ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 56 **2020** **Вып. 4**

УДК 621.391 : 519.728

© 2020 г. **Л.А. Шоломов**

ПОЛИНОМИАЛЬНОЕ АСИМПТОТИЧЕСКИ ОПТИМАЛЬНОЕ КОДИРОВАНИЕ НЕДООПРЕДЕЛЕННЫХ БЕРНУЛЛИЕВСКИХ ИСТОЧНИКОВ ОБЩЕГО ВИДА

Недоопределенный источник Бернулли порождает независимо с некоторыми вероятностями символы заданного недоопределенного алфавита. Каждому недоопределенному символу соответствует некоторое множество основных (полностью определенных) символов, любым из которых он может быть замещен (доопределен). Недоопределенный источник характеризуется энтропией, которая вводится неявно как минимум некоторой функции и играет роль, подобную роли энтропии Шеннона для полностью определенных источников. Кодирование недоопределенного источника должно обеспечить для всякой порождаемой им последовательности воспроизведение какого-либо ее доопределения. Кодирование асимптотически оптимально, если средняя длина кода асимптотически равна энтропии источника. Оно универсально, если не зависит от вероятностей символов источника. В статье описан метод асимптотически оптимального универсального кодирования недоопределенных источников Бернулли, для которого процедуры кодирования и декодирования реализуемы RAM-программами почти линейной сложности.

Ключевые слова: недоопределенный источник, доопределение, энтропия недоопределенного источника, квазиэнтропия слова, комбинаторная энтропия класса, кодирование недоопределенного источника, универсальное кодирование, полиномиальный алгоритм.

DOI: 10.31857/S0555292320040075

§ 1. Определения, постановка задачи и результат

С недоопределенными данными имеют дело во многих областях информатики – в задачах распознавания, хранения и обработки данных, управления, логического синтеза и др. Поэтому целесообразно отдельно исследовать свойства недоопределенных данных, разработать методы и алгоритмы эффективного обращения с ними. Некоторые результаты в этом направлении представлены в [1]. Настоящая статья посвящена теоретически эффективному сжатию недоопределенных данных.

Задан конечный алфавит $A_0 = \{a_i,\ i \in M\}$, $M = \{0, 1, \ldots, m-1\}$, *основных* символов. Каждому непустому $T \subseteq M$ соответствует символ a_T, называемый *недоопределенным. Доопределением символа* a_T считается всякий основной символ a_i, $i \in T$. Символ a_M, доопределимый любым основным символом, называется *неопределенным* и обозначается через $*$. Основные символы a_i можно также рассматривать как недоопределенные символы, соответствующие одноэлементным подмножествам $\{i\} \subseteq M$. Под *доопределением слова* v в алфавите A понимается любое слово в алфавите A_0, полученное из v заменой каждого символа каким-либо его доопределением.

Выделена система $\mathcal{T}$ некоторых непустых подмножеств T множества M, и с ней связан *недоопределенный алфавит* $A = A_{\mathcal{T}} = \{a_T,\, T \in \mathcal{T}\}$. Положим $k = \#A = \#\mathcal{T}$,

где # означает мощность множества. В случае $A = A_0$ недоопределенный алфавит A называется *полностью определенным*, а при $A = A_0 \cup \{*\}$ – *частично определенным*. Будем рассматривать бернуллиевский источник X, порождающий (независимо) символы $a_T \in A$ с вероятностями p_T, $\sum\limits_{T \in \mathcal{T}} p_T = 1$. Он обозначается через (A, P), где $P = (p_T,\ T \in \mathcal{T})$, и называется *недоопределенным источником*. Если алфавит A полностью (частично) определен, будем говорить о полностью (частично) определенном источнике.

Задавшись набором вероятностей $Q = (q_i, i \in M)$, $\sum\limits_{i \in M} q_i = 1$, основных символов, введем функцию

$$\mathcal{H}(P, Q) = -\sum_{T \in \mathcal{T}} p_T \log \sum_{i \in T} q_i \tag{1}$$

(здесь и далее все логарифмы двоичные). *Энтропией источника* X назовем величину

$$\mathcal{H}(X) = \mathcal{H}(P) = \min_Q \mathcal{H}(P, Q). \tag{2}$$

Для полностью определенного источника в силу соотношения

$$\min_Q \left\{ -\sum_{i \in M} p_i \log q_i \right\} = -\sum_{i \in M} p_i \log p_i$$

она совпадает с энтропией Шеннона, а для частично определенного источника $X = (A_0 \cup \{*\}, P)$, $P = ((p_i,\ i \in M), p_*)$, представима [1] в виде

$$\mathcal{H}(X) = (1 - p_*) \log(1 - p_*) - \sum_{i \in M} p_i \log p_i \tag{3}$$

и достигается в (2) на наборе $Q = \left(\frac{p_0}{1 - p_*}, \ldots, \frac{p_{m-1}}{1 - p_*} \right)$.

Функция $\mathcal{H}(P)$ была введена (из эвристических соображений) в [2] в качестве меры неопределенности задач с несколькими ответами. Ее свойства изучены в работе [1]. Некоторые из них совпадают со свойствами энтропии Шеннона, некоторые получаются их модификацией, некоторые являются новыми.

Очевидно, что $\mathcal{H}(X) \geqslant 0$. Нас будет интересовать случай, когда энтропия источника $X = (A, P)$ строго положительна. Необходимым и достаточным для положительности энтропии является условие (см. [1], а также пункт 1° леммы 1 настоящей статьи)

$$\bigcap_{T:\, p_T > 0} T = \varnothing.$$

Будем рассматривать двоичное разделимое блоковое кодирование K недоопределенного источника X, использующее блоки длины n. Кодирование K должно обеспечить для всякого слова $v \in A^n$ возможность восстановления по его коду $K(v)$ какого-либо доопределения слова v. Кодирование источника (A, P) называется *универсальным* (при заданных A и n), если оно не зависит от набора вероятностей P.

Задачу кодирования недоопределенного источника можно рассматривать как специальный случай задачи кодирования с заданным критерием верности [3, 4], когда исходные сообщения в алфавите A должны быть воспроизведены в алфавите A_0, а искажение при воспроизведении символа a_i вместо a_T считается равным 0 при $i \in T$ и равным ∞ при $i \notin T$. Введенная выше энтропия $\mathcal{H}(X)$ оказалась эквивалентной [1] скорости как функции искажения (в другой терминологии – W-энтропии [5])

для этого случая. Более подробно связь кодирования недоопределенных данных и кодирования с заданным критерием верности рассмотрена в [6].

Качество кодирования K будем характеризовать *средней длиной кода*

$$\overline{\ell}_K^{(n)} = \frac{1}{n} \sum_{v \in A^n} p(v)|K(v)|, \tag{4}$$

приходящейся на символ источника. Здесь $p(v) = p_{T_1} \dots p_{T_n}$ – вероятность порождения источником X слова $v = a_{T_1} \dots a_{T_n}$, $|K(v)|$ – длина кодового слова для v.

В теории сложности принято считать процедуру эффективной, если ее сложность (число элементарных операций) не превосходит полинома от размера задачи. В качестве модели вычислений будем использовать РАМ-программу [7] с подходящим набором операций.

Основным результатом статьи является

Теорема. *Для любого недоопределенного источника X с положительной энтропией $\mathcal{H}(X)$ справедливы следующие утверждения*:

1) *При любом способе кодирования K справедливо неравенство*

$$\overline{\ell}_K^{(n)} \geqslant \mathcal{H}(X);$$

2) *Имеется метод K универсального кодирования с оценкой*

$$\overline{\ell}_K^{(n)} \leqslant \mathcal{H}(X) + O\left(\frac{\log\log n}{\log^{1/2} n}\right),$$

для которого кодирование и декодирование реализуемы РАМ-программами сложности $n^{1+o(1)}$.

Возможность асимптотически оптимального универсального кодирования недоопределенного источника общего вида была установлена в [8] методом случайного кодирования. В [9] для решения этой задачи предложена модификация метода арифметического кодирования, использующая случайные последовательности, которая также неэффективна.

В работе [10] описан метод асимптотически оптимального универсального кодирования частично определенных источников, реализуемый РАМ-программами почти линейной сложности. Но он требует многократного вычисления (для различных фрагментов кодируемых слов) энтропии и минимизирующего правую часть в (2) набора Q. В случае частично определенных источников это не вызывает трудностей, поскольку их энтропия и минимизирующий набор выразимы в явном виде (3). Недоопределенные источники общего вида требуют применения приближенных методов. При этом возникают определенные технические трудности, связанные с оценкой необходимой точности и трудоемкости ее достижения. Отметим, что в [11, Добавление] описана сходящаяся итеративная процедура вычисления минимизирующего набора, но скорость ее сходимости не установлена, и это не позволяет использовать ее для оценки трудоемкости кодирования.

В предлагаемой статье развит подход, дающий возможность избежать вычисления энтропии и минимизирующего набора и решить задачу кодирования недоопределенных источников общего вида с теми же оценками средней длины кода и сложности, что и в [10] для частично определенных источников.

Последующая часть статьи посвящена доказательству утверждения 2) теоремы. Доказательство утверждения 1) содержится в [1]. В курсе лекций [12] оценка утверждения 1) распространена на недоопределенные стационарные источники.

Приведем некоторые другие результаты, имеющие отношение к настоящей статье.

Код слова можно воспринимать как (двоичную) программу вычисления слова некоторым алгоритмом (см. классическую работу Колмогорова [13]). Если слово двоичное, то занумеровав его разряды двоичными числами, придем к булевой функции, выражающей зависимость разрядов слова от их двоичных номеров, и в качестве кода слова можно использовать двоичную запись программы вычисления этой функции. Конструктивный подход, связывающий задачи вычисления булевых функций с задачами их специального (локального) кодирования, предложен в [14].

Ряд методов исследования и сжатия недоопределенных данных возник при решении задач реализации недоопределенных булевых функций различными вычислительными средствами. Их обсуждение (с указанием авторства) имеется в работе [15], содержащей решение задачи асимптотически наилучшей реализации булевых функций с заданными числами нулевых, единичных и неопределенных значений. Свойства энтропии $\mathcal{H}(P)$ были впервые изучены в связи с задачей реализации систем недоопределенных булевых функций [11].

Метод кодирования недоопределенных слов, представленный в настоящей статье, использует некоторые подходы, разработанные при синтезе схем. Идея кодирования слова путем разбиения на короткие подслова и совместного кодирования подслов является модификацией идеи Шеннона из метода синтеза контактных схем [16]. Важную роль играет техника Нечипорука [17,18] работы с частично определенными данными, изложенная им применительно к задачам реализации булевых функций и матриц некоторыми типами схем.

Одними из центральных вопросов для настоящей статьи являются вопросы сложности кодирования. В известных исследованиях по схемной реализации недоопределенных функций вопросы трудоемкости методов построения схем фактически не рассматривались. Некоторое исключение составляет работа [19], в которой изложен почти квадратичный по трудоемкости метод построения асимптотически наилучших двоичных программ вычисления частичных булевых функций с заданной областью определения.

Как уже отмечалось, задача кодирования недоопределенных данных тесно связана с задачей кодирования с заданным критерием верности [4,20]. Близкие к сжатию недоопределенных данных постановки играют важную роль в задачах поиска информации с применением хеш-функций [21] и в задачах дерандомизации алгоритмов с использованием генераторов протыкающих множеств [22, 23].

§ 2. Квазиэнтропия и комбинаторная энтропия недоопределенных данных

Дальше считаем, что алфавит A содержит символ $* = a_M$. Если его там изначально нет, добавим, приписав вероятность 0. При этом слова, содержащие символ $*$, будут иметь нулевую вероятность и не повлияют на среднюю длину кода.

Свойства энтропии $\mathcal{H}(P)$ изучены в [1]. Сформулируем и докажем те из них, которые понадобятся в настоящей статье.

Лемма 1. *Справедливы следующие утверждения*:

1°. *Функция* $\mathcal{H}(P)$ *неотрицательна, причем*

$$\mathcal{H}(P) = 0 \iff \bigcap_{T:\, p_T>0} T \neq \varnothing.$$

2°. *Имеет место оценка*

$$\mathcal{H}(P) \leqslant \log m - \sum_{1\leqslant i\leqslant m} p^{(i)} \log i,$$

где $p^{(i)} = \sum\limits_{T:\, \#T=i} p_T$ – *вероятность символов, имеющих* i *доопределений.*

3°. *Функция* $\mathcal{H}(P)$ *вогнута, т.е. для любых* P, P' *и* θ, $0 \leqslant \theta \leqslant 1$,

$$\mathcal{H}(\theta P + (1-\theta)P') \geqslant \theta\mathcal{H}(P) + (1-\theta)\mathcal{H}(P').$$

4°. *Если источник* $X' = (A', P')$ *образован из* $X = (A, P)$ *исключением неопределенного символа и нормировкой вероятностей, т.е.* $A' = A \setminus \{*\}$, $P' = (p'_T$, $T \in \mathcal{T} \setminus \{M\})$, $p'_T = \dfrac{p_T}{1-p_*}$, *то*

$$\mathcal{H}(P) = (1-p_*)\mathcal{H}(P'). \tag{5}$$

Доказательство. 1°. Неотрицательность очевидна. Пусть минимум $\mathcal{H}(P,Q)$ в (2) достигается на наборе Q^0. Положим $T^0 = \{i \in M \mid q_i^0 > 0\}$. Если $\mathcal{H}(P) = \mathcal{H}(P,Q^0) = 0$, то для любого T с $p_T > 0$ выполнено $\sum\limits_{i\in T} q_i^0 = 1$, а потому T включает T^0, и пересечение всех таких T непусто. Обратно, если пересечение непусто, то назначив $q_i = 1$ для некоторого i из этого пересечения и $q_j = 0$ для всех $j \neq i$, получим набор Q, для которого $\mathcal{H}(P,Q) = 0$.

2°. Вычислим $\mathcal{H}(P,Q)$ на наборе $Q_0 = (1/m, \ldots, 1/m)$. Имеем

$$\mathcal{H}(P) \leqslant \mathcal{H}(P,Q_0) = -\sum_T p_T \log\frac{\#T}{m} = \log m - \sum_i \sum_{\#T=i} p_T \log i =$$
$$= \log m - \sum_i p^{(i)} \log i.$$

3°. Пусть минимум функции $\mathcal{H}(\theta P + (1-\theta)P', Q)$ достигается на наборе Q. Тогда

$$\mathcal{H}(\theta P + (1-\theta)P') = -\theta\sum_T p_T \log\sum_{i\in T} q_i - (1-\theta)\sum_T p'_T \log\sum_{i\in T} q_i \geqslant$$
$$\geqslant \theta\mathcal{H}(P) + (1-\theta)\mathcal{H}(P').$$

4°. Для любого набора вероятностей $Q = (q_i,\ i \in M)$ выполнено $\log\sum\limits_{i\in M} q_i = 0$, поэтому

$$-\sum_{T\subseteq M} p_T \log\sum_{i\in T} q_i = -(1-p_*)\sum_{T\subset M}\frac{p_T}{1-p_*}\log\sum_{i\in T} q_i.$$

Взяв минимум по Q, получаем нужное утверждение. ▲

Отметим, что выражение (3) для энтропии частично определенного источника X вытекает из пункта 4° с учетом того, что источник X' в этом случае полностью определен и его энтропия совпадает с энтропией Шеннона.

Для слова $v \in A^n$ обозначим через $r_T(v)$ число появлений в нем символа a_T, $\sum\limits_{T\in\mathcal{T}} r_T(v) = |v|$, и положим $\boldsymbol{r}(v) = (r_T(v),\ T \in \mathcal{T})$. *Квазиэнтропией слова* v назовем величину

$$h(v) = |v|\,\mathcal{H}\left(\frac{\boldsymbol{r}(v)}{|v|}\right).$$

В силу (1), (2) она может быть переписана в виде

$$h(v) = \min_Q\left\{-\sum_{T\in\mathcal{T}} r_T(v)\log\sum_{i\in T} q_i\right\}. \tag{6}$$

Следующая лемма содержит необходимые для дальнейшего свойства квазиэнтропии слов.

Лемма 2. *Справедливы следующие утверждения*:

1°. *Имеет место неравенство* $h(v) \leqslant |v| \log m$.

2°. *Если наборы* $\boldsymbol{r}(v_1)$ *и* $\boldsymbol{r}(v_2)$ *различаются лишь в компоненте* $r_*(\cdot)$, *то* $h(v_1) = h(v_2)$.

3°. *Квазиэнтропия конкатенации* $v_1 v_2$ *удовлетворяет неравенству* $h(v_1 v_2) \geqslant h(v_1) + h(v_2)$.

4°. *Если слово* v' *образовано из слова* v *приписыванием некоторого символа, то* $h(v') \leqslant h(v) + \log|v| + 2$.

Доказательство. 1°. Из пункта 2° леммы 1 получаем неравенство $\mathcal{H}(P) \leqslant \log m$, которое умножением обеих частей на $|v|$ приводит к нужному утверждению.

2°. Обозначим через v_1' и v_2' слова, полученные из v_1 и v_2 удалением символов $*$. Имеем $|v_1'| = |v_1| - r_*(v_1) = |v_2| - r_*(v_2) = |v_2'|$ и $\boldsymbol{r}(v_1') = \boldsymbol{r}(v_2')$, а потому $h(v_1') = h(v_2')$. Из пункта 4° леммы 1 при $p_T = \dfrac{r_T(v_1)}{|v_1|}$ с учетом равенства $1 - \dfrac{r_*(v_1)}{|v_1|} = \dfrac{|v_1'|}{|v_1|}$ выводим

$$\mathcal{H}\biggl(\frac{\boldsymbol{r}(v_1)}{|v_1|}\biggr) = \frac{|v_1'|}{|v_1|}\mathcal{H}\biggl(\frac{\boldsymbol{r}(v_1')}{|v_1'|}\biggr).$$

Домножая обе части на $|v_1|$, приходим к равенству $h(v_1) = h(v_1')$. Аналогично доказывается, что $h(v_2) = h(v_2')$. Отсюда и из $h(v_1') = h(v_2')$ получаем $h(v_1) = h(v_2)$.

3°. Воспользуемся вогнутостью энтропии (пункт 3° леммы 1) при $P = \dfrac{\boldsymbol{r}(v_1)}{|v_1|}$, $P' = \dfrac{\boldsymbol{r}(v_2)}{|v_2|}$ и $\theta = \dfrac{|v_1|}{|v_1| + |v_2|}$. Имеем

$$\frac{|v_1|}{|v_1| + |v_2|}\mathcal{H}\biggl(\frac{\boldsymbol{r}(v_1)}{|v_1|}\biggr) + \frac{|v_2|}{|v_1| + |v_2|}\mathcal{H}\biggl(\frac{\boldsymbol{r}(v_2)}{|v_2|}\biggr) \leqslant \mathcal{H}\biggl(\frac{\boldsymbol{r}(v_1) + \boldsymbol{r}(v_2)}{|v_1| + |v_2|}\biggr) = \mathcal{H}\biggl(\frac{\boldsymbol{r}(v_1 v_2)}{|v_1 v_2|}\biggr).$$

Домножая обе части на $|v_1| + |v_2| = |v_1 v_2|$, получаем $h(v_1) + h(v_2) \leqslant h(v_1 v_2)$.

4°. Пусть $v' = v a_{T'}$. Обозначим через $Q = (q_i,\ i \in M)$ набор, на котором в (6) достигается квазиэнтропия $h(v)$. Возьмем некоторое $s \in T'$ и образуем набор $Q' = (q_i',\ i \in M)$, положив

$$q_s' = \frac{(|v| - 1)q_s}{|v|} + \frac{1}{|v|}, \quad q_i' = \frac{(|v| - 1)q_i}{|v|}, \quad i \neq s.$$

Набор Q' удовлетворяет условию

$$\sum_{i \in M} q_i' = \frac{|v| - 1}{|v|}\sum_{i \in M} q_i + \frac{1}{|v|} = \frac{|v| - 1}{|v|} + \frac{1}{|v|} = 1.$$

С учетом этого имеем

$$\begin{aligned}
h(v') &\leqslant -\sum_{T \in \mathcal{T}} r_T(v') \log \sum_{j \in T} q_j' = -\sum_{T \in \mathcal{T}} r_T(v) \log \sum_{j \in T} q_j' - \log \sum_{j \in T'} q_j' \leqslant \\
&\leqslant -\sum_{T \in \mathcal{T}} r_T(v) \log \sum_{j \in T} \frac{|v| - 1}{|v|} q_j - \log \frac{1}{|v|} = \\
&= -\sum_{T \in \mathcal{T}} r_T(v) \log \sum_{j \in T} q_j - |v| \log \frac{|v| - 1}{|v|} + \log|v| = \\
&= h(v) - |v| \log\biggl(1 - \frac{1}{|v|}\biggr) + \log|v|.
\end{aligned}$$

Принимая во внимание, что при $|v| \geqslant 2$

$$-|v|\log\left(1-\frac{1}{|v|}\right) \leqslant -|v|\left(-\frac{1}{|v|}-\frac{1}{2|v|^2}\right)\log e = \left(1+\frac{1}{2|v|}\right)\log e \leqslant \frac{5}{4}\log e \leqslant 2,$$

приходим к оценке пункта 4°. ▲

Для заданного набора $\boldsymbol{r} = (r_T,\, T \in \mathcal{T})$ натуральных чисел положим

$$\ell = \sum_{T\in\mathcal{T}} r_T$$

и обозначим через $\mathcal{K}_\ell(\boldsymbol{r})$ класс всех слов длины ℓ в алфавите A, в которых символы $a_T \in A$ встречаются r_T раз (с частотой r_T/ℓ). Такие классы называют *частотными*. Все слова $v \in \mathcal{K}_\ell(\boldsymbol{r})$ имеют одинаковую квазиэнтропию $h(v) = \ell\mathcal{H}(\boldsymbol{r}/\ell)$, которую будем обозначать через $h_\ell(\boldsymbol{r})$ и называть *квазиэнтропией класса* $\mathcal{K}_\ell(\boldsymbol{r})$.

Пусть задано конечное множество V недоопределенных слов в алфавите A. Будем говорить, что некоторое *множество слов* в алфавите A_0 образует *доопределение множества* V, если в нем найдется доопределение каждого слова из V. Обозначим через $N(V)$ минимальную мощность множества, доопределяющего V. Величину $\log N(V)$ будем называть *комбинаторной энтропией* множества слов V.

Будем считать, что все слова из V имеют одинаковую длину ℓ. Для доопределения множества слов V может быть использована *градиентная процедура* (жадный алгоритм). Ее удобно описывать в терминах таблицы, строки которой соответствуют словам $v \in V$, столбцы – словам $w \in A_0^\ell$, а в клетке (v, w) содержится 1 либо 0 в зависимости от того, доопределяет w слово v или нет. Будем считать, что столбцы упорядочены в соответствии с лексикографическим упорядочением слов множества A_0^ℓ. Градиентная процедура реализуется в виде последовательности шагов, на каждом из которых в таблице, полученной после предыдущего шага, выбирается первый столбец с наибольшим числом единиц и вычеркивается вместе со строками, содержащими в нем единицы. Совокупность столбцов, выбранных к моменту, когда все строки окажутся вычеркнутыми, задает доопределяющее множество для V. Оно определено однозначно; его мощность обозначим через $N^G(V)$.

При $V = \mathcal{K}_\ell(\boldsymbol{r})$ вместо записей $N(\mathcal{K}_\ell(\boldsymbol{r}))$ и $N^G(\mathcal{K}_\ell(\boldsymbol{r}))$ будем использовать $N_\ell(\boldsymbol{r})$ и $N_\ell^G(\boldsymbol{r})$. Величина $\log N_\ell^G(\boldsymbol{r})$ оценивает сверху комбинаторную энтропию $\log N_\ell(\boldsymbol{r})$ класса $\mathcal{K}_\ell(\boldsymbol{r})$.

Всюду дальше буквами C с индексами и (или) пометками обозначаются некоторые константы, абсолютные или зависящие от мощностей m и k алфавитов A_0 и A. При необходимости они могут быть указаны явно.

Лемма 3. *Имеют место оценки*

$$h_\ell(\boldsymbol{r}) - C_1\log\ell \leqslant \log N_\ell^G(\boldsymbol{r}) \leqslant h_\ell(\boldsymbol{r}) + C_2\log\ell.$$

Доказательство. В работе [8] (см. также [1]) с использованием метода случайного кодирования получены оценки комбинаторной энтропии частотного класса $\mathcal{K}_\ell(\boldsymbol{r})$

$$h_\ell(\boldsymbol{r}) - C_1\log\ell \leqslant \log N_\ell(\boldsymbol{r}) \leqslant h_\ell(\boldsymbol{r}) + C_1'\log\ell. \tag{7}$$

Воспользуемся результатом из [24] (см. также [25]) о точности градиентного алгоритма. Применительно к данному случаю он приобретает вид

$$N_\ell^G(\boldsymbol{r}) \leqslant N_\ell(\boldsymbol{r})\left(1+\ln\frac{\#\mathcal{K}_\ell(\boldsymbol{r})}{N_\ell(\boldsymbol{r})}\right)$$

и дает

$$N_\ell^G(\boldsymbol{r}) \leqslant N_\ell(\boldsymbol{r}) \ln(\#\mathcal{K}_\ell(\boldsymbol{r})) \leqslant N_\ell(\boldsymbol{r})\, \ell \ln k.$$

Отсюда и из (7) получаем верхнюю оценку

$$\log N_\ell^G(\boldsymbol{r}) \leqslant h_\ell(\boldsymbol{r}) + C_1' \log \ell + \log \ell + \log \ln k \leqslant h_\ell(\boldsymbol{r}) + C_2 \log \ell.$$

Нижняя оценка следует из (7) и неравенства $N_\ell(\boldsymbol{r}) \leqslant N_\ell^G(\boldsymbol{r})$. ▲

Будем рассматривать также *t-ограниченную* градиентную процедуру доопределения множества V, где t – заданный натуральный параметр. Она реализуется в соответствии с предыдущим описанием и считается *результативной*, если завершается не более чем за t шагов. В противном случае, если после t шагов остались невычеркнутые строки, процедура прерывается *безрезультатно*.

Лемма 4. *Если t-ограниченная градиентная процедура доопределения класса $\mathcal{K}_\ell(\boldsymbol{r})$ завершается результативно, то*

$$h_\ell(\boldsymbol{r}) \leqslant \log t + C_1 \log \ell,$$

а если безрезультатно, то

$$h_\ell(\boldsymbol{r}) \geqslant \log t - C_2 \log \ell.$$

Доказательство. Установим первое соотношение, второе доказывается аналогично. Если процедура завершилась результативно, то справедливо неравенство $N_\ell^G(\boldsymbol{r}) \leqslant t$, а потому в силу леммы 3 выполнено $\log t \geqslant \log N_\ell^G(\boldsymbol{r}) \geqslant h_\ell(\boldsymbol{r}) - C_1 \log \ell$, откуда $h_\ell(\boldsymbol{r}) \leqslant \log t + C_1 \log \ell$. ▲

§ 3. Кодирование

Алфавиты A_0, A и длину n блоков считаем заданными.

Будем говорить, что слово w в алфавите A_0 *обобщенно доопределяет* слово u в алфавите A, если $|u| \leqslant |w|$ и начало длины $|u|$ слова w доопределяет u. Метод кодирования недоопределенных слов $v \in A^n$ использует некоторый натуральный параметр $\lambda = \lambda(n)$ и некоторое множество $\mathcal{D}$, $\mathcal{D} \subseteq A_0^\lambda$, *допустимых обобщенных доопределений*. Они будут указаны позже, а пока потребуем лишь, чтобы для каждого символа $a_i \in A_0$ в $\mathcal{D}$ имелось слово, начинающееся с a_i. Пусть $\mathcal{D} = \{w_1, w_2, \ldots, w_d\}$, где $d = \#\mathcal{D}$ и слова w_s упорядочены лексикографически.

Для каждого $v \in A^n$ кодовое слово будет иметь вид $K(v) = K_0\, K_1(v)$, где подслово K_0, одинаковое для всех v, называется *справочной частью* кодового слова, а $K_1(v)$ – его *основной частью*. Опишем способ их построения.

При кодировании слова $v \in A^n$ от него последовательно отрезаются слева подслова $v_1, v_2, \ldots$ максимально возможной длины, имеющие в множестве $\mathcal{D}$ обобщенные доопределения. Условие, наложенное на множество $\mathcal{D}$, гарантирует, что процедура не прервется, пока слово v не будет исчерпано. Полученные подслова v_i будем называть *фрагментами* слова v. Если число фрагментов равно $t = t(v)$, то $v = v_1 v_2 \ldots v_t$.

Доопределение фрагмента v_i может быть задано парой чисел (s_i, ℓ_i), где s_i – наименьшее s, при котором слово w_s обобщенно доопределяет v_i, а ℓ_i – длина фрагмента v_i. Положим $\alpha = \lceil \log d \rceil$, $\beta = \lceil \log \lambda \rceil$, где $\lceil x \rceil$ означает наименьшее целое, не меньшее числа x, и фрагменту v_i сопоставим слово (код) $K(v_i) = \widetilde{s}_i^{(\alpha)} \widetilde{\ell}_i^{(\beta)}$, где для натуральных чисел q и γ, $\gamma \geqslant \log(q+1)$, через $\widetilde{q}^{(\gamma)}$ обозначена γ-разрядная двоичная запись числа q. Основная часть кодового слова $K(v)$ образуется путем приписывания друг к другу кодов фрагментов

$$K_1(v) = K(v_1) \ldots K(v_t) = \widetilde{s}_1^{(\alpha)} \widetilde{\ell}_1^{(\beta)} \ldots \widetilde{s}_t^{(\alpha)} \widetilde{\ell}_t^{(\beta)}. \tag{8}$$

Отметим, что указанное кодирование фрагментов обладает полезным свойством равномерности по выходу: фрагменты v_i могут иметь разную длину, но кодируются двоичными словами $K(v_i)$ одинаковой длины $\alpha+\beta$.

Справочная часть K_0 кодирует множество $\mathcal{D}$. Она содержит двоичные представления его параметров λ и d и двоичные представления входящих в него слов w_s. Для представления натурального числа q используется двоичное слово $\widehat{q} = q_1 q_1 \dots q_r q_r 01$, где $\overline{q_1 \dots q_r}$ – двоичная запись числа q минимальной длины. Такое представление чисел позволяет по слову $\widehat{q}u$, где u – некоторое двоичное слово, однозначно восстановить q и u. Очевидно, что

$$|\widehat{q}| \leqslant 2(\log q + 2). \tag{9}$$

Слово $w_s \in \mathcal{D} \subseteq A_0^\lambda$ будем представлять двоичным словом $\widetilde{w}_s$ длины $m\lambda$, полученным из w_s заменой символов a_i, $i = 0, 1, \dots, m-1$, словами $0 \dots 010 \dots 0$ длины m, содержащими 1 в разряде i. Справочная часть имеет вид

$$K_0 = \widehat{\lambda}\widehat{d}\widetilde{w}_1 \dots \widetilde{w}_d. \tag{10}$$

Лемма 5. *При любом выборе множества $\mathcal{D}$ обобщенных доопределений описанное кодирование слов $v \in A^n$ разделимо (префиксно) и позволяет восстановить по коду $K(v)$ слова v некоторое его доопределение.*

Доказательство. Опишем способ декодирования. Пусть имеется двоичное слово, начинающееся с кодового слова $K(v) = K_0\, K_1(v)$. Отрежем от него подслова $\widehat{\lambda}$ и $\widehat{d}$ и по ним найдем параметры λ, d, α и β. Затем отрежем d подслов длины $m\lambda$. Они соответствуют словам $\widetilde{w}_s$ и позволяют найти слова w_s множества $\mathcal{D}$. Далее от оставшейся части слова будем последовательно отрезать подслова длины $\alpha+\beta$. Они являются кодами $K(v_i) = \widetilde{s}_i^{(\alpha)}\widetilde{\ell}_i^{(\beta)}$ фрагментов v_i и задают доопределения u_i этих фрагментов. Процедура завершится после того, как при некотором t длина слова $u_1 \dots u_t$ окажется равной n. Это слово доопределяет v.

Префиксность кода следует из того, что по двоичной последовательности, начинающейся с кодового слова, это слово находится однозначно. ▲

Качество кодирования зависит от выбора множества $\mathcal{D}$. Будем использовать следующий способ построения этого множества. Зададимся некоторыми натуральными параметрами $\lambda = \lambda(n)$ и $\tau = \tau(n)$, удовлетворяющими условию

$$\log \lambda = o(\tau). \tag{11}$$

При этих λ и τ к каждому частотному классу $\mathcal{K}_\lambda(\boldsymbol{r})$ применим 2^τ-ограниченную градиентную процедуру и обозначим через $\mathcal{D}_{\lambda,\tau}$ объединение полученных доопределений для всех классов $\mathcal{K}_\lambda(\boldsymbol{r})$, в применении к которым процедура завершилась результативно. В качестве $\mathcal{D}$ возьмем множество $\mathcal{D}_{\lambda,\tau}$.

Оценим мощность $d = d_{\lambda,\tau}$ множества $\mathcal{D}_{\lambda,\tau}$. Для всякого класса $\mathcal{K}_\lambda(\boldsymbol{r})$, для которого 2^τ-ограниченная градиентная процедура завершилась результативно, мощность доопределяющего множества не превосходит 2^τ. Число таких классов не больше общего числа частотных классов в A^λ, которое не превышает $\binom{\lambda+k-1}{k-1}$, где $k = \#A$. Поэтому

$$d \leqslant (\lambda+k-1)^{k-1} 2^\tau \leqslant \lambda^{C_3} 2^\tau. \tag{12}$$

При этом параметры $\alpha = \lceil \log d \rceil$ и $\beta = \lceil \log \lambda \rceil$ будут удовлетворять оценкам

$$\alpha \leqslant \tau + C_3 \log \lambda, \quad \beta \leqslant \log \lambda + 1. \tag{13}$$

Без ограничения общности можно считать, что 2^τ не превосходит числа m^λ столбцов градиентной таблицы, а потому

$$\tau \leqslant \lambda \log m. \tag{14}$$

Приведем некоторые предварительные (не совсем строгие) соображения, объясняющие выбор $\mathcal{D}_{\lambda,\tau}$ в качестве множества допустимых обобщенных доопределений. Из леммы 4 при $t = 2^\tau$ вытекает, что квазиэнтропия фрагментов v_i слова v не превышает величины $\tau + C_1 \log \lambda$, которая в силу условия (11) асимптотически равна τ. При этом, вследствие максимальности отрезаемых фрагментов, их квазиэнтропия в типичном случае окажется асимптотически равной τ. Длина $\alpha + \beta$ кода фрагментов, асимптотически равная τ в силу (13), будет в типичном случае асимптотически минимальной, что приведет к асимптотически наилучшему кодированию источника. Далее, если назначить параметр λ удовлетворяющим условию $\lambda = o(\log n)$, то сложность построения множества $\mathcal{D}_{\lambda,\tau}$ составит $n^{o(1)}$, а кодирование и декодирование слова v при заданном явно множестве $\mathcal{D}_{\lambda,\tau}$ будут выполнимы с трудоемкостью $n^{1+o(1)}$.

§ 4. Длина кодовых слов

Лемма 6. *Длина справочной части кода удовлетворяет оценке*

$$|K_0| \leqslant \lambda^{C_4} 2^\tau. \tag{15}$$

Доказательство. Воспользуемся представлением (10). Все слова $\widetilde{w}_s$ имеют длину $m\lambda$, а потому в силу (9) и (12)

$$|K_0| \leqslant 2(\log \lambda + 2) + 2(C_3 \log \lambda + \tau + 2) + \lambda^{C_3} 2^\tau m\lambda \leqslant \lambda^{C_4} 2^\tau. \quad \blacktriangle$$

Лемма 7. *Длина основной части $K_1(v)$ кодового слова для $v \in A^n$ удовлетворяет оценке*

$$|K_1(v)| \leqslant h(v) + n\left(\frac{C_6 \log \lambda}{\tau - C_5 \log \lambda} + \frac{\tau + C_6 \log \lambda}{\lambda}\right) + \tau, \tag{16}$$

где $h(v)$ – квазиэнтропия слова v.

Доказательство. Будем говорить, что фрагмент v_i слова v имеет тип 1, если $|v_i| < \lambda$, и тип 2, если $|v_i| = \lambda$. Числа фрагментов типа 1 и типа 2 в слове v обозначим, соответственно, через t_1 и t_2. Очевидно,

$$t_2 \leqslant \frac{n}{\lambda}. \tag{17}$$

Оценим t_1. Пусть v_i – фрагмент типа 1, не являющийся заключительным в слове v, а a_T – следующий за ним символ слова v. Образуем слово $v' = v_i a_T * \ldots *$ длины λ путем приписывания к слову $v_i a_T$ подходящего числа символов $*$. Слово $v_i a_T$ в силу максимальности фрагмента v_i не имеет обобщенных доопределений в классе $\mathcal{D}_{\lambda,\tau}$, а потому слово v' не доопределимо в этом классе. Это означает, что 2^τ-ограниченная процедура доопределения в применении к частотному классу $\mathcal{K}_\lambda(\boldsymbol{r}')$, содержащему слово v', закончилась безрезультатно, а потому в силу леммы 4

$$h(v') = h_\lambda(\boldsymbol{r}') \geqslant \tau - C_2 \log \lambda.$$

Слово v' образовано из $v_i a_T$ добавлением символов $*$, и в соответствии с пунктом 2° леммы 2 выполнено

$$h(v_i a_T) = h(v') \geqslant \tau - C_2 \log \lambda.$$

Воспользовавшись пунктом 4° леммы 2, получаем

$$h(v_i) \geqslant h(v_i a_T) - \log|v_i| - 2 \geqslant \tau - (C_2+1)\log\lambda - 2 \geqslant \tau - C_5\log\lambda.$$

Пункт 3° леммы 2 и это неравенство приводят к оценке

$$h(v) = h(v_1 \dots v_t) \geqslant \sum_{1\leqslant i\leqslant t} h(v_i) \geqslant \sum_{i:\ i<t, |v_i|<\lambda} h(v_i) \geqslant (t_1 - 1)(\tau - C_5\log\lambda),$$

из которой следует соотношение

$$t_1 \leqslant \frac{h(v)}{\tau - C_5\log\lambda} + 1.$$

Оно в сочетании с (17) дает оценку числа t фрагментов v_i в слове v:

$$t = t_1 + t_2 \leqslant \frac{h(v)}{\tau - C_5\log\lambda} + \frac{n}{\lambda} + 1. \tag{18}$$

Оценим длину $|K_1(v)|$ основной части кодового слова. Из представления (8) следует, что $|K_1(v)| = t(\alpha+\beta)$. Подставляя сюда (18) и оценку $\alpha+\beta \leqslant \tau + C'\log\lambda$, вытекающую из (13), получаем

$$|K_1(v)| \leqslant \frac{h(v)(\tau + C'\log\lambda)}{\tau - C_5\log\lambda} + \frac{n(\tau + C'\log\lambda)}{\lambda} + \tau + C'\log\lambda. \tag{19}$$

Первое слагаемое этой суммы, которое обозначим через B, преобразуется к виду

$$B = h(v) + h(v)\frac{(C_5 + C')\log\lambda}{\tau - C_5\log\lambda}$$

и с учетом пункта 1° леммы 2 допускает оценку

$$B \leqslant h(v) + \frac{n(C_5 + C')\log m\log\lambda}{\tau - C_5\log\lambda} \leqslant h(v) + \frac{C_6 n\log\lambda}{\tau - C_5\log\lambda}.$$

Ее подстановка в (19) приводит к (16). ▲

Из (15) и (16) вытекает оценка длины кодовых слов

$$|K(v)| \leqslant h(v) + G(n,\lambda,\tau), \tag{20}$$

где

$$G(n,\lambda,\tau) = n\left(\frac{C_6\log\lambda}{\tau - C_5\log\lambda} + \frac{\tau + C_6\log\lambda}{\lambda}\right) + \lambda^{C_4}2^{\tau}. \tag{21}$$

§ 5. Средняя длина кода

Лемма 8. *При заданных параметрах $\lambda = \lambda(n)$ и $\tau = \tau(n)$ средняя длина кода удовлетворяет оценке*

$$\overline{\ell}_K^{(n)} \leqslant \mathcal{H}(X) + \frac{G(n,\lambda,\tau)}{n},$$

где функция $G(n,\lambda,\tau)$ задается равенством (21).

Доказательство. Подстановка оценок (20) в выражение (4) средней длины кода дает

$$\overline{\ell}_K^{(n)} \leqslant \frac{1}{n}\sum_{v\in A^n} p(v)h(v) + \frac{G(n,\lambda,\tau)}{n}. \tag{22}$$

Обозначим через $p_n(\boldsymbol{r})$ вероятность порождения источником $X = (A, P)$ слов класса $\mathcal{K}_n(\boldsymbol{r})$, $\boldsymbol{r} = (r_T,\, T \in \mathcal{T})$. Вероятности $p_n(\boldsymbol{r})$ образуют полиномиальное (мультиномиальное) распределение [26]

$$p_n(\boldsymbol{r}) = \frac{n!}{\prod\limits_{T\in\mathcal{T}} r_T!}\prod_{T\in\mathcal{T}} p_T^{r_T}.$$

Оно обладает свойством

$$\sum_{\boldsymbol{r}} p_n(\boldsymbol{r})\frac{r_T}{n} = p_T, \quad T \in \mathcal{T}, \tag{23}$$

где r_T и p_T – компоненты наборов $\boldsymbol{r}$ и P.

В выражении (22) сгруппируем слова $v \in A^n$ по их принадлежности частотным классам $\mathcal{K}_n(\boldsymbol{r})$. Поскольку слова класса $\mathcal{K}_n(\boldsymbol{r})$ имеют одинаковую квазиэнтропию $h(v) = h_n(\boldsymbol{r}) = n\mathcal{H}\Big(\dfrac{\boldsymbol{r}}{n}\Big)$, это дает

$$\overline{\ell}_K^{(n)} \leqslant \sum_{\boldsymbol{r}} p_n(\boldsymbol{r})\mathcal{H}\left(\frac{\boldsymbol{r}}{n}\right) + \frac{G(n,\lambda,\tau)}{n}.$$

Применив к вогнутой функции $\mathcal{H}$ (пункт 3° леммы 1) неравенство Йенсена и используя (23), получаем

$$\overline{\ell}_K^{(n)} \leqslant \mathcal{H}\left(\sum_{\boldsymbol{r}} p_n(\boldsymbol{r})\frac{\boldsymbol{r}}{n}\right) + \frac{G(n,\lambda,\tau)}{n} = \mathcal{H}(P) + \frac{G(n,\lambda,\tau)}{n}. \quad \blacktriangle$$

§ 6. Сложность кодирования и декодирования

Лемма 9. *Справедливы следующие утверждения*:

1. *Сложность построения справочной части кодового слова не превосходит величину* $\lambda^{C_7}\big((C_8)^\lambda + 2^{2\tau}\big)$;
2. *Сложность построения основной части кодового слова при заданной справочной части не превосходит* $n\lambda^{C_9}2^\tau$;
3. *Сложность декодирования не превосходит* $(n + 2^\tau)\lambda^{C_{10}}$.

Доказательство. 1. Сложность градиентной процедуры доопределения класса $\mathcal{K}_\lambda(\boldsymbol{r})$ полиномиальна относительно числа клеток градиентной таблицы. Числа строк и столбцов таблицы не более чем экспоненциальны по λ, а потому сложность процедуры оценивается величиной C^λ. То же относится и к 2^τ-ограниченной градиентной процедуре. Общее число классов $\mathcal{K}_\lambda(\boldsymbol{r})$ не превосходит $\lambda^{C'}$, и следовательно, на реализацию 2^τ-ограниченных процедур для этих классов затрачивается не более $\lambda^{C'}C^\lambda$ операций. При этом суммарное число слов, вошедших в доопределения классов $\mathcal{K}_\lambda(\boldsymbol{r})$, для которых 2^τ-ограниченная процедура завершилась результативно, не превышает $\lambda^{C'}2^\tau$. Их лексикографическое упорядочивание с одновременным удалением повторяющихся слов требует не более $\lambda^{C''}2^{2\tau}$ операций. Нетрудно понять, что суммарная трудоемкость других операций, используемых при построении

справочной части, ограничена величиной $\lambda^{\widehat{C}}2^\tau$. Будем считать, что она учтена константой C'', и таким образом, общая сложность построения справочной части оценивается величиной $\lambda^{C'}C^\lambda + \lambda^{C''}2^{2\tau}$. Полагая $C_7 = \max\{C', C''\}$, $C_8 = C$, приходим к утверждению пункта 1.

2. При построении основной части $K_1(v)$ кодового слова вначале по справочной части K_0 восстанавливается множество $\mathcal{D}_{\lambda,\tau}$ и словам w_s этого множества приписываются двоичные номера $\widetilde{s}^{(\alpha)}$. С учетом оценки (15) нетрудно понять, что трудоемкость этого не превосходит $\lambda^C 2^\tau$ при некотором C.

Далее формируется основная часть $K_1(v)$ в результате последовательности шагов i. Результатом шага i являются слово $v^{(i)}$, полученное из v удалением фрагментов $v_1, \ldots, v_i$, и слово

$$K(v_1 \ldots v_i) = \widetilde{s}_1^{(\alpha)}\widetilde{\ell}_1^{(\beta)} \ldots \widetilde{s}_i^{(\alpha)}\widetilde{\ell}_i^{(\beta)}.$$

На шаге i последовательно образуются слова $v_{i,1}, v_{i,2}, \ldots$, где $v_{i,j}$ – начальное подслово длины j слова $v^{(i-1)}$, и для каждого из них ищется первое в порядке расположения в $\mathcal{D}$ его обобщенное доопределение w_s. Шаг i завершится после того, как при некотором $j = \ell$ окажется, что обобщенных доопределений слова $v_{i,\ell}$ словами w_s нет. Тогда полагаем $v_i = v_{i,\ell-1}$, $\ell_i = \ell - 1$ и берем в качестве s_i номер слова, обобщенно доопределяющего слово $v_{i,\ell-1}$. Слово $v^{(i)}$ образуем из $v^{(i-1)}$ удалением фрагмента v_i, а слово $K(v_1 \ldots v_i)$ – дописыванием к $K(v_1 \ldots v_{i-1})$ слова $\widetilde{s}_i^{(\alpha)}\widetilde{\ell}_i^{(\beta)}$. Процедура завершится шагом t, после которого слово $v^{(t)}$ окажется пустым.

Основная трудоемкость этого этапа приходится на перебор пар $(v_{i,j}, w_s)$ и выяснение, является ли w_s обобщенным доопределением слова $v_{i,\ell}$. При формировании фрагмента v_i используется $\ell_i + 1$ слов $v_{i,j}$, а потому общее число слов $v_{i,j}$, участвующих в сравнениях, не превосходит

$$\ell_1 + \ldots + \ell_t + t \leqslant n + t \leqslant 2n.$$

Число слов w_s ограничено величиной $\lambda^{C_3}2^\tau$ (см. (12)). Поэтому общее число пар $(v_{i,j}, w_s)$ не больше $2n\lambda^{C_3}2^\tau$. Число операций, связанных с одним сравнением, полиномиально по λ, и следовательно, на все сравнения затрачивается не более $n\lambda^{C'}2^\tau$ операций. Нетрудно видеть, что учет других операций не изменяет характера этой оценки, и будем считать, что она их учитывает.

Суммирование полученных оценок дает величину $\lambda^C 2^\tau + n\lambda^{C'}2^\tau$, которая не превосходит $n\lambda^{C_9}2^\tau$ при некотором C_9.

3. При декодировании вначале по справочной части K_0 восстанавливается множество $\mathcal{D}_{\lambda,\tau}$ и словам w_s этого множества приписываются двоичные номера $\widetilde{s}^{(\alpha)}$. На это затрачивается $\lambda^C 2^\tau$ операций. Затем основная часть разбивается на подслова $\widetilde{s}_i^{(\alpha)}\widetilde{\ell}_i^{(\beta)}$ длины $\alpha + \beta$, а они, в свою очередь, делятся на части длины α и β. Эти части задают слово w_{s_i}, обобщенно доопределяющее фрагмент v_i, и длину ℓ_i его подслова, доопределяющего v_i. Доопределение слова v образуется заменой фрагментов v_i найденными доопределениями.

Доопределение одного фрагмента v_i выполнимо с полиномиальной относительно λ сложностью. Поскольку число t фрагментов не превосходит n, на все фрагменты затрачивается не более $n\lambda^{C'}$ операций. Нетрудно видеть, что эта величина при подходящем выборе C' покрывает также сложность других операций, используемых на втором этапе. Суммарная сложность декодирования составляет $\lambda^C 2^\tau + n\lambda^{C'}$. Выбирая $C_{10} = \max(C, C')$, приходим к утверждению 3 леммы. ▲

§ 7. Выбор параметров и завершение доказательства

Задавшись функцией $\varphi(n)$, удовлетворяющей условиям

$$\varphi(n) \to \infty, \quad \varphi(n) = o\big(\log^{1/2}(n)\big), \tag{24}$$

назначим параметры $\lambda = \lambda(n)$ и $\tau = \tau(n)$, положив

$$\lambda(n) = \left\lfloor \frac{\log n}{\varphi^2(n)} \right\rfloor, \tag{25}$$

$$\tau(n) = \left\lfloor \frac{(\log n \log\log n)^{1/2}}{\varphi(n)} \right\rfloor, \tag{26}$$

где через $\lfloor x \rfloor$ обозначено наибольшее целое, не большее числа x.

Лемма 10. *Если функция $\varphi(n)$ удовлетворяет условиям* (24), *то метод кодирования K при значениях* (25) *и* (26) *параметров λ и τ*

1) *обеспечивает среднюю длину кода*

$$\overline{\ell}_K^{(n)} \leqslant \mathcal{H}(X) + O\left(\varphi(n)\left(\frac{\log\log n}{\log n}\right)^{1/2}\right);$$

2) *допускает кодирование и декодирование со сложностью $n^{1+o(1)}$.*

Доказательство. 1. Из леммы 8 следует, что

$$\overline{\ell}_K^{(n)} \leqslant \mathcal{H}(X) + \frac{G(n)}{n}, \tag{27}$$

где $G(n) = G(n, \lambda(n), \tau(n))$ – результат подстановки в функцию (21) значений (25) и (26). Представим $G(n, \lambda, \tau)$ в виде $A_1 + A_2 + A_3$, где

$$A_1 = \frac{C_6\, n \log\lambda}{\tau - C_5 \log\lambda}, \quad A_2 = \frac{n(\tau + C_6 \log\lambda)}{\lambda}, \quad A_3 = \lambda^{C_4} 2^{\tau}.$$

Из (25) с учетом (24) следует, что $\lambda \to \infty$ и $\log\lambda \leqslant \log\log n$. Сравнение равенств (25) и (26) показывает, что $\tau \geqslant (\lambda \log\lambda)^{1/2} - 1$, а потому $\log\lambda = o(\tau)$. Это означает, что выполнено условие (11) и имеют место асимптотические равенства $\tau - C_5 \log\lambda \sim \tau$ и $\tau + C_6 \log\lambda \sim \tau$. Принимая их во внимание, получаем

$$A_1 \sim \frac{C_6 n \log\lambda}{\tau} \lesssim \frac{C_6\, n \log\log n\, \varphi(n)}{(\log n \log\log n)^{1/2}} \lesssim C_6 n\, \varphi(n)\left(\frac{\log\log n}{\log n}\right)^{1/2},$$

$$A_2 \sim \frac{n\tau}{\lambda} \sim \frac{n(\log n \log\log n)^{1/2}\varphi(n)}{\log n} \sim n\, \varphi(n)\left(\frac{\log\log n}{\log n}\right)^{1/2}.$$

Из последнего соотношения следует, в частности, что $\log A_2 \sim \log n$. В то же время, $\log A_3 = C_4 \log\lambda + \tau \sim \tau = o(\log n)$, а потому $\log A_3 = o(\log A_2)$, и тем более, $A_3 = o(A_2)$.

Суммируя оценки и опуская малые члены, получаем

$$D(n) \lesssim (C_6 + 1) n \varphi(n)\left(\frac{\log\log n}{\log n}\right)^{1/2},$$

а потому

$$\frac{D(n)}{n} = O\left(\varphi(n)\left(\frac{\log\log n}{\log n}\right)^{1/2}\right).$$

Подстановка этого соотношения в (27) дает первое утверждение леммы.

2. Параметры λ и τ, заданные равенствами (25) и (26), удовлетворяют условиям $\lambda = o(\log n)$ и $\tau = o(\log n)$. Подстановка этих соотношений в оценки сложности кодирования и декодирования из леммы 9 показывают, что каждая из этих оценок не превосходит $n^{1+o(1)}$. ▲

Основная теорема вытекает из леммы 10 при $\varphi(n) = (\log\log n)^{1/2}$.

СПИСОК ЛИТЕРАТУРЫ

1. *Шоломов Л.А.* Элементы теории недоопределенной информации // Прикл. дискр. матем. 2009. Приложение № 2 (Лекции, прочитанные на Международной конференции с элементами научной школы для молодежи “Компьютерная безопасность и криптография”. Омск, ОмГТУ. 7–12 сентября 2009 г.). С. 18–42.
2. *Бонгард М.М.* О понятии “полезная информация” // Проблемы кибернетики. Вып. 9. М.: Физматгиз, 1963. С. 71–102.
3. *Shannon C.E.* Coding Theorems for a Discrete Source with Fidelity Criterion // IRE Nat. Conv. Rec. 1959. V. 7. № 4. P. 142–163. (Русск. перевод в *Шеннон К.Э.* Работы по теории информации и кибернетике. М.: Изд-во иностр. лит., 1963. С. 587–621).
4. *Галлагер Р.* Теория информации и надежная связь. М.: Сов. радио, 1974.
5. Вероятность и математическая статистика. Энциклопедический словарь / Под ред. Ю.В. Прохорова. М: Большая российская энциклопедия, 1999.
6. *Шоломов Л.А.* О кодировании недоопределенных последовательностей с заданной точностью воспроизведения // ДАН. 2009. Т. 429. № 5. С. 605–609.
7. *Ахо А., Хопкрофт Дж., Ульман Дж.* Построение и анализ вычислительных алгоритмов. М.: Мир, 1979.
8. *Шоломов Л.А.* Сжатие частично определенной информации // Нелинейная динамика и управление. Вып. 4 / Под ред. С.В. Емельянова, С.К. Коровина. М.: Физматлит, 2004. С. 377–396.
9. *Потапов В.Н.* Арифметическое кодирование сообщений с использованием случайных последовательностей // Прикл. дискр. матем. 2008. № 2 (2). С. 131–133.
10. *Шоломов Л.А.* Теоретически эффективное асимптотически оптимальное универсальное кодирование частично определенных источников // Прикл. дискр. матем. 2020. № 47. С. 30–56.
11. *Шоломов Л.А.* Информационные свойства функционалов сложности для систем недоопределенных булевых функций // Проблемы кибернетики. Вып. 34. М.: Наука, 1978. С. 133–150.
12. *Потапов В.Н.* Введение в теорию информации. Ижевск: НИЦ “Регулярная и хаотическая динамика”, 2014.
13. *Колмогоров А.Н.* Три подхода к определению понятия “количество информации” // Пробл. передачи информ. 1965. Т. 1. № 1. С. 3–11.
14. *Лупанов О.Б.* Об одном подходе к синтезу схем – принципе локального кодирования // Проблемы кибернетики. Вып. 14. М.: Наука, 1965. С. 31–110.
15. *Чашкин А.В.* Методы вычисления частичных булевых функций // Тр. VII Межд. конф. “Дискретные модели в теории управляющих систем” (Покровское, Моск. обл., 4–6 марта 2006 г.). М.: МАКС Пресс, 2006. С. 390–404.
16. *Shannon C.E.* The Synthesis of Two-Terminal Switching Circuits // Bell Syst. Tech. J. 1949. V. 98. № 1. P. 59–98. (Русск. перевод в *Шеннон К.Э.* Работы по теории информации и кибернетике. М.: Изд-во иностр. лит., 1963. С. 59–101.)

17. *Нечипорук Э.И.* О сложности вентильных схем, реализующих булевские матрицы с неопределенными элементами // ДАН СССР. 1965. Т. 163. № 1. С. 40–42.
18. *Нечипорук Э.И.* О топологических принципах самокорректирования // Проблемы кибернетики. Вып. 21. М.: Наука, 1969. С. 5–102.
19. *Krichevsky R.E.* Occam's Razor, Partially Specified Boolean Functions, String Matching, and Independent Sets // Inform. and Comput. 1994. V. 108. № 1. P. 158–174.
20. *Berger T.* Rate Distortion Theory: A Mathematical Basis for Data Compression. Englewood Cliffs, NJ: Prentice-Hall, 1971.
21. *Krichevsky R.* Universal Compression and Retrieval. Dordrecht: Kluwer, 1994.
22. *Andreev A.E., Clementi A.E.F., Rolim J.D.P.* Hitting Sets Derandomize BPP // Proc. 23rd Int. Colloq. on Automata, Languages and Programming (ICALP'96). Paderborn, Germany. July 8–12, 1996. Lect. Notes Comp. Sci. V. 1099. Berlin: Springer, 1996. P. 357–368.
23. *Goldreich O., Wigderson A.* Improved Derandomization of BPP Using a Hitting Set Generator // Randomization, Approximation, and Combinatorial Optimization: Algorithms and Techniques (Proc. 3rd Int. Workshop on Randomization and Approximation Techniques in Computer Science, and 2nd Int. Workshop on Approximation Algorithms for Combinatorial Optimization Problems [RANDOM-APPROX'99]. Berkeley, CA, USA. August 8–11, 1999). Lect. Notes Comp. Sci. V. 1671. Berlin: Springer, 1999. P. 131–137.
24. *Нигматуллин Р.Г.* Метод наискорейшего спуска в задачах на покрытие // Тр. симпоз. "Вопросы точности и эффективности вычислительных алгоритмов". Киев: Ин-т кибернетики АН УССР, 1969. Т. 5. С. 116–126.
25. *Нигматуллин Р.Г.* Сложность булевых функций. М.: Наука, 1991.
26. *Крамер Г.* Математические методы статистики. М.: Мир, 1975.

Шоломов Лев Абрамович
ФИЦ "Информатика и управление" РАН
Институт системного анализа РАН
levshol@mail.ru

Поступила в редакцию
28.05.2020
После доработки
13.11.2020
Принята к публикации
23.11.2020