(n) = \bigcup\limits_{i=1}^{k} R_i(n)$ и

$$R_i(n) = \{A_i \xrightarrow{p'_{ij}} \alpha'_{ij} \mid \alpha'_{ij} \in M_i^n\},$$

где $j = 1, \dots, n_i'$ и число правил n_i' в $R_i(n)$ равно числу слов в M_i^n.

Каждому правилу в $R_i(n)$ приписывается вероятность p_{ij}', равная вероятности вывода слова α_{ij}' из A_i в исходной грамматике G. Нетрудно заметить, что $\mathscr{L}_G = \mathscr{L}_{G(n)}$ и $G(n)$ — грамматика с однозначным выводом.

Для $G(n)$ матрица первых моментов совпадает с n-й степенью матрицы A для G, т. е. равна A^n [7]. Для множества правил $R_i(n)$ суммарная вероятность незаключительных правил при $n \to \infty$ имеет следующий вид:

$$P\left(R_i(n)^{\text{н}}\right) = k_0 u_i r^n + o\left(r^n\right),$$

где k_0 — некоторая константа и u_i — координата правого собственного вектора для перронова корня r. Этот результат легко может быть получен интерпретацией результатов теории ветвящихся процессов [7] применительно к процессу порождения слов языка. Вероятности множества правил $R_i(n)^{\text{н}}$ соответствует вероятность продолжения соответствующего ветвящегося процесса в момент времени n.

Таким образом, $P(R_i(n)^{\text{н}}) = O(r^n)$ и, следовательно, $P\left(R_i(n)^{\text{з}}\right) = 1 + O\left(r^n\right)$. Используем эти оценки для (8). Так как одному ярусу дерева вывода в грамматике $G(n)$ соответствует n ярусов дерева вывода в исходной грамматике G, то при переходе к грамматике $G(n)$ величина h переходит в nh. Перронов корень матрицы первых моментов A^n для грамматики $G(n)$ равен r^n [7], а значения величин w и w_i при переходе от грамматики G к грамматике $G(n)$ не изменяются.

Обозначим через $\Delta(n)$ избыточность кодирования при использовании грамматики $G(n)$. Тогда

$$\Delta(n) \leqslant \frac{w}{nh} - \frac{n\log r}{nh} + \frac{1}{nh}\sum_{i=1}^{k} w_i \log w_i + \frac{n\log r}{nh}\sum_{i=1}^{k}\sum_{R_i^{\text{н}}} w_{ij} + \frac{\log e}{nh}\sum_{i=1}^{k}\sum_{R_i^{\text{н}}}\frac{p_{ij}}{r^n} + \frac{\log e}{nh}\sum_{i=1}^{k}\sum_{R_i^{\text{з}}} p_{ij} \qquad (9)$$

(здесь величины w_{ij}, p_{ij} и множества $R_i^{\text{н}}$, $R_i^{\text{з}}$ соответствуют новой грамматике $G(n)$).

Оценим в (9) величины $\delta_1 = \sum\limits_{R_i^{\text{н}}} w_{ij}$, $\delta_2 = \sum\limits_{R_i^{\text{н}}} p_{ij}$ и $\delta_3 = \sum\limits_{i=1}^{k} \sum_{R_i^{\text{з}}} p_{ij}$:

$$\delta_1 = \sum_{R_i^{\text{н}}} w_{ij} = w_i - \sum_{R_i^{\text{з}}} w_{ij} = w_i - \sum_{R_i^{\text{з}}} p_{ij} B_i$$
$$= w_i - B_i\left(1 + O\left(r^n\right)\right) = u_i v_i + O\left(r^n\right),$$

$$\delta_2 = \sum_{R_i^{\text{н}}} p_{ij} = O\left(r^n\right), \quad \delta_3 = \sum_{i=1}^{k} \sum_{R_i^{\text{э}}} p_{ij} = k + O\left(r^n\right).$$

(При оценке δ_1 мы использовали равенство $w_i = u_i v_i + B_i$.) Подставив δ_1, δ_2 и δ_3 в (9), получаем

$$\Delta(n) \leqslant \frac{w}{nh} - \frac{n\log r}{nh} + \frac{1}{nh}\sum_{i=1}^{k} w_i \log w_i$$
$$+ \frac{n\log r}{nh} \cdot \sum_{i=1}^{k} \left(u_i v_i + O\left(r^n\right)\right) + \frac{\log e}{nhr^n}\sum_{i=1}^{k} O\left(r^n\right) + \frac{\log e}{nh}\left(k + O\left(r^n\right)\right).$$

Так как $\sum\limits_{i=1}^{k} u_i v_i = 1$, то

$$\Delta(n) \leqslant \frac{w}{nh} - \frac{\log r}{h} + \frac{1}{nh}\sum_{i=1}^{k} w_i \log w_i + \frac{\log r}{h} \cdot \left(1 + O\left(r^n\right)\right)$$
$$+ \frac{\log e}{nh} \cdot O(1) = O\left(\frac{1}{n}\right).$$

Поэтому $\Delta(n) \to 0$ при $n \to \infty$.

Таким образом, $C^*\left(\mathscr{L}\right) = C_0\left(\mathscr{L}\right)$ и справедлива следующая

Теорема 2. *Пусть $\mathscr{L}$ — язык, порожденный стохастической КС-грамматикой G с однозначным выводом, матрица первых моментов которой неразложима, непериодична и перронов корень строго меньше единицы. Тогда существует последовательность кодирований $\{f_n \mid f_n \in F(\mathscr{L}), n = 1, 2, \ldots\}$ такая, что*

$$C\left(\mathscr{L}, f_n\right) - C^*\left(\mathscr{L}\right) \to 0 \quad \text{при} \quad n \to \infty.$$

Доказательство дает алгоритм асимптотически оптимального кодирования, который состоит в переходе от исходной грамматики G к грамматике $G(n)$ с «укрупненными» правилами и в применении алгоритма алфавитного кодирования Шеннона к каждому подмножеству $R_i(n)$ правил с одинаковым нетерминалом A_i в левой части правил. Заметим, что для более быстрой сходимости стоимости кодирования к нижней оценке вместо алгоритма Шеннона можно использовать известный алгоритм Хаффмена (см. [12]), который дает минимальное значение стоимости алфавитного кодирования.

Для кодирования слова α языка достаточно построить левый вывод этого слова в грамматике и затем каждое правило в выводе заменить его элементарным кодом в соответствии с построенной схемой локально-префиксного кодирования.

Проиллюстрируем алгоритм асимптотически оптимального кодирования на примере рассмотренной выше грамматики G_0. Для грамматики G_0 с двумя правилами

$$r_1 : N \xrightarrow{p} xN\bar{x}N,$$
$$r_2 : N \xrightarrow{1-p} \lambda \ (\lambda - \text{пустое слово})$$

выпишем производящую функцию:

$$F_1(s_1) = p \cdot s_1^2 + (1-p) \cdot s_1^0 = p \cdot s_1^2 + (1-p).$$

Определим первый и второй моменты a_{11} и b_{111} соответственно:

$$a_{11} = \frac{\partial F_1(s_1)}{\partial s_1} \Big|_{s_1=1} = 2p,$$
$$b_{111} = \frac{\partial^2 F_1(s_1)}{\partial s_1^2} \Big|_{s_1=1} = 2p.$$

Так как матрица первых моментов для грамматики с одним нетерминальным символом состоит из одного элемента, то его значение является перроновым корнем. Поэтому $r = a_{11} = 2p$. Следовательно, $r < 1$ при $p < 1/2$. Очевидно, что для грамматики с одним нетерминалом $u_1 = v_1 = 1$. Отсюда и из (3) следует, что

$$B_1 = \frac{1}{2p} \cdot b_{111} \sum_{\tau=1}^{\infty} (2p)^{\tau-1} = \frac{1}{1-2p}.$$

Найдем значения величин w_{11} и w_{12} по формуле (2), учитывая, что $\tilde{s}_1 = 2$ и $\tilde{s}_2 = 0$, так как правило r_1 содержит два нетерминала в правой части, а r_2 не содержит нетерминалов в правой части:

$$w_{11} = p \cdot \left(\frac{\tilde{s}_1}{r} + B_1\right) = p \cdot \left(\frac{2}{2p} + \frac{1}{1-2p}\right) = 1 + \frac{p}{1-2p};$$
$$w_{12} = (1-p) \cdot \left(\frac{\tilde{s}_2}{r} + B_1\right) = (1-p) \cdot \left(0 + \frac{1}{1-2p}\right) = 1 + \frac{p}{1-2p}.$$

Заметим, что w_{11} равно w_{12} независимо от значения вероятности p, т. е. правила r_1 и r_2 применяются с одинаковой частотой в выводах слов, имеющих деревья вывода большой высоты.

Для математического ожидания среднего числа правил, приходящегося на один ярус дерева вывода, при $t \to \infty$ имеем следующее значение:

$$w_1 = w_{11} + w_{12} = 2 \cdot \left(1 + \frac{p}{1-2p}\right) = 1 + \frac{1}{1-2p}.$$

Определим значение величины h, учитывая, что число l_1 терминальных символов в правой части правила r_1 равно двум и l_2 равно нулю для правила r_2 :

$$h = l_1 w_{11} + l_2 w_{12} = 2\Big(1 + \frac{p}{1-2p}\Big) = 1 + \frac{1}{1-2p}.$$

Определим стоимость оптимального кодирования для языка $\mathscr{L}_{G_0}$:

$$C_0(\mathscr{L}_{G_0}) = \frac{\log r}{h} - \frac{1}{h}\left(w_{11}\log p + w_{12}\log(1-p)\right)$$
$$= \frac{1+\log p}{h} - \frac{1}{2}(\log p + \log(1-p)) = \frac{1-2p}{2(1-p)} + \frac{1}{2(1-p)} \cdot H(p, 1-p),$$

где $H(p,1-p) = -(p\log p + (1-p)\log(1-p))$. Отметим, что $C_0(\mathscr{L}_{G_0}) \to 1$ при $p \to \frac{1}{2}$ и $C_0(\mathscr{L}_{G_0}) \to \frac{1}{2}$ при $p \to 0$.

Пусть $p = 1/8$. Тогда $r = 1/4$ и $C_0(\mathscr{L}_{G_0}) = \frac{3}{7} + \frac{4}{7} \cdot H\left(\frac{1}{8} + \frac{7}{8}\right) \approx 0,739$.

Применим алгоритм асимптотически оптимального кодирования для $n = 1$ и $n = 2$. Очевидно, что $C(\mathscr{L}_{G_0}, f_{sh}) = 1$ при $n = 1$ и одно из правил следует кодировать символом 0, а другое — символом 1.

Пусть $n = 2$. Построим правила грамматики $G_0(2)$:

$$r_1 : N \xrightarrow{p_1} \lambda,\ p_1 = \frac{7}{8} = 0,875,$$
$$r_2 : N \xrightarrow{p_2} x\bar{x},\ p_2 = \frac{1}{8} \cdot \left(\frac{7}{8}\right)^2 \approx 0,95700,$$
$$r_3 : N \xrightarrow{p_3} xxN\bar{x}N\bar{x},\ p_3 = \left(\frac{1}{8}\right)^2 \cdot \frac{7}{8} \approx 0,01367,$$
$$r_4 : N \xrightarrow{p_4} x\bar{x}xN\bar{x}N,\ p_4 = \left(\frac{1}{8}\right)^2 \cdot \frac{7}{8} \approx 0,01367,$$
$$r_5 : N \xrightarrow{p_5} xxN\bar{x}N\bar{x}xN\bar{x}N,\ p_5 = \left(\frac{1}{8}\right)^3 \approx 0,00195.$$

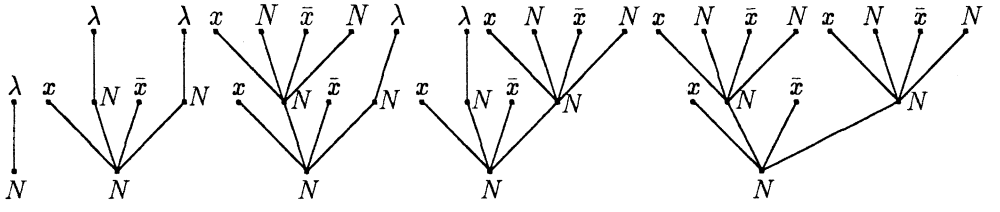

Рис. 2

Деревья вывода в грамматике G_0, соответствующие правилам грамматики $G_0(2)$, изображены на рис. 2.

Перронов корень для грамматики $G_0(2)$ равен $r^2 = (\frac{1}{4})^2 = \frac{1}{16}$. Для грамматики $G_0(2)$ производящей является функция

$$F_1(s_1) = p_1 + p_2 + p_3 \cdot s_1^2 + p_4 \cdot s_1^2 + p_5 \cdot s_1^4.$$

Далее имеем

$$b_{111} = \frac{\partial^2 F_1(s_1)}{\partial s_1^2}\Big|_{s_1=1} = 2p_3 + 2p_4 + 12p_5 = \frac{5}{64},$$

$$B_1 = \frac{1}{r} \cdot b_{111} \cdot \sum_{\tau=1}^{\infty} r^{\tau-1} = \frac{5}{64 \cdot \frac{1}{16} \cdot \frac{15}{16}} = \frac{4}{3}.$$

Определим значения w_{11}, w_{12}, w_{13}, w_{14} и w_{15}, используя формулу (2):

$$w_{11} \approx 1,1666, \quad w_{12} \approx 0,1276, \quad w_{13} \approx 0,4557, \quad w_{14} \approx 0,4557, \quad w_{15} \approx 0,1276.$$

Следовательно,

$$w_1 = \sum_{j=1}^{5} w_{1j} \approx 2,3332.$$

Подсчитаем частоты применения правил грамматики $G_0(2)$:

$$p'_1 = \frac{w_{11}}{w_1} = 0,5,$$
$$p'_2 = \frac{w_{12}}{w_1} \approx 0,0547,$$
$$p'_3 = \frac{w_{13}}{w_1} \approx 0,1953,$$
$$p'_4 = \frac{w_{14}}{w_1} \approx 0,1953,$$
$$p'_5 = \frac{w_{15}}{w_1} \approx 0,0547.$$

Применив алгоритм Хаффмена к полученным частотам, получим следующую схему кодирования f:

$$v_1 = 0, \quad v_2 = 1010, \quad v_3 = 11, \quad v_4 = 100, \quad v_5 = 1011.$$

Здесь v_i — элементарный код для правила r_i $(i = 1, \ldots, 5)$.

Определим стоимость кодирования. После несложных преобразований получим

$$C\left(\mathscr{L}_{G_0(2)}, f\right) \approx 0,957.$$

Таким образом, при $n = 2$ мы получили меньшее значение стоимости кодирования, чем при $n = 1$.

Заключение

Несколько слов о временной сложности построенного алгоритма асимптотически оптимального кодирования. Сложность этого алгоритма можно характеризовать с двух сторон. Во-первых, можно рассматривать сложность построения схемы локально-префиксного кодирования по заданной грамматике. Во-вторых, алгоритм можно характеризовать сложностью кодирования и декодирования сообщения, являющегося словом языка из рассматриваемого класса.

Известно, что временная сложность алгоритма Хаффмена не более чем квадратично зависит от числа букв в алфавите. Если КС-грамматика содержит k нетерминалов, то алгоритм Хаффмена применяется k раз, отдельно для каждого множества правил R_i $(i = 1, 2, \dots, k)$.

Пусть l_i — число правил в множестве R_i и $l = \max\{l_1, l_2, \dots, l_k\}$. Тогда алгоритм построения локально-префиксного кода по грамматике имеет временную сложность $O(kl^2)$.

При переходе к грамматике $G(n)$ число правил в множестве $R_i(n)$ может расти экспоненциально. Однако строить схему локально-префиксного кодирования для ее последующего использования придется один раз.

Рассмотрим временную сложность кодирования сообщения α, являющегося словом КС-языка. Алгоритм кодирования сообщения можно разбить на два этапа.

Этап 1. Построение левого вывода слова α в грамматике. Для КС-грамматики с однозначным выводом левый вывод строится за время $O(|\alpha|^2)$, где $|\alpha|$ — длина слова α [11].

Этап 2. Кодирование левого вывода в соответствии со схемой локально-префиксного кодирования. Временная сложность кодирования на этапе 2 имеет линейный порядок от длины вывода. Так как длина вывода для слова в случае КС-грамматики с однозначным выводом имеет порядок $O(|\alpha|)$, выполнение этапа 2 требует не более $O(|\alpha|)$ операций.

Суммарная временная сложность алгоритма асимптотически оптимального кодирования сообщения длины m равна $O(m^2)$.

Нетрудно показать, что алгоритм декодирования для асимптотически оптимального кодирования также имеет квадратичную временную сложность от длины сообщения.

ЛИТЕРАТУРА

1. **Ахо А., Ульман Дж.** Теория синтаксического анализа, перевода и компиляции. Т. 1. М.: Мир, 1978.
2. **Гантмахер Ф. Р.** Теория матриц. М.: Наука, 1967.

3. **Жильцова Л. П.** Кодирование стохастических контекстно-свободных языков с однозначным выводом // Дискретная математика. 1994. Т. 6, вып. 3. С. 73–88.
4. **Жильцова Л. П.** Закономерности применения правил грамматики в выводах слов стохастического контекстно-свободного языка // Математические вопросы кибернетики. Вып. 9. М.: Наука, 2000. С. 101–126.
5. **Кричевский Р. Е.** Сжатие и поиск информации. М.: Радио и связь, 1989.
6. **Марков А. А.** Введение в теорию кодирования. М.: Наука, 1982.
7. **Севастьянов В. А.** Ветвящиеся процессы. М.: Наука, 1971.
8. **Феллер В.** Введение в теорию вероятностей и ее приложения. Т. 1. М.: Мир, 1984.
9. **Фу К.** Структурные методы в распознавании образов. М.: Мир, 1977.
10. **Шеннон К.** Математическая теория связи // Работы по теории информации и кибернетике. М.: Изд-во иностр. лит., 1963. С. 243–333.
11. **Эрли Дж.** Эффективный алгоритм анализа контекстно-свободных языков // Языки и автоматы. М.: Мир, 1975. С. 47–70.
12. **Яблонский С. В.** Введение в дискретную математику. М.: Наука, 1986.

Адрес автора:

Нижегородский
государственный
педагогический университет,
ул. Ульянова, 1,
603005 Нижний Новгород,
Россия

Статья поступила
6 июня 2001 г.